Measurement in Research on Perceptions of Probability and Risk

Dissertation

zur Erlangung des akademischen Grades
Doktor der Philosophie (Dr. phil.)

der
Erziehungswissenschaftlichen Fakultät
der Universität Erfurt

vorgelegt von
Niels Haase

Erfurt 2016

Erstes Gutachten: Prof. Dr. Tilmann Betsch (Universität Erfurt)
Zweites Gutachten: Prof. Dr. Claudia Steinbrink (Universität Erfurt)
Drittes Gutachten: Prof. Dr. Ulf-Dietrich Reips (Universität Konstanz)

Tag der Disputation: 21.04.2017
Datum der Promotion: 21.04.2017

urn:nbn:de:gbv:547-201700095

Ehrenwörtliche Erklärung

Ich erkläre hiermit ehrenwörtlich, dass ich die vorliegende Arbeit ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Bei der Auswahl und Auswertung des Materials sowie bei der Herstellung des Manuskripts habe ich Unterstützungsleistung von folgenden Personen erhalten:

1. Prof. Dr. Tilmann Betsch
2. PD Dr. Cornelia Betsch
3. Dr. Frank Renkewitz

Weitere Personen waren an der geistigen Herstellung der vorliegenden Arbeit nicht beteiligt. Insbesondere habe ich nicht die Hilfe einer Promotionsberaterin bzw. eines Promotionsberaters in Anspruch genommen. Dritte haben von mir weder unmittelbar noch mittelbar geldwerte Leistungen für Arbeiten erhalten, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen.

Die Arbeit oder Teile davon wurden bisher weder im Inland noch im Ausland in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde als Dissertation vorgelegt. Ferner erkläre ich, dass ich nicht bereits eine gleichartige Doktorprüfung an einer Hochschule endgültig nicht bestanden habe.

Zusammenfassung

Der Schwerpunkt dieser Dissertation liegt auf Fragen der Messung in der Erforschung von Wahrscheinlichkeits- und Risikourteilen. Ausgangspunkt hierfür waren eine Reihe von Studien zum sogenannten Einzelfalleffekt—einer systematischen Verzerrung von Urteilen über statistisch vermittelte Risiken durch persönliche Erfahrungsberichte—, an denen ich beteiligt war (Betsch, Renkewitz & Haase, 2013; Betsch, Ulshöfer, Renkewitz & Betsch, 2011). Der erste Artikel (Haase, Betsch & Renkewitz, 2015) ist ein Beispiel für diese Forschung: In dem experimentellen Paradigma erhalten die Teilnehmer sowohl eine statistische Information über die Auftretenswahrscheinlichkeit eines unerwünschten Ereignisses, als auch eine Reihe von kurzen Einzelfallerzählungen. Variiert wird die relative Häufigkeit der Berichte, die das Auftreten des unerwünschten Ereignisses schildern. Der Einzelfalleffekt bezeichnet den Befund, dass eine höhere Frequenz solcher Einzelfallberichte die wahrgenommene Auftretenswahrscheinlichkeit und das wahrgenommene Risiko des unerwünschten Ereignisses signifikant steigert. Normativ betrachtet sollten die Einzelfälle aufgrund der geringen Stichprobengröße im Vergleich zur Statistik jedoch keinen Einfluss auf das Urteilsverhalten haben.

In allen Studien zu diesem Effekt sind die subjektive Wahrscheinlichkeit und das wahrgenommene Risiko die zentralen abhängigen Variablen. In der Forschung herrscht jedoch nur wenig Einigkeit darüber, wie beide Konstrukte am besten zu messen sind. Daher beschäftigt sich der zweite Artikel (Betsch, Haase, Renkewitz & Schmid, 2015) unter anderem mit der Frage, ob der Einzelfalleffekt aufgabenabhängig ist, das heißt, ob er sich in Abhängigkeit vom verwendeten Messinstrument ändert. Artikel 3 und 4 (Haase & Betsch, 2016; Haase, Renkewitz & Betsch, 2013) sind methodisch ausgerichtet. In insgesamt vier Experimenten habe ich verschiedene Messformate zur Erfassung subjektiver Wahrscheinlichkeitsurteile unter streng kontrollierten Bedingungen evaluiert. Beide Artikel befassen sich mit psychometrischen Eigenschaften der untersuchten Messinstrumente, wie Sensitivität und Kontextabhängigkeit.

Artikel 1: Quellenglaubwürdigkeit und der verzerrende Einfluss von Einzelfall-informationen auf die Wahrnehmung von Impfrisiken

Während sich viele Bereiche für die Erforschung von Risikowahrnehmung eignen, sind Fragen des Gesundheitsverhaltens in jüngerer Zeit stärker in den Fokus psychologischer Forschung gerückt. Ein Grund hierfür liegt im Aufkommen der sogenannten partizipativen Entscheidungsfindung in der Medizin (Makoul & Clayman, 2006). Dies bedeutet, dass Patienten in zunehmendem Maße in medizinische und gesundheitsrelevante Entscheidungen

einbezogen werden. Dafür ist die adäquate Verarbeitung und Gewichtung probabilistischer Informationen eine wichtige Voraussetzung. Ein prominentes Beispiel hierfür ist die Impfentscheidung und die in den letzten Jahren wieder verstärkte Diskussion über das Risiko von Nebenwirkungen bei Impfungen (Dubé, Vivion & MacDonald, 2015). Darüber hinaus bezieht eine zunehmende Anzahl von Menschen gesundheitsrelevante Informationen aus dem Internet (Fox & Duggan, 2013), wo sich vermehrt impfkritische Internetseiten finden lassen, die von sogenannten Impfgegnern als Plattform genutzt werden. Ein häufiges Merkmal dieser Seiten sind individuelle Erfahrungsberichte von oder über Personen, die angeblich durch Impfstoffe geschädigt wurden (Guidry, Carlyle, Messner & Jin, 2015; Kata, 2010). Solche narrativen Informationen können in der Gesundheitskommunikation sehr effektiv sein (Hinyard & Kreuter, 2007; Winterbottom, Bekker, Conner & Mooney, 2008). Um Gesundheitsentscheidungen unter diesen Bedingungen experimentell zu erforschen, entwickelten Betsch et al. (2011) das oben beschriebene Forschungsparadigma.

In der vorliegenden Studie wurde getestet, ob die Glaubwürdigkeit der Informationsquellen den Effekt von Einzelfällen auf die Wahrnehmung von Impfrisiken moderiert. Die Teilnehmer sahen zunächst eine Statistik über die Auftretens-wahrscheinlichkeit von Nebenwirkungen (20%) nach einer Impfung gegen eine fiktive Krankheit. Im Anschluss lasen sie 20 personalisierte Erfahrungsberichte aus einem Online-Forum zu dieser Impfung. In einem between-subjects Versuchsplan wurden die relative Häufigkeit der Einzelfälle, die von Impfnebenwirkungen berichten (35% vs. 85%), die Glaubwürdigkeit der Quelle von Einzelfällen (impfkritische Internetseite vs. neutrales Gesundheitsforum) und die Glaubwürdigkeit der Statistik (verlässliche Daten vs. nicht verlässliche Daten vs. Kontrolle) variiert. Im Anschluss beurteilten die Teilnehmer das Risiko der Impfung, ihre Impfintention sowie die wahrgenommene Wahrscheinlichkeit und den wahrgenommenen Schweregrad der Nebenwirkungen. Die Ergebnisse zeigten einen stabilen Einzelfalleffekt auf allen abhängigen Variablen, der nicht von den Glaubwürdigkeits-manipulationen moderiert wurde. Allerdings führten Einzelfälle aus einem impfkritischen Forum zu einer generell niedrigeren Wahrnehmung von Impfrisiken. Zusätzliche Analysen zeigten, dass die beiden Risikokomponenten, Wahrscheinlichkeit und Schweregrad, nicht unabhängig voneinander wahrgenommen wurden.

Artikel 2: Was treibt den verzerrenden Einfluss von Einzelfällen auf die Risikowahrnehmung?

In zwei Experimenten wurden verschiedene methodische und prozedurale Faktoren untersucht, welche den Einzelfalleffekt beeinflussen oder erklären könnten. Dazu wurden unterschiedliche Messformate zur Erfassung des Effekts verglichen und getestet, ob die relative oder aber die absolute Anzahl von Einzelfällen, die das relevante Ereignis berichten, zu der Verzerrung der Urteile führt. Darüber hinaus wurde untersucht, ob die Einzelfälle Urteile in gleichem Maße erhöhen und reduzieren, und demzufolge eine symmetrische Verzerrung auftritt. Außerdem wurde die Bedeutung von Konversationsnormen für das Auftreten des Effekts erforscht.

Der Effekt der Einzelfälle war auf einem nichtnumerischen Risikomaß am stärksten, während zwei Skalen für subjektive Wahrscheinlichkeit hauptsächlich, jedoch in unterschiedlichem Maße, Manipulationen der statistischen Information abbildeten. Darüber hinaus erwies sich das Risikomaß als bester Prädiktor für Verhaltensintentionen. Carry-Over-Effekte zwischen den Instrumenten zeigten, dass Risiko- und Wahrscheinlichkeitsurteile ad hoc konstruiert werden und daher anfällig für Formulierungs- und Framing-Effekte sind. Risikowahrnehmung wurde durch die Einzelfälle in stärkerem Maße erhöht als vermindert, während der Einfluss auf Wahrscheinlichkeitsurteile symmetrisch war. Es fand sich keine Evidenz dafür, dass der Effekt durch Konversationsnormen zustande kommt. Die Manipulation der absoluten Anzahl von relevanten Einzelfällen bei gleichzeitiger Konstanthaltung ihrer relativen Häufigkeit hatte keinen Einfluss auf den Einzelfalleffekt.

Insgesamt unterstreichen die Ergebnisse die wichtige konzeptuelle Unterscheidung zwischen wahrgenommenem Risiko und wahrgenommener Wahrscheinlichkeit. Ferner ist es besonders bemerkenswert, dass der Einzelfalleffekt auf einer Repräsentation von relativer Häufigkeit, d.h. Wahrscheinlichkeit, beruht, jedoch eine stärkere Ausprägung auf einem breiten Risikomaß findet als auf einem Instrument für Wahrscheinlichkeitsurteile.

Artikel 3: Die Messung von subjektiven Wahrscheinlichkeitsurteilen: Evaluation der Sensitivität und Genauigkeit verschiedener Skalenformate

Subjektive Wahrscheinlichkeit ist eine zentrale Variable in vielen Studien zu Risikowahrnehmung—einschließlich unserer eigenen. Es existiert jedoch kein allgemein anerkanntes Maß für subjektive Wahrscheinlichkeitsurteile. Unter anderem wird über die optimale Anzahl von Antwortkategorien, d. h. die Auflösung der Skala, und darüber, ob man wahrgenommene Wahrscheinlichkeiten besser verbal oder numerisch messen sollte, diskutiert (z. B. Diefenbach, Weinstein & O'Reilly, 1993; Windschitl & Wells, 1996). Ferner bestimmen Evaluierungsstudien die Skalenleistung in der Regel in Bezug auf reale Daten (z. B. Eibner, Barth, Helmes & Bengel, 2006). Dieser Ansatz ermöglicht es jedoch nicht, zwischen tatsächlichen Verzerrungen in der Wahrnehmung von Wahrscheinlichkeit und vom Messinstrument produzierten Verzerrungen, zu unterscheiden.

In dieser Studie kontrollierten wir daher die objektiven, zu beurteilenden Wahrscheinlichkeiten und evaluierten fünf übliche Abfrageformate—eine verbale 7-stufige Ratingskala, eine verbal-numerische 11-stufige Rating Skale, eine visuelle Analogskala sowie Schätzungen relativer Häufigkeit und Prozenturteile—hinsichtlich ihrer Sensitivität und Genauigkeit. Variiert wurden dabei das Wahrscheinlichkeitsspektrum (niedrige vs. mittlere Wahrscheinlichkeiten), der Schweregrad der unsicheren Ereignisse (niedrig vs. hoch; beides between-subjects) sowie die Art der Enkodierung der objektiven Wahrscheinlichkeiten (sequentiell vs. aggregiert, within-subjects).

Der Schweregrad hatte keinen Einfluss auf die Wahrscheinlichkeitsurteile. Grundsätzlich waren die numerischen Maße den restlichen Formaten in allen Kriterien überlegen. Die Unterschiede zwischen den Instrumenten hingen jedoch von der Art der Darbietung objektiver Wahrscheinlichkeiten ab. Das aggregierte Format erlaubte eine fehlerfreie Enkodierung. Hier waren die Sensitivitätsunterschiede über beide Wahrscheinlichkeitsbereiche stabil. Die sequentielle Darbietung hingegen führte zu einer gewissen Ungenauigkeit, welche im mittleren Spektrum höher ausgeprägt war als bei niedrigen Wahrscheinlichkeiten. Während bei geringem Fehler die Unterschiede zwischen den Skalen mit den Unterschieden in der fehlerfreien Enkodierungsbedingung vergleichbar waren, glich sich die Skalensensitivität unter höherer Fehlerlast auf allgemein niedrigerem Niveau an.

Darüber hinaus fanden sich deutliche Kontexteffekte, d. h., unterschiedliche Antwortfunktionen bei den beiden Wahrscheinlichkeitsbereichen, auf den Ratingskalen und der Analogskala, wenn die enkodierten Wahrscheinlichkeiten ungenau waren. Bei fehlerfreier Enkodierung waren diese stark reduziert. Die Befunde zeigen, dass Unterschiede der Skalenleistung nicht nur von inhärenten Eigenschaften der Instrumente abhängen, sondern auch von der Genauigkeit der den Urteilen zugrundeliegenden Repräsentationen.

Artikel 4: Die Messung von subjektiven Wahrscheinlichkeitsurteilen: Fehler- und Ankereffekte

Dieser Artikel beschreibt eine Folgestudie zu Artikel 3. In drei Experimenten wurde das leistungsstärkste Format, die Prozentschätzungen, dem leistungsschwächsten Format, der verbalen 7-stufigen Ratingskale gegenübergestellt. Das Experimentaldesign unterschied sich in folgenden Punkten von dem vorhergegangenen: Objektive Wahrscheinlichkeiten wurden nur sequentiell enkodiert und die Darbietung wurde angepasst, um den Enkodierfehler zu erhöhen. Ferner wurde das Spektrum der objektiven Wahrscheinlichkeiten erweitert (10%–90%), bei der Enkodierung jedoch wieder in zwei Bereiche getrennt (niedrig vs. hoch). Die entscheidende Neuerung war, dass beide Bereiche einen gemeinsamen Stimulus (50%)

enthielten und within-subjects manipuliert wurden. Der Grund hierfür war die Annahme, dass die in Artikel 3 beschriebenen Kontexteffekte nicht auf die Stimulusverteilung, sondern auf die Interaktion von Enkodierfehler und Skalenformat zurückzuführen seien.

Experiment 1 zeigte, dass die hohe Auflösung des Prozentformats unter extremer Fehlerlast keinerlei Vorteile hinsichtlich der Urteilssensitivität bietet und im Aggregat sogar nachteilig sein kann. Weiterhin zeigte sich, dass fehlerbehaftete Repräsentationen zu scheinbaren Kontexteffekten auf der Ratingskala führen können. In Experiment 2 zeigte sich, dass ein Anker in den Stimuli die Urteile auf der Ratingskala in Relation zum Skalenmittelpunkt fixieren kann, dafür jedoch zu inkonsistenten Antwortfunktionen bei Prozentschätzungen führt, die durch eine starke Regression zur Mitte erklärt werden können. Experiment 3 demonstrierte schließlich, dass ungenaue Wahrscheinlichkeitsrepräsentationen zu einer Veränderung des Konstrukts führen, das auf der Ratingskala abbildet wird, während Prozenturteile konsistent bleiben. Die Befunde zeigen, dass eine verbale Ratingskala weder eine sinnvolle Quantifizierung noch einen sinnvollen Vergleich von Wahrscheinlichkeitsurteilen zulässt und daher für diesen Zweck nicht verwendet werden sollte. Prozentschätzungen lassen sich hingegen eindeutig interpretieren, sie erfassen jedoch auch die Ungenauigkeiten in den zugrundeliegenden Repräsentationen und können daher zu klassischen Regressionsfehlschlüssen führen.

Literaturverzeichnis

Betsch, C., Haase, N., Renkewitz, F. & Schmid, P. (2015). The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions? *Judgment and Decision Making, 10*(3), 241–264. Verfügbar unter http://journal.sjdm.org/14/141206a/jdm141206a.pdf

Betsch, C., Renkewitz, F. & Haase, N. (2013). Effect of narrative reports about vaccine adverse events and bias-awareness disclaimers on vaccine decisions: A simulation of an online patient social network. *Medical Decision Making, 33*(1), 14–25. https://doi.org/10.1177/0272989X12452342

Betsch, C., Ulshöfer, C., Renkewitz, F. & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making, 31*(5), 742–753. https://doi.org/10.1177/0272989X11400419

Diefenbach, M. A., Weinstein, N. D. & O’Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research, 8*(2), 181–192. https://doi.org/10.1093/her/8.2.181

Dubé, E., Vivion, M. & MacDonald, N. E. (2015). Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert Review of Vaccines, 14*(1), 99–117. https://doi.org/10.1586/14760584.2015.964212

Eibner, F., Barth, J., Helmes, A. & Bengel, J. (2006). Variations in subjective breast cancer risk estimations when using different measurements for assessing breast cancer risk perception. *Health, Risk & Society, 8*(2), 197–210. https://doi.org/10.1080/13698570600677407

Fox, S. & Duggan, M. (2013). *Health online 2013*. Verfügbar unter http://www.pewinternet.org/2013/01/ 15/health-online-2013

Guidry, J. P. D., Carlyle, K., Messner, M. & Jin, Y. (2015). On pins and needles: How vaccines are portrayed on Pinterest. *Vaccine, 33*(39), 5051–5056. https://doi.org/10.1016/j.vaccine.2015.08.064

Haase, N., Betsch, C. & Renkewitz, F. (2015). Source credibility and the biasing effect of narrative information on the perception of vaccination risks. *Journal of Health Communication, 20*(8), 920–929. https://doi.org/10.1080/10810730.2015.1018605

Haase, N. & Betsch, T. (2016). *Self-report measures of subjective probability: Error and anchor effects.* Unveröffentlichtes Manuskript.

Haase, N., Renkewitz, F. & Betsch, C. (2013). The measurement of subjective probability: Evaluating the sensitivity and accuracy of various scales. *Risk Analysis, 33*(10), 1812–1828. https://doi.org/10.1111/risa.12025

Hinyard, L. J. & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior, 34*(5), 777–792. https://doi.org/10.1177/1090198106291963

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine, 28*(7), 1709–1716. https://doi.org/10.1016/j.vaccine.2009.12.022

Makoul, G. & Clayman, M. L. (2006). An integrative model of shared decision making in medical encounters. *Patient Education and Counseling, 60*(3), 301–312. https://doi.org/10.1016/j.pec.2005.06.010

Windschitl, P. D. & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364. https://doi.org/10.1037//1076-898X.2.4.343

Winterbottom, A., Bekker, H. L., Conner, M. & Mooney, A. (2008). Does narrative information bias individual's decision making? A systematic review. *Social Science & Medicine, 67*(12), 2079–2088. https://doi.org/10.1016/j.socscimed.2008.09.037

Contents

Overview xiii

Research Highlights xiii

Scope xiv

Author Contributions xv

General Introduction 1

Article 1: Source Credibility and the Biasing Effect of Narrative Information on the Perception of Vaccination Risks 13

Article 2: The Narrative Bias Revisited: What Drives the Biasing Influence of Narrative Information on Risk Perceptions? 39

Article 3: The Measurement of Subjective Probability: Evaluating the Sensitivity and Accuracy of Various Scales 85

Article 4: Self-Report Measures of Subjective Probability: Error and Anchor Effects 121

General Discussion 189

Acknowledgments 197

Curriculum Vitae 198

Overview

The dissertation presents four separate articles on biases in probabilistic reasoning and on methodological issues in assessing such biases. Articles 1–3 have already been published in peer-reviewed academic journals. References and links to the definitive versions are provided on the respective chapter title pages.

Research Highlights

Article 1

- A small sample of single-case narratives biases risk perceptions even when statistical information is provided.
- Manipulating the credibility of the narrative or statistical information has no moderating effect on the bias.

Article 2

- Single-case narratives have different effects on perceived risk and perceptions of probability.
- The effect varies as a function of concomitantly assessed related constructs.
- The effect is driven by representations of likelihood.

Article 3

- Numeric scale formats for subjective probability show higher sensitivity and less context dependency than rating scales of a visual analog scale.
- The performance differences between scale formats differ themselves as a function of the noise in the underlying representations.

Article 4

- Noisy representations of subjective probability can create apparent context effects in within-subjects designs.
- The responsible mechanism differs between different scale formats.
- Noise changes the level of measurement for a low-resolution verbal scale.
- On a high-resolution numeric scale, noise can lead to classic regression fallacies.

Scope

Article	Experiment	Research Question/Subject Matter	Data Collection	Analyzed *N*
1	1	Does the biasing effect of single-case narratives on risk perception vary as a function of the credibility of statistical and narrative information?	Lab/ online	84 181
2	2	Is the biasing effect of single-case narratives on risk perception dependent on the measure used to assess it?	Lab	277
	3	Is the biasing effect of single-case narratives on risk perception a function of the relative or absolute frequency of narratives reporting the focal event?	online	464
3	4	Sensitivity and accuracy of five scale formats for subjective probability as a function of stimulus range, stimulus severity, and presentation mode of objective probabilities.	Lab	373
4	5	Domain specificity, sensitivity, and context dependency of two different scale formats for subjective probability when representations are highly imprecise.	Lab	87
	6	Anchor effects on two different scale formats for subjective probability when representations are highly imprecise.	Lab	105
	7	Anchor effects on two different scale formats for subjective probability as a function of error in representations.	Lab	92
			Total:	**1663**

Author Contributions

Article 1 Haase, N., Betsch, C., & Renkewitz, F. (2015). Source credibility and the biasing effect of narrative information on the perception of vaccination risks. *Journal of Health Communication, 20*(8), 920–929. https://doi.org/10.1080/10810730.2015.1018605

Conceived and designed the experiments: NH, CB, FR
Performed the experiments: NH
Analyzed the data: NH
Wrote the paper: NH
Approved the final draft: NH, CB, FR

Article 2 Betsch, C.*, Haase, N.*, Renkewitz, F., & Schmid, P. (2015). The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions? *Judgment and Decision Making, 10*(3), 241–264.***shared first authorship**

Conceived and designed the experiments: CB, NH, FR
Performed the experiments: NH
Analyzed the data: NH, CB
Wrote the paper: NH, CB
Approved the final draft: NH, CB, FR, PS

Article 3 Haase, N., Renkewitz, F., & Betsch, C. (2013). The measurement of subjective probability: Evaluating the sensitivity and accuracy of various scales. *Risk Analysis, 33*(10), 1812–1828. https://doi.org/10.1111/risa.12025

Conceived and designed the experiments: NH, FR, CB
Performed the experiments: NH
Analyzed the data: NH
Wrote the paper: NH
Approved the final draft: NH, FR, CB

Article 4 Haase, N., & Betsch, T. (2016). *Self-report measures of subjective probability: Error and anchor effects*. Manuscript in preparation.

Conceived and designed the experiments: NH, TB
Performed the experiments: NH
Analyzed the data: NH
Wrote the paper: NH
Approved the final draft: NH, TB

General Introduction

Risk is a fundamental part of life. From the oft-cited and somewhat trivial risk of getting caught in the rain without an umbrella to the more serious risk of losing money in an investment or even the existential risk of a terrorist attack, individuals are constantly faced with situations that demand an appraisal of a given risk and a choice based on this appraisal. Consequently, risk perception and decision making under uncertainty have been studied extensively by many different scientific disciplines. Especially the fields of psychology and economics have generated two research traditions that have been at odds with as well as influenced each other to varying degrees over the last 150 years (Bruni & Sugden, 2007; Loewenstein, 1992). The research presented in this dissertation addresses biases in risk perception and methodological issues in assessing such biases. However, the many different approaches to this topic have varied markedly in their interpretation and operationalization of the relevant constructs. Therefore, it is necessary that I briefly clarify my understanding before I describe the studies in more detail.

Interpretations of Risk

What is risk? The Merriam-Webster dictionary defines risk as “the possibility that something bad or unpleasant (such as an injury or a loss) will happen” (Risk, 2015). Though somewhat broad, this definition includes two distinct aspects of an event. First, a risk is a possibility of an event, that is, it is uncertain whether the event will occur or not. Second, the event itself is unpleasant, that is, a risk entails an evaluative notion. Thus, following this common understanding, it is then crucial to keep in mind that when individuals judge a risk they actually make at least two distinct, though not necessarily independent judgments: one regarding the probability of an outcome, another regarding its severity. This duality is analogous to the concept of expected value, or rather expected utility, which still forms the basis of decision theory. The different scientific disciplines, however, do not necessarily follow this interpretation of risk.

For instance, in the economic literature risk signifies a certain interpretation of probability. Knight (1921) proposed that the term risk should indicate that the probabilities attached to the possible outcomes of a choice are measurable, that is, they can either be derived a priori as in the throw of a die or they can be obtained statistically through the empirical analysis of the frequency of occurrence. In contrast, the term uncertainty should refer to situations where the likelihood of outcomes cannot be measured but must be estimated. This classification does not include the notion of an unwanted outcome but applies to any uncertain prospect. It does, however, mirror three different interpretations of probability—the classical, the frequentist, and the subjective—which I address shortly.

The psychological literature, on the other hand, is highly inconsistent in its understanding of what constitutes risk. While some argue that there is an implicit agreement that risk includes likelihood and severity (Yates & Stone, 1992), numerous publications refer to risk perception when in fact only judgments of probability were studied (e.g., Eibner, Barth, Helmes, & Bengel, 2006; Lapinski, Rimal, Klein, & Shulman, 2009; Schapira, Davids, McAuliffe, & Nattinger, 2004). Other approaches make it a point to expand the construct. Accordingly, risk judgments may include affective reactions (Loewenstein, Weber, Hsee, & Welch, 2001; Slovic, Finucane, Peters, & MacGregor, 2004), a moral appraisal of the source of a risk (Gardoni & Murphy, 2013), a sense of personal susceptibility to an outcome (Brewer et al., 2007), the voluntariness of, knowledge about, and control over a risk (Slovic, Fischhoff, & Lichtenstein, 1979; see Brun, 1994 and Yates & Stone, 1992 for discussions of this topic). Nonetheless, common and central to all definitions of risk is the concept of probability and much of the pertinent research discusses risk perceptions in relation to probability theory. The work in this dissertation is no exception. The biases in risk perception addressed in Articles 1 and 2 are essentially biases in probabilistic reasoning. However, determining whether this type of reasoning is biased depends on one's interpretation of probability.

Interpretations of Probability

The important distinction here is between the objectivist and the subjectivist view. Very broadly, the former holds that probabilities are intrinsic properties of events, that is, it assumes that there is a true probability of an event independent of the observer. In some cases, such as games of chance, the probability is simply defined as the ratio of the cases where the focal event can occur to the total number of possible outcomes. This is known as the classical interpretation of probability and closely associated with Blaise Pascal and Pierre de Fermat (Edwards, 1982; Ore, 1960). It is based on physical symmetry, that is, the equal chance of occurrence for all possible outcomes. In cases where the symmetry of outcomes is unknown, the probabilities can be derived from the relative frequency of occurrence in a number of repeated trials. The basis for this frequentist interpretation of probability is Jacob Bernoulli's law of large numbers (Hacking, 1971) which states that as the number of trials increases the relative frequency will converge to the true probability.

In contrast, according to the subjectivist view, probability does not exist independently of the human mind. Probability is understood as the degree of belief in the occurrence of an event. While classical or frequentist reasoning may be informative, what matters are a person's belief about the case in question. It follows that the only way to assess probability is to measure a person's behavior based on this belief, usually in the form of betting rates (de

Finetti, 1970). Note that Knight's (1921) definition of uncertainty differs from the subjectivist interpretation of probability insofar as it is based on ignorance not on the refusal of true probabilities.

Interpretations of Bias

Judgments and decisions are assumed to be biased, that is, normatively wrong, when they violate the tenets of probability theory. However, different interpretations of probability entail different normative standards and thus disparate ideas of what constitutes a bias (see Beach & Braun, 1994 for a discussion). An objective interpretation of probability allows for the assessment of accuracy, that is, the difference between objective and subjective probabilities. The research on overconfidence is a classic example of this approach (e.g., Lichtenstein, Fischhoff, & Phillips, 1982). In contrast, the subjective view's rejection of objective probabilities renders the question of accuracy meaningless. It does, however, recognize that different probability judgments must be coherent. Judgments are assumed to be biased if they are not internally consistent. The conjunction fallacy (Tversky & Kahneman, 1983) is a prominent example of this.

Understanding of Constructs in the Present Research

The research in this dissertation is grounded in the frequentist interpretation of probability. In the methodological studies (Articles 3 and 4) we evaluated the performance of different self-report measures for subjective probability and controlled the information that subjects based their judgments on. Objective probabilities were operationalized as relative frequencies in sequences of option-outcome pairs. The studies in Articles 1 and 2 provide a frame of content to the methodological research. The experiments in those sections addressed the biasing effect of concrete exemplars on risk perception. Subjects were provided with a statistical base-rate as well a small sample of single cases pertaining to a risky prospect. Variations in the relative frequency of single cases reporting the focal event had an effect on judgments of perceived probability and perceived risk. This kind of reasoning is considered to be biased because it fails to weight different information according to the sample size on which it is based (e.g., Tversky & Kahneman, 1971). Thus, this bias hinges on the assumption of the law of large numbers. But note that we did not measure this bias as variations in the accuracy of judgments but rather as the presence of any influence of the exemplars, which is more akin to an assessment of coherence. Furthermore, we had a broad understanding of the risk construct, that is, judgments of the perceived risk of an outcome (Articles 1 and 2) were assumed to include an assessment of its likelihood as well as of its severity and possibly any other number of aspects. The experiment in Article 1 touches upon the varying influence of different features of risk information on different aspects of the risk judgment. The experiments in Article 2 directly address the distinction between perceived probability and perceived risk.

Summary of the Research

The principal focus of this dissertation is on questions of measurement in research of risk perception. These questions were inspired by some studies of the above mentioned bias that I was involved in (Betsch, Renkewitz, & Haase, 2013; Betsch, Ulshöfer, Renkewitz, & Betsch, 2011). Article 1 (Haase, Betsch, & Renkewitz, 2015) is an example of this research and serves as a jumping-off point for the remainder of the dissertation. The main manipulation in this paradigm—next to the exploration of a number of possible moderating variables—is the relative frequency of single case narratives and thus probabilistic in nature. Consequently, subjective probability and perceived risk are central dependent variables in all studies. However, there is no agreement on how to assess either one. Therefore, in Article 2 (Betsch, Haase, Renkewitz, & Schmid, 2015) we studied, among other things, whether the bias is task-dependent, that is, whether it changes as a function of the instrument that is used to measure it. In Articles 3 and 4 (Haase & Betsch, 2016; Haase, Renkewitz, & Betsch, 2013) we took a more methodological approach to the question of measurement and evaluated the performance of different self-report formats for judgments of subjective probability under highly controlled conditions. Both articles mainly addressed psychometric properties such as the sensitivity and context dependency of the instruments.

Article 1: Source Credibility and the Biasing Effect of Narrative Information on the Perception of Vaccination Risks

While many areas offer themselves for the study of risk perception, the medical domain has become a recent focus of research because the advent of the shared decision making approach (Makoul & Clayman, 2006) means that patients are increasingly included in treatment and preventative decisions that involve collecting, processing, and weighting probabilistic information. Especially the debate about vaccinations and the risk of adverse events has flared up again over recent years (Dubé, Vivion, & MacDonald, 2015). Additionally, an increasing number of people seek health information on the Internet (Fox & Duggan, 2013) and a common strategy of anti-vaccination proponents is the online dissemination of testimonials from or about individuals who have allegedly been harmed by vaccines (Guidry, Carlyle, Messner, & Jin, 2015; Kata, 2010). Such narrative information can be very persuasive in health communication (Hinyard & Kreuter, 2007; Winterbottom, Bekker, Conner, & Mooney, 2008). Thus, Betsch et al. (2011) created the above described research paradigm with a focus on vaccine adverse events (VAE). In this study we additionally tested whether the credibility of the information sources moderates the biasing effect of narrative information regarding the perception of vaccination risks. 265 participants were provided with statistical information (20%) regarding the occurrence of VAE after vaccination against a fictitious disease. This was followed by 20 personalized narratives from an online forum on vaccination experiences. We varied the relative frequency of narratives reporting vaccine adverse events (35% vs. 85%), narrative source credibility (anti-vaccination website vs. neutral health forum), and the credibility of the statistical information (reliable data vs. unreliable data vs. control) in a between-subjects design. We assessed risk perceptions and vaccination intentions as well as the perceived probability and the perceived severity of VAE. Results showed a stable narrative bias on all dependent variables that was not affected by credibility cues. However, narratives from an anti-vaccination website led to generally lower perceptions of vaccination risks. Additional analyses revealed that the two assumed constituents of risk (i.e., probability and severity) were not perceived independently.

Article 2: The Narrative Bias Revisited – What Drives the Biasing Influence of Narrative Information on Risk Perceptions?

In this work we investigated various methodological and procedural factors that may influence the biasing effect of single-case narratives. We compared different measures to assess the bias. We further investigated whether the absolute or the relative number of narratives reporting the focal event drives the bias. Additionally, we examined whether narratives increase and decrease judgments to the same degree, that is, whether the bias is symmetric. Finally, we explored the impact of conversational norms on the occurrence of the bias. We found that narratives had the strongest effect on a non-numerical risk measure, whereas two scales for subjective probability reflected primarily statistical variations, though to different degrees. Further, the risk measure was the best predictor of behavioral intentions. Moreover, two-way carry-over effects between the measures indicated that judgments of perceived risk and subjective probability are ad hoc constructions and thus susceptible to wording and framing effects. We observed a negativity bias on the risk measure, that is, the narratives rather increased than decreased risk perceptions while the effect on probability judgments was symmetric. Additionally, we found no evidence that the narrative bias is solely produced by adherence to conversational norms. Finally, changing the absolute number of narratives reporting the focal event, while keeping their relative frequency constant, had no effect. Thus, individuals extracted a representation of likelihood from a sample of single-case narratives, which drove the bias. The results not only underline the important conceptual distinction between subjective probability and perceived risk as people use the same information differently when asked for a judgment of one or the other. They also show that the relation between representations of subjective probability and perceived risk is not yet fully understood.

Article 3: The Measurement of Subjective Probability – Evaluating the Sensitivity and Accuracy of Various Scales

Subjective probability is a central variable in many studies of risk perception (including our own). However, there is no standard measure for subjective probability estimates. Two disputed points are the optimal number of response categories, i.e., the scale's resolution (e.g., Diefenbach, Weinstein, & O'Reilly, 1993) and whether one should assess perceptions of probability verbally or numerically (e.g., Windschitl & Wells, 1996). Additionally, many evaluation studies compare scale performance in relation to real-life data (e.g., Eibner et al., 2006). However, this approach does not allow differentiating measurement bias, i.e., the actual functioning of the scale, from biases in representations. Thus, in this study, we compared five commonly used measurement formats—a verbally labeled 7-point rating scale, a verbally anchored and numerically labeled 11-point rating scale, a visual analog scale, estimates of relative frequency, and of percent—in terms of their ability to assess subjective probability when objective probabilities have been provided. We varied two context variables: the range of objective probabilities (low vs. moderate) and the severity (low vs. high) of the events to be judged (both between-subjects), as well as the presentation mode of objective probabilities (sequential presentation of singular events vs. graphical presentation of aggregated information, within-subjects). We assessed scale sensitivity, scale accuracy, and consistent scale use across different contexts. The severity of events had no effect. The numeric formats generally outperformed all other measures. However, differences depended on how the objective probabilities were encoded. Pictographs ensured perfect information and the differences between scales were stable across contexts. In contrast, sequential encoding introduced sampling error into the probability representations to varying degrees. For low range probabilities the error was lower than for moderate probabilities because spotting a few highly salient events in a sequence is easier than tracking many. When the error was low, the differences between scale formats were similar to the graphical condition. When error was higher, on the other hand, the differences in performance were markedly decreased. Moreover, when representations were error-free, all scale formats remained reasonably consistent across probability ranges while sampling error resulted in very different judgment functions on the rating scales and the analog measure but not on the numeric formats. We concluded that differences in performance between scales are caused only in part by characteristics of the scales themselves—they also depend on the error in the underlying representations.

Article 4: Self-Report Measures of Subjective Probability – Error and Anchor Effects

In this study we directly followed up and expanded on the findings reported in the second article. In three experiments we compared the worst and the best performing scale formats—the verbally labeled 7-point rating scale and the percent format. Subjects encoded two ranges of objective probabilities with one common stimulus as sequences of option-outcome pairs. We increased the sampling error and broadened the range of objective probabilities to cover almost the entire continuum. Crucially, we presented the two judgment ranges within-subjects as we reasoned that the previously observed context effects were a function of the encoding error and the scale format rather than of the distribution of stimuli. In Experiment 1 we observed that under highly error-prone conditions the high resolution of the percent format offers no advantage in terms of sensitivity. Additionally, imprecise representations, rather than the stimulus distribution, can result in apparent context effects on the rating scale. In Experiment 2 we found that an anchor in the stimuli can tether judgments on the rating scale to the scale's midpoint but in turn result in inconsistent judgment functions on the percent format that could be misinterpreted as context effects, though they are an expression of regression toward the mean. In Experiment 3 we discovered that imprecise representations change the way the rating scale is used while percent estimates remain consistent. We concluded that a verbal rating scale does not allow for a meaningful quantification or meaningful comparisons between experimental conditions and should not be used in research on subjective probability. The percent format captures the underlying representations reliably and consistently but is very sensitive to noise and can lead to classic regression fallacies.

Overview

The articles in this dissertation are presented in a conceptual rather than a chronological order—from an applied context to a purely methodological approach. Although the research is closely connected, each article is self-contained and includes a discussion of the respective findings. Therefore, I close my dissertation with a brief overarching discussion of the conclusions drawn from the individual manuscripts and the theoretical implications they entail.

References

Beach, L., & Braun, G. (1994). Laboratory studies of subjective probability: A status report. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 107–127). Chichester, England: Wiley.

Betsch, C., Haase, N., Renkewitz, F., & Schmid, P. (2015). The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions? *Judgment and Decision Making, 10*(3), 241–264. Retrieved from http://journal.sjdm.org/14/141206a/jdm141206a.pdf

Betsch, C., Renkewitz, F., & Haase, N. (2013). Effect of narrative reports about vaccine adverse events and bias-awareness disclaimers on vaccine decisions: A simulation of an online patient social network. *Medical Decision Making, 33*(1), 14–25. https://doi.org/10.1177/0272989X12452342

Betsch, C., Ulshöfer, C., Renkewitz, F., & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making, 31*(5), 742–753. https://doi.org/10.1177/0272989X11400419

Brewer, N. T., Chapman, G. B., Gibbons, F. X., Gerrard, M., McCaul, K. D., & Weinstein, N. D. (2007). Meta-analysis of the relationship between risk perception and health behavior: The example of vaccination. *Health Psychology, 26*(2), 136–145. https://doi.org/10.1037/0278-6133.26.2.136

Brun, W. (1994). Risk perception: Main issues, approaches and findings. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 295–320). Chichester, England: Wiley.

Bruni, L., & Sugden, R. (2007). The road not taken: How psychology was removed from economics, and how it might be brought back. *The Economic Journal, 117*(516), 146–173. https://doi.org/10.1111/j.1468-0297.2007.02005.x

de Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica, 34*, 129–145. https://doi.org/10.1016/0001-6918(70)90012-0

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research, 8*(2), 181–192. https://doi.org/10.1093/her/8.2.181

Dubé, E., Vivion, M., & MacDonald, N. E. (2015). Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. *Expert Review of Vaccines, 14*(1), 99–117. https://doi.org/10.1586/14760584.2015.964212

Edwards, A. W. F. (1982). Pascal and the problem of points. *International Statistical Review, 50*(3), 259–266. https://doi.org/10.2307/1402496

Eibner, F., Barth, J., Helmes, A., & Bengel, J. (2006). Variations in subjective breast cancer risk estimations when using different measurements for assessing breast cancer risk perception. *Health, Risk & Society, 8*(2), 197–210. https://doi.org/10.1080/13698570600677407

Fox, S., & Duggan, M. (2013). *Health online 2013*. Retrieved from PEW Internet & American Life Project website: http://www.pewinternet.org/2013/01/ 15/health-online-2013

Gardoni, P., & Murphy, C. (2013). A scale of risk. *Risk Analysis, 34*(7), 1208–1227. https://doi.org/10.1111/risa.12150

Guidry, J. P. D., Carlyle, K., Messner, M., & Jin, Y. (2015). On pins and needles: How vaccines are portrayed on Pinterest. *Vaccine, 33*(39), 5051–5056. https://doi.org/10.1016/j.vaccine.2015.08.064

Haase, N., Betsch, C., & Renkewitz, F. (2015). Source credibility and the biasing effect of narrative information on the perception of vaccination risks. *Journal of Health Communication, 20*(8), 920–929. https://doi.org/10.1080/10810730.2015.1018605

Haase, N., & Betsch, T. (2016). *Self-report measures of subjective probability: Error and anchor effects.* Manuscript in preparation.

Haase, N., Renkewitz, F., & Betsch, C. (2013). The measurement of subjective probability: Evaluating the sensitivity and accuracy of various scales. *Risk Analysis, 33*(10), 1812–1828. https://doi.org/10.1111/risa.12025

Hacking, I. (1971). Jacques Bernoulli's art of conjecturing. *The British Journal for the Philosophy of Science, 22*(3), 209–229. Retrieved from https://www.jstor.org/stable/686744

Hinyard, L. J., & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior, 34*(5), 777–792. https://doi.org/10.1177/1090198106291963

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine, 28*(7), 1709–1716. https://doi.org/10.1016/j.vaccine.2009.12.022

Knight, F. (1921). *Risk, uncertainty and profit*. Boston, MA: Houghton Mifflin Company.

Lapinski, M. K., Rimal, R. N., Klein, K. A., & Shulman, H. C. (2009). Risk perceptions of people living with HIV/AIDS: How similarity affects optimistic bias. *Journal of Health Psychology, 14*(2), 251–257. https://doi.org/10.1177/1359105308100209

Lichtenstein, S., Fischhoff, B., & Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgement under uncertainty: Heuristics and biases* (pp. 306–334). Cambridge, England: University Press. https://doi.org/10.1017/CBO9780511809477.023

Loewenstein, G. (1992). The fall and rise of psychological explanations in the economics of intertemporal choice. In G. Loewenstein & J. Elster (Eds.), *Choice over time* (pp. 3–34). New York, NY: Russell Sage Foundation. Retrieved from http://www.jstor.org/stable/10.7758/9781610443654.5

Loewenstein, G., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127*(2), 267–286. https://doi.org/10.1037//0033-2909.127.2.267

Makoul, G., & Clayman, M. L. (2006). An integrative model of shared decision making in medical encounters. *Patient Education and Counseling, 60*(3), 301–312. https://doi.org/10.1016/j.pec.2005.06.010

Ore, O. (1960). Pascal and the invention of probability theory. *The American Mathematical Monthly, 67*(5), 409–419. https://doi.org/10.2307/2309286

Risk. (2015). In *Merriam-Webster.com*. Retrieved September 26 2016 from http://www.merriam-webster.com/dictionary/risk

Schapira, M., Davids, S., McAuliffe, T. L., & Nattinger, A. B. (2004). Agreement between scales in the measurement of breast cancer risk perceptions. *Risk Analysis, 24*(3), 665–673. https://doi.org/10.1111/j.0272-4332.2004.00466.x

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis, 24*(2), 311–322. https://doi.org/10.1111/j.0272-4332.2004.00433.x

Slovic, P., Fischhoff, B., & Lichtenstein, S. (1979). Rating the risks. *Environment, 21*(3), 14-39. https://doi.org/10.1080/00139157.1979.9933091

Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin, 76*(2), 105–110. https://doi.org/10.1037/h0031322

Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review, 90*(4), 293–315.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364. https://doi.org/10.1037//1076-898X.2.4.343

Winterbottom, A., Bekker, H. L., Conner, M., & Mooney, A. (2008). Does narrative information bias individual's decision making? A systematic review. *Social Science & Medicine, 67*(12), 2079–2088. https://doi.org/10.1016/j.socscimed.2008.09.037

Yates, J. F., & Stone, E. R. (1992). The risk construct. In J. F. Yates (Ed.), *Risk-taking behavior*. (pp. 1–25). Oxford, England: Wiley.

Article 1

Source Credibility and the Biasing Effect of Narrative Information on the Perception of Vaccination Risks

Reference:

Haase, N., Betsch, C., & Renkewitz, F. (2015). Source credibility and the biasing effect of narrative information on the perception of vaccination risks. *Journal of Health Communication, 20*(8), 920–929. https://doi.org/10.1080/10810730.2015.1018605

The definitive version is available at:

https://www.tandfonline.com/doi/abs/10.1080/10810730.2015.1018605

Source Credibility and the Biasing Effect of Narrative Information on the Perception of Vaccination Risks

Niels Haase, Cornelia Betsch, and Frank Renkewitz
University of Erfurt, Germany

Author Note

Niels Haase, Department of Psychology, University of Erfurt, Germany; Cornelia Betsch, Center of Empirical Research in Economics and Behavioral Sciences (CEREB), University of Erfurt, Germany; Frank Renkewitz, Department of Psychology, University of Erfurt, Germany

This research was financed by a research grant from the German Research Foundation (BE 3970/4-2) to the second and third authors. The authors are grateful to Karl-Philipp Henschelmann and Philipp Schmid for their help in conducting the study as well as Heather Fuchs for helpful comments on a previous draft of this article. The authors also gratefully acknowledge http://www.sparbaby.de, http://www.onmeda.de, and http://estudy.zpid.de/ for posting a link to the study on their websites.

Correspondence concerning this article should be addressed to Niels Haase, Department of Psychology, University of Erfurt, Nordhaeuser Strasse 63, 99089 Erfurt, Germany. E-mail: niels.haase@uni-erfurt.de

Abstract

Immunization rates are below the Global Immunization Vision and Strategy established by the World Health Organization. One reason for this are anti-vaccination activists, who use the Internet to disseminate their agenda, frequently by publishing narrative reports about alleged vaccine adverse events. In health communication, the use of narrative information has been shown to be effectively persuasive. Furthermore, persuasion research indicates that the credibility of an information source may serve as a cue to discount or augment the communicated message. Thus, the present study investigated the effect of source credibility on the biasing effect of narrative information regarding the perception of vaccination risks. 265 participants were provided with statistical information (20%) regarding the occurrence of vaccine adverse events after vaccination against a fictitious disease. This was followed by 20 personalized narratives from an online forum on vaccination experiences. The authors varied the relative frequency of narratives reporting vaccine adverse events (35% vs. 85%), narrative source credibility (anti-vaccination website vs. neutral health forum), and the credibility of the statistical information (reliable data vs. unreliable data vs. control) in a between-subjects design. Results showed a stable narrative bias on risk perception that was not affected by credibility cues. However, narratives from an anti-vaccination website generally led to lower perceptions of vaccination risks.

Keywords: Persuasion, testimonials, immunization, discounting

Source Credibility and the Biasing Effect of Narrative Information on the Perception of Vaccination Risks

Immunization has been one of the most successful endeavors of modern medicine, eradicating smallpox and averting an estimated 2.5 million deaths per year (World Health Organization, 2013). Still, global vaccination uptake remains below the goals set forth in the Global Immunization Vision and Strategy launched by the World Health Organization in 2006 (Brown et al., 2011). While this shortcoming is due to many reasons including a lack of resources in less developed countries, suboptimal immunization rates also occur in parts of the world where the availability and cost of vaccines are no hindrance. This is exemplified by recent and recurring outbreaks of measles in various member-states of the European Union (Burki, 2013). In these countries, vaccinations have, in a way, become a victim of their own success. Because many vaccine-preventable diseases occur only rarely, the risks associated with these diseases remain invisible to most individuals (Omer, Orenstein, & Koplan, 2013). As perceived risk has been shown to be a reliable predictor of vaccination behavior (Brewer et al., 2007), this lacking awareness of risky diseases may contribute to the low vaccination rates. In addition, many individuals focus on the risk of vaccine adverse events—first, because the absolute frequency of vaccine adverse events (VAE) increases with the number of individuals receiving vaccines (Chen, 1999) and, second, because anti-vaccination activists propagate severe alleged VAE, for example, the refuted myth that the MMR-vaccination may lead to autism (The Lancet, 2010; Wakefield et al., 1998). A typical strategy of anti-vaccination activists is to promote emotional narratives from and about individuals, preferably children, who have allegedly been harmed by vaccinations (Kata, 2010, 2012). Recent web 2.0 technology (O'Reily, 2005) offers the ideal platform for this approach (Betsch et al., 2012).

In health communication, the use of narrative information has also been shown to be effective in persuasion (Hinyard & Kreuter, 2007). With regard to likelihood information, in particular, the fact that concrete examples have a greater effect on decisions than abstract base-rate information has long been observed (Borgida & Nisbett, 1977; Brosius & Bathelt, 1994; de Wit, Das, & Vet, 2008). This so-called *narrative bias* has also been demonstrated within the specific context of perceived vaccination risks: even when participants are provided with base-rate information regarding the occurrence of VAE, normatively irrelevant narrative evidence increases participants' perceptions regarding the probability and risk of VAE (Betsch, Renkewitz, & Haase, 2013; Betsch, Ulshöfer, Renkewitz, & Betsch, 2011).

Theoretical accounts of the narrative bias focus on different aspects of narrative communication. Some argue that the mere frequency with which one encounters narrative information determines its effect, irrespective of the sample size to which the information pertains (Obrecht, Chapman, & Gelman, 2009). This approach relates to the probability dimension of risk. In general, risk is understood as a combination of an event's probability of occurrence and its severity (Yates & Stone, 1992). Other accounts focus on the quality of narratives such as the sense of transportation, that is, immersion in the story, that a narrative evokes in the reader (Green & Brock, 2000, 2002). This aspect may also affect the perceived severity of the event.

While the narrative bias has been demonstrated frequently, little is known regarding the importance of the sources that communicate the narrative and statistical information. Persuasion research has shown that the credibility of an information source may serve as a cue to discount or augment the communicated message (Hovland & Weiss, 1951; Pornpitakpan, 2004). That is, irrespective of the amount of information learned from a source, the persuasive impact of the provided information may be significantly smaller if the source is perceived to be low in credibility (discounting); at the same time, the persuasive impact may be significantly greater if the source is perceived to be highly credible (augmenting). Thus, we assessed the potential moderating effect of statistic and narrative source credibility on the narrative bias.

Overview

In the following experiment, we presented participants with statistical information regarding the occurrence of VAE. The statistical information was combined with either a discounting or an augmenting cue in the experimental conditions. No extra information was provided in the control condition. Participants then read 20 personalized narrative reports, framed as contributions in an online forum, either on an anti-vaccination website (discounting cue) or an online health forum (control). The relative frequency of positive narratives (i.e., narratives reporting VAE) was varied between subjects to test for the narrative bias.

Hypotheses

On the basis of previous research regarding the biasing influence of narrative information, we hypothesize the following:

H1: A higher relative frequency of narratives reporting VAE will increase the perceived risk of vaccination (H1a) and perceived probability of VAE (H1b, narrative bias).

As the credibility of the source of information can lead to augmenting or discounting the communicated information, we expect that the narrative bias will depend on the credibility of the sources of the statistical and narrative information:

H2: The narrative bias on risk and probability perception will be moderated by the credibility of the statistical information. The narrative bias will be strongest when the statistical information is discredited and weakest when it is praised.

H3: The narrative bias on risk and probability perception will be moderated by the credibility of the narrative information. The narrative bias will be stronger when the narratives are provided by a neutral source as opposed to an interest group.

Perceived risk is an important predictor of preventive health behavior. Based on previous research (Betsch et al., 2013), we therefore posit the following:

H4: A higher relative frequency of narratives reporting VAE will decrease the intention to get vaccinated (H4a). This effect will be mediated by the perceived risk of VAE (H4b).

In addition to probability, we also explored the manipulations' effects on the perceived severity of VAE, as perceived risk is constituted by both variables (Yates & Stone, 1992).

Method

Participants and Design

The study was conducted both in the laboratory and as an online survey. Participants in the laboratory were students at a German university between the ages of 19 and 36 years who took part in exchange for either a payment of 3 € (ca. US$4) or course credit. Online participants (18–47 years) were recruited through a number of websites (Facebook, websites related to online research, health, and childhood topics). As an incentive, we raffled off ten 20 € (ca. US$27) vouchers for a large Internet store. Participants were randomly distributed to one of the 12 cells in the 3 (low credibility vs. high credibility of statistical information vs. no credibility information) × 2 (low credibility vs. high credibility of narrative information) × 2 (35% vs. 85% relative frequency of narratives reporting VAE) between-subjects design.

Procedure

Participants were presented with information regarding a fictitious disease called dysomeria, a serious and highly contagious disease with symptoms including vomiting and fever as well as meningitis and, in some cases, permanent paralysis. Participants were informed that a vaccination against dysomeria exists that effectively protects against infection and is highly recommended by the STIKO (National Immunization Technical Advisory Group, German equivalent to the Advisory Committee on Immunization Practices).

Participants were also told that a study had observed VAE such as fever, rash, restlessness, and dizziness in 20% of all vaccination cases. Depending on the condition, participants received a discounting or augmenting cue or no information regarding the credibility of the statistics. Subsequently, participants read 20 short narratives from an online forum, which described personal experiences with the dysomeria vaccination. The forum's source varied in credibility between subjects. Participants then proceeded to the dependent and control variables and the manipulation checks.

Statistical Base-Rate Information

The base-rate information (20% VAE) was also provided in the form of an icon array, that is, a 10 × 10 matrix of 100 rectangles colored either blue (VAE) or gray (no VAE, generated using http://www.iconarray.com). This type of display has been shown to reduce the effect of anecdotal information (Fagerlin, Wang, & Ubel, 2005) and therefore yields a strong test for the effect of narratives on risk judgments. The VAE (fever, rash, restlessness, dizziness) were selected from 66 medical conditions that were pretested for severity on a 7-point rating scale. Their mean severity scores ranged between 1.97 (SD = 1.20) and 2.77 (SD = 1.31) and did not differ significantly from values of either 2 or 3 ($ts \leq 1.60$).

Credibility of Statistical Information

Below the icon array, a short evaluation of the research methods employed in the study was presented. The text stated that two national scientific bodies had described the scientific methods and data analysis used in the study to be either poor (discounting) or exemplary (augmenting), resulting in unreliable or reliable data, respectively (see the Appendix for the full texts). In the control condition, no additional information pertaining to the statistics was provided.

Credibility of Narrative Information

We manipulated narrative credibility by stating that they had been posted on either a neutral online health forum called *gesundheit-net.de* (health-net.de) or an anti-vaccination website called *impfen-schadet.de* (vaccination-harms.de). First, participants were presented with an introductory text explaining that, in the age of web 2.0, it had become common for individuals to share their experiences with others in online forums (e.g., regarding vaccinations). To increase the salience of the narratives' source all participants then read short texts describing both websites in counterbalanced order. The description of the anti-vaccination website stated that, to further their agenda, the website's goal was to collect as many cases of harm through vaccination as possible (see the Appendix for the full texts). Subsequently, participants proceeded to the actual narratives, which were the same in both conditions. A mock online banner with the name of the forum was constantly visible at the top of the page displaying the narrative.

Relative Frequency of VAE

The online forum presented a sequence of 20 narratives in a randomized order, one on each page, describing personal experiences with the dysomeria vaccine. The narratives reported either the occurrence or nonoccurrence of VAE. To test for the narrative bias, the relative frequency of positive VAE implied in the forum was higher than in the statistical information (20%). In the 35% condition, 7 out of 20 narratives reported VAE compared to 17 such cases in the 85% condition (see the Appendix for sample narratives). All narratives had a length of 53 words and were balanced for author gender. The VAE mentioned in the narratives corresponded to the ones provided in the statistical information.

Measures

Dependent variables. Participants judged the riskiness of the vaccination on a scroll bar ranging from *not risky at all* (score = 0) to *very risky* (= 100). Numeric scale anchors were not provided. Probability of VAE was assessed through a percentage statement. Participants then rated on 7-point rating scales the severity of VAE (1 = *not severe*, 7 = *very severe*) and their intention to get vaccinated against dysomeria if they had the chance to do so in the following week (1 = *definitely not vaccinate*, 7 = *definitely vaccinate*).

Control variables. Because attitude is a strong predictor for vaccination behavior (Glasman & Albarracín, 2006), participants were asked to rate their general attitude toward vaccination on a 7-point rating scale (1 = *fully against vaccination*, 7 = *fully in favor of vaccination*).

In addition, participants rated the extent to which they weighted narrative vs. statistical information in their judgments on a scroll-bar ranging from *judgment based fully on statistics* (= 0) to *based fully on narratives* (= 100). The scale was anchored at the midpoint (= 50) in the beginning to indicate equal weighting of both sources. Again, no numeric anchors were provided.

To assess whether numeracy affects risk judgments (Reyna, Nelson, Han, & Dieckmann, 2009) participants completed the Berlin Numeracy Test for the general population (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Schwartz, Woloshin, Black, & Welch, 1997)—a 7-item instrument consisting of tests of probabilistic inference abilities, for example, “Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?”

Because of its length, the numeracy test was administered at the very end of the study, that is, after manipulation checks were completed.

Manipulation checks. Participants rated four items pertaining to the credibility of either information source on 7-point rating scales. They judged the credibility (1 = not credible at all, 7 = absolutely credible) and representativeness of the statistics and narratives when considering all vaccinations that are administered in Germany within a year (1 = not representative at all, 7 = very representative) as well as the expertise (1 = very bad, 7 = very good) and trustworthiness (1 = not at all trustworthy, 7 = absolutely trustworthy) of the study and the narratives' authors.

To ensure that all information was properly encoded, participants were also asked to recall the statistically conveyed base-rate of VAE, the absolute number of narratives reporting VAE, and the name of the online forum in which the narratives were published.

Results

Participants

102 individuals took part in the web-administered lab experiment. 18 stated that they had participated in a similar study previously and were thus excluded, resulting in a sample of 84 individuals, 67 (79.8%) of which were female. Because of the high number of participants reporting previous participation, we decided to make the study available for participants on the Internet.

After being published online, 390 individuals clicked the link to the study and 244 finished the survey, 45 of which stated that they had previously participated in a similar study and were therefore excluded. 18 participants who completed the survey in less than the mean time minus one standard deviation (4 minutes 42 seconds) were also excluded, resulting in a sample of 181 participants, 163 (90.1%) of which were female.[1]

The laboratory and online samples were combined for analyses, resulting in a final sample of 265 participants, 230 (86.8%) of which were female. The mean (SD) age was 28.42 (6.11) years. 33 participants (12.6%) did not complete the German university entrance exam (Abitur), 127 (47.9%) completed the Abitur, 92 (34.7%) completed some higher education,

[1] The time required to complete the study was calculated as the time between having read the initial instructions and having answered the last of the manipulation checks (recall of forum name). The numeracy test was excluded from this calculation, as the seven items took approximately half as long to complete as the entire rest of the survey, with a very large standard deviation, that is, 6 min 9 s (4 min 20 s).

and 10 (3.8%) completed a PhD. Mean time (SD) to complete the survey for the combined sample was 9 min 41 s (3 minutes 34 seconds).[2]

Manipulation Checks

Encoding variables. The majority of participants ($n = 232$; 87.6%) correctly recalled the probability of VAE (20%) as presented in the icon array (mode = 20; $M = 20.70$, $SD = 9.45$). Likewise, the majority of the participants correctly or approximately remembered the number of narratives reporting VAE. In the 35% condition, 72.5% of participants reported remembering 5, 6, 7, or 8 positive narratives (correct answer = 7) with a mode of 6 ($M = 7.11$, $SD = 3.71$). In the 85% condition, 84.3% recalled 15–18 narratives (correct answer = 17) with a mode of 15 ($M = 15.70$, $SD = 2.22$).

Participants were also asked to recall the name of the online forum. Recalled names were coded as correct if they included the term *health* in the health forum condition and if they clearly indicated an anti-vaccination stance in the anti-vaccination website condition. In the anti-vaccination condition 117 participants (86.7%) correctly remembered the name of the online forum, 111 (85.4%) individuals correctly remembered the online health forum.

Source credibility. The respective four items pertaining to source credibility were all moderately to strongly positively correlated (statistics: rs = .35–.61, ps < .001; narratives: rs = .36–.73, ps < .001). Therefore, we averaged the items to form a single credibility score for each information source (statistics: Cronbach's $\alpha = .76$; narratives: Cronbach's $\alpha = .76$).

Compared with the control condition without additional information ($M = 4.99$, $SD = 1.00$), participants perceived the statistical information to be less credible when it was described as unreliable ($M = 4.40$, $SD = 1.14$). Praising its reliability ($M = 5.15$, $SD = 0.98$) did not increase credibility, $F(2, 262) = 12.22$, $p < .001$, $\eta^2 = .09$, Bonferroni's test ps $\leq .001$. These results indicate that the discounting cue lowered source credibility for the statistical information, whereas the augmenting cue had no effect.

In contrast, the source of the narrative information had neither an effect on the credibility index (health forum: $M = 3.52$, $SD = 1.12$; anti-vaccination: $M = 3.47$, $SD = 1.17$;

[2] In the laboratory sample, the mean (SD) age was 22.69 (3.77) years. 81 participants (96.4%) reported an Abitur grade with a sample mean (SD) of 2.15 (0.44). Grades on this exam range between 1.0 and 4.0, with 1.0 being the best possible grade. The mean time (SD) to complete the study was 10 min 24 s (2 min 16 s). In the online sample, the mean (SD) age was 31.07 (5.09) years. Thirty-three participants (18.2%) did not complete the Abitur, 46 (25.4%) completed the Abitur, 92 (50.8%) received a higher education degree, and 10 (5.5%) received a PhD. The mean time (SD) to complete the study was 8 min 51 s (4 min 9 s).

F < 1) nor on any of the four single items, questioning if this manipulation worked as anticipated.

To explore whether the different sources affected the use of the information differently, we compared the self-reported weighting of the sources between the conditions. We found no significant differences in self-reported source weighting between the three sources of statistical information, $F(2, 259) = 1.09$, $p = .338$, and a small effect of narrative source (health forum: $M = 48.33$, $SD = 29.29$; anti-vaccination: $M = 40.73$, $SD = 28.50$), $F(1, 259) = 3.84$, $p = .051$, $\eta_p^2 = .02$.

Hypothesis Testing

Regression analyses were used for all analyses. We created two dummy variables to assess the effect of the statistical information credibility; the condition without any additional information served as the reference or control group. Thus, we compare praising the statistical information to providing no information. Correspondingly, we compare criticizing its credibility to providing no information. The resulting variables are "Credible statistical information" (coded as 1 with the remaining two conditions coded as 0) and correspondingly "Not credible statistical information" (= 1, rest = 0). Likewise, narrative source was coded as 0 = anti-vaccination website and 1 = online health forum. All predictors were standardized (Cohen, Cohen, West, & Aiken, 2003). Interaction terms are mathematical products of the respective predictors. In a first step, we predicted perceived vaccination risk, perceived probability of VAE, and vaccination intention with the three manipulated factors (relative frequency of narratives, source of narrative and statistical information) and assessed whether the narrative bias was moderated by either source. In a second step, we explored whether numeracy, general attitude toward vaccination, and source used for the judgment moderated the narrative bias

Perceived risk of vaccination. Results of the first regression analysis are summarized in Table 1. Figure 1 displays mean ratings across conditions. A higher relative frequency led to higher perceptions of vaccination risk (H1a). Furthermore, we observed a main effect of narrative source, that is, participants generally reported greater perceived risk of vaccination when narratives originated from a neutral health forum as compared to an anti-vaccination website. However, in contrast with our expectations, we found no interaction between relative frequency and narrative source (contradicting H3). Furthermore, source credibility of the statistical information also did not moderate the narrative bias (contradicting H2).

Table 1
Regression analysis 1 predicts perceived risk of vaccination with the independent variables

	Perceived risk of vaccination $R^2 = .17$		
Predictors	*B*	*SE*	β
Constant	34.097	1.375	
Relative frequency of narratives reporting VAE (35%, 85%)	9.071	1.378	.374***
Credible statistical information (=1, rest = 0)	−0.704	1.563	−.029
Not credible statistical information (=1, rest = 0)	1.470	1.558	.061
Narrative source (0 = anti-vaccination website, 1 = online health forum)	3.600	1.386	.148*
Relative frequency × credible statistical information	−1.598	1.567	−.066
Relative frequency × not credible statistical information	−0.701	1.561	−.029
Relative frequency × narrative source	0.244	1.389	.010

*$p < .05$. ***$p < .001$.

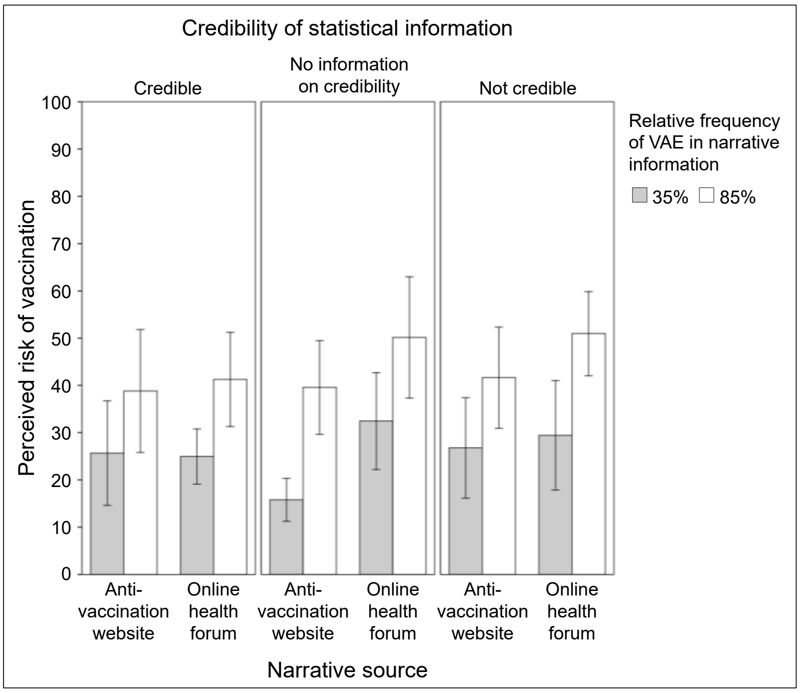

Figure 1. Perceived risk of vaccination as a function of the relative frequency of narratives reporting VAE, the credibility of the narratives' source and the credibility of the statistical information. Error bars represent 95% confidence intervals.

Table 2
Regression analysis 1 predicts perceived probability of VAE with the independent variables

	Perceived probability of VAE $R^2 = .17$		
Predictors	*B*	*SE*	β
Constant	28.407	1.172	
Relative frequency of narratives reporting VAE (35%, 85%)	7.970	1.175	.387***
Credible statistical information (=1, rest = 0)	−0.417	1.332	−.020
Not credible statistical information (=1, rest = 0)	0.234	1.328	.011
Narrative source (0 = anti-vaccination website, 1 = online health forum)	2.484	1.181	.120*
Relative frequency × credible statistical information	0.267	1.336	.013
Relative frequency × not credible statistical information	0.616	1.330	.030
Relative frequency × narrative source	0.160	1.184	.008

Note: VAE = vaccine adverse events.
$^*p < .05$. $^{***}p < .001$.

Perceived probability of VAE. As displayed in Table 2, the results for the perceived probability of VAE mirrored the results for perceived risk: more positive narratives resulted in higher probability judgments (H1b). The mean (*SD*) perceived probability of VAE was 20.75 (9.94) in the 35% conditions and 36.74 (25.48) in the 85% conditions, demonstrating the biasing effect of narratives. Again, the two source manipulations did not lead to the expected interactions (contradicting H2 and H3). We again observed the aforementioned main effect of narrative source, indicating that narratives from the neutral source led to higher perceived probability of VAE.

Intention to get vaccinated. Hypothesis 4a predicts that the narrative bias also manifests itself in intentions to get vaccinated. To control for the other manipulated variables, we included intention in the same regression as the risk variables (Table 3). The results support Hypothesis 4a: A higher number of narratives reporting VAE decreased intentions to get vaccinated. In addition, when the narratives originated from a neutral online health forum, as compared to an anti-vaccination website, intention decreased.

Table 3
Regression analysis 1 predicts intention to get vaccinated with the independent variables

	Intention to get vaccinated $R^2 = .09$		
Predictors	*B*	*SE*	β
Constant	4.614	0.103	
Relative frequency of narratives reporting VAE (35%, 85%)	−0.336	0.104	−.193**
Credible statistical information (=1, rest = 0)	0.048	0.117	.028
Not credible statistical information (=1, rest = 0)	−0.208	0.117	−.119
Narrative source (0 = anti-vaccination website, 1 = online health forum)	−0.297	0.104	−.171**
Relative frequency × credible statistical information	−0.082	0.118	−.047
Relative frequency × not credible statistical information	−0.029	0.117	−.017
Relative frequency × narrative source	0.081	0.104	.046

Note: VAE = vaccine adverse events.
***p* < .01.

Mediation analysis. We predicted that the biasing effect of the narrative information on vaccination intention would be mediated by risk perception (H4b). To test this assumption, we estimated the indirect effect employing the method described by Preacher and Hayes (2004) using bootstrapped estimates of confidence intervals (5,000 samples).[3] As predicted, the indirect effect of the relative frequency of positive narratives on vaccination intention through perceived vaccination risk was significant (indirect effect = −.22, standard error = .04, 95% CI [−.30, −.15]), indicating mediation. Figure 2 represents a path model of the mediation, illustrating that there was no direct effect of relative frequency on vaccination intention.

Exploratory Analysis

We calculated a hierarchical regression entering the control variables in the first block and the simple model variables in the second. Table 4 displays the explorative model's coefficients. Inclusion of the control variables significantly increased the amount of variance that could be explained by the model, change in $R^2 = 0.24$, $F(7, 257) = 11.42$, $p < .001$. The predictors' variance inflation factors ranged from 1.00 to 1.32 with an average variance

[3] We used the SPSS macro PROCESS provided by Andrew Hayes at http://afhayes.com.

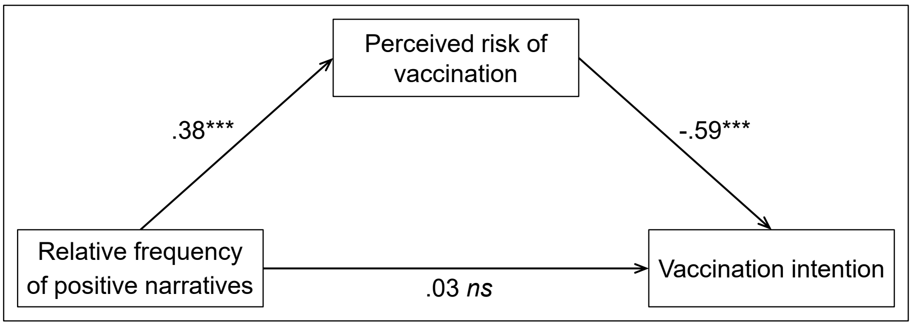

Figure 2. Path model of the indirect effect of the relative frequency of positive narratives on vaccination intention through perceived risk of vaccination. The numbers are standardized regression coefficients. ***$p \leq .001$.

inflation factor of 1.14, indicating a lack of multicollinearity (Cohen et al., 2003). Neither the narrative bias nor the main effect of narrative source could be explained by any of the control variables.

Considering the control variables' independent contributions, the results indicate that participants with a greater ability to understand and use numeric information judged the risk of vaccination to be lower. This effect was qualified by a significant interaction between the relative frequency of positive narratives and numeracy, that is, only when most narratives reported VAE did higher numeracy scores reduce risk perception (85%: $r = -.32$, $p < .001$, 35%: $r = .00$, $p = .959$).

Participants who were generally in favor of vaccination judged the risk of vaccination to be significantly lower than participants who were more opposed to vaccination. This effect held when only a few narratives reported VAE (35%: $r = -.53$, $p < .001$). When almost all narratives reported VAE, the general attitude was no longer related to risk perception (85%: $r = -.14$, $p = .115$).

The self-report measure regarding the information sources used to make judgments confirmed the observed effects: Participants indicating that they attributed more weight to the narrative information in their judgments reported higher perceived vaccination risk. As indicated by two significant interactions, weighting narratives more strongly than statistical information only led to higher risk perceptions when most narratives reported VAE ($r = .48$, $p < .001$; 35%: $r = .09$, $p = .311$) and when narratives originated from a neutral online health forum, ($r = .35$, $p < .001$; anti-vaccination: $r = .12$, $p = .180$).

Table 4

Regression analysis 2 predicts perceived risk with the extended model including numeracy, general attitude toward vaccination, and self-reported weighting of information sources

	Perceived risk of VAE $R^2 = .41$		
Predictors	*B*	*SE*	β
Constant	33.980	1.189	
Relative frequency of narratives reporting VAE (35%, 85%)	9.520	1.182	.392***
Credible statistical information (=1, rest = 0)	−1.516	1.348	−.062
Not credible statistical information (=1, rest = 0)	−0.364	1.353	−.015
Narrative source (0 = anti-vaccination website, 1 = online health forum)	2.686	1.204	.111*
Relative frequency × credible statistical information	−1.351	1.360	−.056
Relative frequency × not credible statistical information	−1.300	1.354	−.053
Relative frequency × narrative source	0.653	1.207	.027
Numeracy	−2.477	1.233	−.102†
Relative frequency × numeracy	−2.804	1.238	−.115*
General attitude	−6.272	1.201	−.258***
Relative frequency × general attitude	4.165	1.213	.171**
Source weighting (lower values indicate more weight given to the statistical information)	5.026	1.253	.207***
Relative frequency × source weighting	4.228	1.253	.174**
Narrative source × source weighting	2.646	1.217	.108*

† $p = .05$. *$p < .05$. **$p < .01$. ***$p < .001$.

Perceived Severity of VAE

We calculated the same simple regression model with perceived severity as the dependent variable and found the same pattern of significant predictors ($R^2 = .08$) as for probability and risk perception. The effect of the relative frequency of positive narratives on perceived severity resulted in a standardized regression coefficient of $\beta = .168$, $p = .006$, as compared to the frequency's effect on perceived probability of $\beta = .387$, $p < .001$. Narrative source, on the other hand, predicted perceived severity with $\beta = .206$, $p = .001$, whereas the effect on perceived probability was substantially smaller, $\beta = .120$, $p = .036$. Estimating the indirect effect of narrative source on perceived risk of vaccination through perceived probability of VAE and perceived severity of VAE indicated a significant mediation (total indirect effect = .13, standard error = .04, 95% CI [.06, .21]), in which the indirect effect through severity (indirect effect = .09, standard error = .03, [.04, .15]) was roughly twice as large as the indirect effect through probability (indirect effect = .04, standard error = .02, [.00, .10]). A pair-wise contrast, however, revealed that the two indirect effects did not differ significantly (contrast = −.04, standard error = .04, [−.12, .03]). There was no significant direct effect of narrative source on vaccination risk perception (direct effect = .02, standard error = .05, [−.08, .12]; Preacher & Hayes, 2008).

Discussion

In this study we tested whether the biasing effect of narrative information on risk perceptions is moderated by source credibility—that is, the credibility of the narrative as well as that of the statistical information delivering the base-rate. In all conditions, we observed a stable narrative bias: even though participants had statistical base-rate information about the occurrence of vaccine adverse events (VAE), normatively irrelevant information conveyed by narratives from online discussion boards systematically increased perceptions of vaccination risks. Contrary to our expectations, informing participants that the statistical base-rate was based on unreliable or very reliable data did not moderate the narrative bias. Furthermore, the narrative bias occurred irrespective of whether the narratives originated from an anti-vaccination website or a neutral health forum. This lack of both expected moderator effects is mirrored in the self-reported weighting of information sources, as there were no significant differences in reported source weighting between the three statistic conditions as well as a negligible effect of narrative source. Thus, it appears that informing individuals about the credibility of different types of risk information does not affect the integration thereof. This is in line with previous research, which showed that the type of information (i.e., statistic and narrative) appears to determine their relative weighting (Betsch et al., 2013).

It is important to note that we found no effect of narrative source manipulation on the manipulation check variables. However, the manipulation check items regrettably pertained to the narratives rather than the whole forum. Even though the narratives were presented on a website with an agenda, there was no indication that the individual narratives had also been authored for this purpose. Thus, it appears that reports of vaccination experiences by “people like me” are trusted irrespective of the source that delivers them (Haase & Betsch, 2012). Future research should manipulate the credibility of the narratives and the forum independently to assess whether the expected moderator effect occurs when the credibility of the narratives is low.

While the narrative source manipulation did not eliminate the narrative bias, we found that presenting the narratives in the context of an anti-vaccination forum led to a general decrease in vaccination risk perceptions. Thus, even when the narratives originated from a non-credible source, their biasing effect remained stable—although on a generally lower level of perceived risk. Considering the two constituents of risk separately—probability and severity—might help to partially explain this finding. The additional analyses reported above revealed that the main effect of relative frequency on perceived probability was approximately twice as large as that on perceived severity. In contrast, narrative source affected severity perceptions almost twice as strongly as it affected perceived probability. Furthermore, the main effect of narrative source on perceived risk primarily occurred through its effect on severity. There is no compelling theoretical reason to assume an interaction effect of narrative source and the relative frequency of positive cases on perceived severity. Given that narrative source affected perceived vaccination risk primarily through perceived severity of VAE, the observed main effect of narrative source on perceived risk (rather than an interaction) would be expected. However, this explanation requires that both risk constituents are perceived independently—both the current study and previous research (Betsch et al., 2013; Harris, Corner, & Hahn, 2009) have demonstrated that this is not the case, for example, we observed an effect of relative frequency (a pure likelihood manipulation) on perceived severity as well as an effect of narrative source on perceived probability.

Nonetheless, the observed main effect indicates that the source of the narratives offers an interpretative frame for the narratives' content, rather than for their frequency. Anti-vaccination websites often indicate their agenda in the title (e.g., *www.vaccineinjury.info*). The names of the forums in the current study closely followed this pattern and were constantly visible at the top of the screen while participants read the narratives. This consistent display of an agenda together with the persuasive information might have induced psychological reactance and resulted in reduced persuasion, that is, lower risk perception (Pavey & Sparks, 2009). Future research should investigate whether comparable effects are observed when a neutral source is compared with providers with an explicit pro-vaccination title (such as *@vaccineswork* on Twitter).

In line with previous research (Reyna et al., 2009), participants lower in numeracy overestimated the vaccination risk more so than those with greater numerical skills. However, this effect was limited to the 85% condition, that is, when almost all narratives implied vaccination risks. Fuzzy-trace theory might offer an explanation for this apparent discrepancy. According to this theory, individuals extract two kinds of representations from information:

Verbatim representations are precise, whereas gist representations capture the bottom line meaning. Judgments and decisions are informed by gist rather than verbatim representations (Reyna, 2008). The manipulation check indicated that most participants were able to recall absolute numbers of positive narratives that were in a range of ±2 to the correct answer. Considering this somewhat imprecise verbatim recall of positive narratives, we assume that, in the 35% condition, the gist of the vaccination risk was simply *low* and, therefore, more or less congruent with the statistical information that also proposed *low* riskiness (20%). Consequently, participants may have felt no need to adjust their risk representations that they had already constructed based on the statistical information. Evidence for this notion can be found in the probability ratings. In the 35% condition, the perceived probability ratings did not differ significantly from 20%, the percentage conveyed in the statistic, $t(137) = 0.89$, $p = .374$. In the 85% condition, the gist of the risk of vaccination may have been *high*. Hence, the original idea of how risky vaccination is differed from the gist derived from the online discussion board. Only in this case did participants need to adjust their risk perceptions. This adjustment was affected by individual numeracy: Participants low in numeracy were more prone to the biasing effect of the narratives, which is in line with previous findings (Dieckmann, Slovic, & Peters, 2009). Accordingly, we interpret the finding that participants' general attitude toward vaccination only affected risk judgments in the 35% condition: When statistical and narrative evidence were more or less congruent, vaccination risk perceptions were primarily a function of general attitudes. The overwhelming narrative evidence for VAE in the 85% condition, however, made an adjustment necessary and appeared to convince even participants with a pro-vaccination attitude that this vaccination is risky.

Limitations

First, we used self-report measures pertaining to a hypothetical scenario. While ethical considerations prescribe this approach, the potential lack of external validity should be kept in mind when interpreting results. Future research could try to improve upon this issue by paying participants dependent on performance (making the bias costly) and thereby rendering the scenario more akin to a real-life health decision (Hertwig & Ortmann, 2001).

Secondly, while finding the same results in two different samples would usually be supportive of external validity, the precise effects of combining the two samples—laboratory and online—in the current study cannot be meaningfully assessed due to small cell sizes. There were, however, no main effects of the sample on any of the three dependent variables (all ts < 1).

Conclusions

We found an extremely stable bias of normatively irrelevant narrative information on perceptions of a statistically conveyed vaccination risk. This bias affects perceptions of vaccination risks through both the perceived probability and severity of VAE and indirectly affects individuals' intentions to get vaccinated. Informing people about the credibility of a statistic had no correcting effect on the bias. This implies that health communicators cannot counter the narrative bias by underlining the excellent reliability of their data (see Nyhan, Reifler, Richey, & Freed, 2014, for comparable results). Rather, it appears that techniques pertaining to the narratives themselves are needed. In a previous study, we were able to show that bias awareness disclaimers emphasizing the biased sampling of narratives can decrease the effect of narratives—although the disclaimer's effect was very small; and the narrative bias still occurred (Betsch et al., 2013). The present study adds to evidence that informing individuals about the ulterior motive behind presenting narratives does not impede their biasing effect.

Narrative evidence is a common feature of anti-vaccination websites and, unlike in the current study, contains almost exclusively reports of very severe VAE (Kata, 2010, 2012). Apparently, these reports affect individuals' risk perceptions and vaccination intentions, irrespective of whether they were collected with a specific agenda in mind. Fortunately, it appears that some anti-vaccination websites may inadvertently reduce this effect by stressing their agenda, which potentially creates reactance within the reader. Overall, the present research demonstrates once more how difficult it is to counter the biasing effect of narrative information.

References

Betsch, C., Brewer, N. T., Brocard, P., Davies, P., Gaissmaier, W., Haase, N., … Stryk, M. (2012). Opportunities and challenges of Web 2.0 for vaccination decisions. *Vaccine, 30*(25), 3727–3733. https://doi.org/10.1016/j.vaccine.2012.02.025

Betsch, C., Renkewitz, F., & Haase, N. (2013). Effect of narrative reports about vaccine adverse events and bias-awareness disclaimers on vaccine decisions: A simulation of an online patient social network. *Medical Decision Making, 33*(1), 14–25. https://doi.org/10.1177/0272989X12452342

Betsch, C., Ulshöfer, C., Renkewitz, F., & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making, 31*(5), 742–753. https://doi.org/10.1177/0272989X11400419

Borgida, E., & Nisbett, R. E. (1977). The differential impact of abstract vs. concrete information on decisions. *Journal of Applied Social Psychology, 7*(3), 258–271. https://doi.org/10.1111/j.1559-1816.1977.tb00750.x

Brewer, N. T., Chapman, G. B., Gibbons, F. X., Gerrard, M., McCaul, K. D., & Weinstein, N. D. (2007). Meta-analysis of the relationship between risk perception and health behavior: The example of vaccination. *Health Psychology, 26*(2), 136–145. https://doi.org/10.1037/0278-6133.26.2.136

Brosius, H.-B., & Bathelt, A. (1994). The utility of exemplars in persuasive communications. *Communication Research, 21*(1), 48–78. https://doi.org/10.1177/009365094021001004

Brown, D. W., Burton, A., Gacic-Dobo, M., Karimov, R. I., Vandelaer, J., & Okwo-Bele, J. M. (2011). A mid-term assessment of progress towards the immunization coverage goal of the Global Immunization Vision and Strategy (GIVS). *BMC Public Health, 11*(806), 1–7. https://doi.org/10.1186/1471-2458-11-806

Burki, T. (2013). Challenges and targets for measles elimination. *The Lancet Infectious Diseases, 13*(6), 479–480. https://doi.org/10.1016/S1473-3099(13)70133-6

Chen, R. T. (1999). Vaccine risks: Real, perceived and unknown. *Vaccine, 17*(Suppl 3), S41–S46. https://doi.org/10.1016/S0264-410X(99)00292-3

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ, US: Lawrence Erlbaum.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making, 7*(1), 25–47. Retrieved from http://journal.sjdm.org/11/11808/jdm11808.html

de Wit, J. B. F., Das, E., & Vet, R. (2008). What works best: Objective statistics or a personal testimonial? An assessment of the persuasive effects of different types of message evidence on risk perception. *Health Psychology, 27*(1), 110–115. https://doi.org/10.1037/0278-6133.27.1.110

Dieckmann, N. F., Slovic, P., & Peters, E. M. (2009). The use of narrative evidence and explicit likelihood by decisionmakers varying in numeracy. *Risk Analysis, 29*(10), 1473–1488. https://doi.org/10.1111/j.1539-6924.2009.01279.x

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making, 25*(4), 398–405. https://doi.org/10.1177/0272989X05278931

Glasman, L. R., & Albarracín, D. (2006). Forming attitudes that predict future behavior: a meta-analysis of the attitude-behavior relation. *Psychological Bulletin, 132*(5), 778–822. https://doi.org/10.1037/0033-2909.132.5.778

Green, M. C., & Brock, T. C. (2000). The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology, 79*(5), 701–721. https://doi.org/10.1037//0022-3514.79.5.701

Green, M. C., & Brock, T. C. (2002). In the mind's eye: Transportation-imagery model of narrative persuasion. In M. C. Green, J. J. Strange, & T. C. Brock (Eds.), *Narrative impact: Social and cognitive foundations*. (pp. 315–341). Mahwah, NJ, US: Lawrence Erlbaum.

Haase, N., & Betsch, C. (2012). Parents trust other parents: Lay vaccination narratives on the Web may create doubt about vaccination safety. *Medical Decision Making, 32*(4), 645. https://doi.org/10.1177/0272989X12445286

Harris, A. J. L., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition, 110*(1), 51–64. https://doi.org/10.1016/j.cognition.2008.10.006

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(3), 383–451.

Hinyard, L. J., & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior, 34*(5), 777–792. https://doi.org/10.1177/1090198106291963

Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly, 15*(4), 635–650. https://doi.org/10.1007/BF02716996

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine, 28*(7), 1709–1716. https://doi.org/10.1016/j.vaccine.2009.12.022

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine, 30*(25), 3778–3789. https://doi.org/10.1016/j.vaccine.2011.11.112

Nyhan, B., Reifler, J., Richey, S., & Freed, G. L. (2014). Effective messages in vaccine promotion: A randomized trial. *Pediatrics, 133*(4), e835–e842. https://doi.org/10.1542/peds.2013-2365

O'Reily, T. (2005). *What is web 2.0 design patterns and business models for the next generation of software.* Retrieved from https://oreilly.com/web2/archive/what-is-web-20.html

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition, 37*(5), 632–643. https://doi.org/10.3758/MC.37.5.632

Omer, S. B., Orenstein, W. A., & Koplan, J. P. (2013). Go big and go fast — vaccine refusal and disease eradication. *The New England Journal of Medicine, 368*(15), 1374–1376. https://doi.org/10.1056/NEJMp1300765

Pavey, L., & Sparks, P. (2009). Reactance, autonomy and paths to persuasion: Examining perceptions of threats to freedom and informational value. *Motivation and Emotion, 33*(3), 277–290. https://doi.org/10.1007/s11031-009-9137-1

Pornpitakpan, C. (2004). The persuasiveness of source credibility: A critical review of five decades' evidence. *Journal of Applied Social Psychology, 34*(2), 243–281. https://doi.org/10.1111/j.1559-1816.2004.tb02547.x

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. *Behavior Research Methods, Instruments, & Computers, 36*(4), 717–731. https://doi.org/10.3758/BF03206553

Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making, 28*(6), 850–865. https://doi.org/10.1177/0272989X08327066

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943–973. https://doi.org/10.1037/a0017327

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*(11), 966–972. https://doi.org/10.7326/0003-4819-127-11-199712010-00003

The Lancet. (2010). Retraction—Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet, 375*(9713), 445. https://doi.org/10.1016/S0140-6736(10)60175-4

Wakefield, A. J., Murch, S. H., Anthony, A., Linnell, J., Casson, D. M., Malik, M., … Walker-Smith, J. A. (1998). RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. *The Lancet, 351*(9103), 637–641. https://doi.org/10.1016/S0140-6736(97)11096-0

World Health Organization. (2013). *Global Vaccine Action Plan 2011–2020*. Retrieved from http://www.who.int/immunization/global_vaccine_action_plan/GVAP_doc_2011_2020/en

Yates, J. F., & Stone, E. R. (1992). The risk construct. In J. F. Yates (Ed.), *Risk-taking behavior*. (pp. 1–25). Oxford England: John Wiley & Sons.

Appendix: Texts Used for Manipulated Factors

Note that the original materials were in German.

Discounting Cue for Statistical Information

Some time after its publication, the German Research Foundation and the Robert-Koch-Institute described this study as being poor with regard to the scientific methods employed and the data analysis. Accordingly, this study does not provide reliable or generally valid data about the occurrence of VAE following the vaccination against dysomeria.

Augmenting Cue for Statistical Information

Some time after its publication, the German Research Foundation and the Robert-Koch-Institute described this study as being exemplary with regard to the scientific methods employed and the data analysis. Accordingly, this study provides highly reliable and generally valid data about the occurrence of VAE following the vaccination against dysomeria.

Description of Both Websites, Neutral Health Forum Mentioned First

On Health-net.de people can share their experiences with medications and adverse events. The explicit goal of Health-net.de is not only to support people with their health decisions but also to provide real-life data to manufacturers and regulatory authorities and thereby increase others' safety. Vaccination-harms.de is a forum, run by so called anti-vaccination activists. Anti-vaccination activists reject any kind of vaccination, mostly on ideological grounds. One key argument is that, due to the high risk of severe adverse events and subsequent damages, vaccinations are very dangerous—a claim that cannot be substantiated scientifically. The explicit goal of Vaccination-harms.de is to collect as many cases of harm through vaccination as possible in order to provide support to this claim.

Example for a Narrative Reporting the Occurrence of VAE (Positive)

Hey! I am Anna and I got vaccinated against dysomeria a month ago. Apart from the pricking I didn't notice anything bad during the first few days following the vaccination. After a week however I started feeling dizzy constantly. It got so bad that my boss sent me home.

Example for a Narrative Reporting the Nonoccurrence of VAE (Negative)

Hi! I got vaccinated against dysomeria about a month ago. I hadn't gotten any shots in years and had forgotten what a quick procedure this is. Everything went really well. I just didn't like the shot so much but I guess that's normal. Jens

Article 2

The Narrative Bias Revisited: What Drives the Biasing Influence of Narrative Information on Risk Perceptions?

Reference:
Betsch, C., Haase, N., Renkewitz, F., & Schmid, P. (2015). The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions? *Judgment and Decision Making, 10*(3), 241–264.

The definitive version is available at:
http://journal.sjdm.org/14/141206a/jdm141206a.pdf

The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions?

Cornelia Betsch, Niels Haase, Frank Renkewitz, and Philipp Schmid
University of Erfurt, Germany

Author Note

Cornelia Betsch, Center of Empirical Research in Economics and Behavioral Sciences (CEREB), University of Erfurt, Germany; Niels Haase, Department of Psychology, University of Erfurt, Germany; Frank Renkewitz, Department of Psychology, University of Erfurt, Germany; Philipp Schmid, Department of Psychology, University of Erfurt

Cornelia Betsch and Niels Haase contributed equally to this paper. This research was financed by a research grant from the German Research Foundation (BE 3970/4-1) to the first and third authors. The authors are grateful to Alexandra Schmitterer for her help in conducting the study as well as to Heather Fuchs, Jonathan Baron, Edward Cokely, Gary Brase and one anonymous reviewer for helpful comments on a previous draft of this article.

Correspondence concerning this article should be addressed to Cornelia Betsch, Center for Empirical Research in Economics and Behavioral Sciences (CEREB), University of Erfurt, Nordhaeuser Strasse 63, 99089 Erfurt, Germany.
E-mail: cornelia.betsch@uni erfurt.de

Abstract

When people judge risk or the probability of a risky prospect, single case narratives can bias judgments when a statistical base-rate is also provided. In this work we investigate various methodological and procedural factors that may influence this narrative bias. We found that narratives had the strongest effect on a non-numerical risk measure, which was also the best predictor of behavioral intentions. In contrast, two scales for subjective probability reflected primarily statistical variations. We observed a negativity bias on the risk measure, such that the narratives increased rather than decreased risk perceptions, whereas the effect on probability judgments was symmetric. Additionally, we found no evidence that the narrative bias is solely produced by adherence to conversational norms. Finally, changing the absolute number of narratives reporting the focal event, while keeping their relative frequency constant, had no effect. Thus, individuals extract a representation of likelihood from a sample of single-case narratives, which drives the bias. These results show that the narrative bias is in part dependent on the measure used to assess it and underline the conceptual distinction between subjective probability and perceived risk.

Keywords: Risk perception, subjective probability, narratives, cognitive bias, negativity bias

The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions?

Every day we encounter and act upon probabilistic information. From the weather forecast to consumer reports, individuals are regularly confronted with likelihood information about risks (e.g., the chance of rain) to inform their behavior in one way or another (e.g., whether to leave the house with or without an umbrella). In the medical domain, the advent of the modern shared decision making approach means that patients are increasingly involved in treatment and preventative decisions such as choosing between bypass surgery and angioplasty or deciding for or against vaccinations. All such decisions involve collecting, processing, and weighing probabilistic information. As a result, individual risk perception about medical matters has have been a recent focus of research.

At least 40 years of psychological research have produced an extensive catalog of situations in which likelihood estimates deviate from the prescriptions of probability theory (see Gilovich, Griffin, & Kahneman, 2002 for an overview). One such bias is the excessive influence of narrative information, exemplars, and testimonies, which we refer to as *narrative bias*. In a classic example, Borgida and Nisbett (1977) found that a few brief personal accounts had a far stronger impact on students' course choices than mean course evaluations. Such reasoning is considered to be biased, that is, formally incorrect, because it fails to weigh different samples of data according to the respective sample size.

Assessing the Narrative Bias

One difficulty in understanding the mechanisms behind the narrative bias and in coherently summarizing findings lies in the different measures used to assess the influence of narrative information. Dependent variables vary from subjective probability to perceived risk or actual decisions (Betsch, Renkewitz, & Haase, 2013; Betsch, Ulshöfer, Renkewitz, & Betsch, 2011; Fagerlin, Wang, & Ubel, 2005; Obrecht, Chapman, & Gelman, 2009).

Researchers on biases in risk perception commonly collect some sort of magnitude judgment regarding the likelihood of a specified event (e.g., de Wit, Das, & Vet, 2008; Knapp, Gardner, Raynor, Woolf, & McMillan, 2010; Lee, Schwarz, Taubman, & Hou, 2010). However, even the most parsimonious, and also most common, definition of perceived risk (following expected value theory) additionally includes a value dimension, that is, the significance or severity of a loss. Other concepts include the affective reaction to an outcome, the perceived source of a risk, the susceptibility to a risk, and degree of belief. Further, the voluntariness of risk, the knowledge about, and control over risk can also play a role in risk judgments (Brewer et al., 2007; Eiser, 1994; Gardoni & Murphy, 2013; Loewenstein, Weber, Hsee, & Welch, 2001; Slovic, Finucane, Peters, & MacGregor, 2004; Slovic, Fischhoff, & Lichtenstein, 1979; see Brun, 1994 for a comprehensive discussion of this topic). Thus, it seems prudent to distinguish between subjective probability and perceived risk.

In addition, although subjective probability and perceived risk are central variables in many studies, there is no consensus regarding their measurement. Methods include inferences from bets (Beach & Phillips, 1967), balls and bins tasks (Goldstein & Rothschild, 2014), risk matrices (Ball & Watt, 2013), and various self-report formats. The latter typically elicit a type of magnitude judgment and include numeric estimates, rating scales, and visual analog scales. One goal of this paper is to compare narrative biases across different measures used in previous research (Betsch et al., 2013, 2011; Obrecht et al., 2009).

Theoretical Accounts

Theoretical accounts of the highly persuasive effect of narrative evidence vary in focus and scope. Some explanations focus on the content of the narrative itself, which elicits affective reactions and immersion (see Hinyard & Kreuter, 2007 for an overview). Indeed, findings from previous research have shown that highly emotional narratives reporting vaccine adverse events increase the perceived risk of vaccination compared to less emotional narratives (Betsch et al., 2011). However, other findings show that the narrative bias occurs even when the content of the narrative is free of emotion and contains only the statement that the critical event occurred (Betsch et al., 2013; Obrecht et al., 2009). In this paper, we focus on a more formal approach that explains the narrative bias based solely on the structure of statistical and narrative information regardless of the narratives' qualitative content.

Previous research that presented both statistical and narrative evidence to subjects has led to comparable results but differed regarding the causal explanation put forward by the authors. Ubel, Jepson, and Baron (2001) examined the importance of the match between statistical and narrative information and found that narratives were especially influential when the ratio of narratives indicating success vs. failure of a treatment was incongruent with previously presented statistical evidence. The effect, however, disappeared when controlling for the absolute number of narratives. Nevertheless, this finding indicates that individuals may perceive a set of narratives as a single unit of information—comparable to statistical information—that conveys the relative frequency of events.

Contrary to this idea, Obrecht et al. (2009) developed the encounter frequency theory, which assumes that each piece of information, be it a statistic or a single narrative case, is attributed equal weight when forming a judgment. Accordingly, individuals simply count each piece of information indicating the (non)occurrence of an event. Encounter frequency theory does not specify the process of how positive and negative counts are integrated. However, this account suggests that changing the absolute number of narratives reporting the occurrence of a focal event while keeping their relative proportion constant will affect judgments or decisions. A similar notion can be found in research on the ratio-bias or denominator neglect—that is, the phenomenon that individuals tend to prefer a gamble with a $\frac{9}{100}$ likelihood of winning over a gamble with a $\frac{1}{10}$ likelihood, because they tend to ignore the denominator (Denes-Raj & Epstein, 1994; Reyna & Brainerd, 2008).

Thus, the second goal of this paper is to clarify whether the narrative bias relies on the relative or absolute number of narratives reporting the critical event.

Negativity Bias

There is some evidence that individuals tend to weigh information regarding the presence of a risk more strongly than information concerning its absence (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001; Rozin & Royzman, 2001; Siegrist & Cvetkovich, 2001). This negativity bias means that narratives may have an asymmetric effect. Narratives implying a higher risk than the provided statistical information would have a stronger influence on risk perceptions than narratives implying a lower risk than the statistical information. Therefore, a third goal is to test whether narratives can both increase and decrease risk perceptions, relative to the perception resulting from the statistical information alone.

Experimental Artifact

We also strive to test the narrative bias against two potential and related alternative explanations that are inherent in the experimental procedure. First, it is possible that the narrative bias occurs simply because subjects follow conversational norms. That is, as experimenters we assume that the statistical information is the most or even only relevant information for the judgment. We expect individuals to attribute less weight to or even ignore the less reliable narrative information. However, Grice (1975) argues that conversation follows certain norms of cooperation, one of which states that communicated information is to be relevant. Thus, from the subjects' point of view, all information provided by the experimenter may appear relevant for judgment due to the simple fact that it has been provided.

Second, in most studies comparing the influence of statistical and individuating information, statistical information is presented first followed by individuating information. Thus, it is possible that the narrative bias is at least partially caused by a recency effect. Expanding on the idea of conversational norms, Krosnick, Li, and Lehman (1990) argue that more informative and thus more important information is typically provided last, especially when two contradictory pieces of information are presented. Accordingly, subjects may assume that the experimenter considers the second piece of information, that is, the narrative information, to be more important and that they should, in turn, do the same when making a judgment.

Finally, and related to this, we will investigate whether subjects would seek the narrative information at all if it were not provided. The act of seeking more information when sound statistical evidence is already available results in added costs to the individual—at least in terms of time. From a *homo oeconomicus* point of view, statistics provide the necessary likelihood information to quickly make a decision and should, therefore, be preferred over the time consuming evaluation of narrative reports.

Summary of Research Questions and Overview

Perceived risk and subjective probability are conceptually different; and there is no consensus on how to measure either one. Previous research has studied the narrative bias effect on both subjective probability, assessed either as percent estimates or by rating scales, and perceived risk measured using a visual analog scale. Experiment 1 asks whether the narrative bias is dependent on the task. Specifically:

RQ1: Do narratives and statistical information have different relative effects on a numeric and a verbal measure of subjective probability and a visual analog scale measure of perceived risk?

In previous studies, the relative frequency of the critical event as implied by the narratives typically exceeded that given by the statistical information, which led to an increase in perceived risk (Betsch et al., 2013, 2011). Research on the negativity bias (Siegrist & Cvetkovich, 2001) suggests that narratives may have an asymmetrical influence on risk perceptions such that they will have a greater impact when they exceed rather than fall below statistical risk information. Experiment 1 will therefore address the following research question:

RQ2: Is the narrative bias symmetric or asymmetric?

The same number of narratives indicating the occurrence of an event will lead to different risk perceptions if we assume that the relative rather than absolute frequency influences risk perceptions. Therefore, in Experiment 2, we strive to answer the question:

RQ3: Is the narrative bias caused by the relative or absolute number of narratives reporting the focal event?

Experiment 2 also explores questions related to the experimental procedure aimed to rule out the possibility that the narrative bias is an experimental artifact:

RQ4: Is the narrative bias partially an artifact produced by conversational norms in that the narratives have to be encoded and therefore appear to be relevant for judgments?

RQ5: Is the narrative bias partially caused by a recency effect when narratives appear after the statistical information?

RQ6: Is narrative information an attractive source of information when statistical information about a risk is already provided?

For the experiment content domain, we use vaccination risks. The example of vaccination risks seems particularly relevant in this context for two reasons: a growing number of individuals facing health related decision consult the Internet for information (Fox & Duggan, 2013); and narrative evidence is a common feature on anti-vaccination activist websites that propagate alleged adverse events and high risks of vaccination (Betsch et al., 2012; Haase & Betsch, 2012; Kata, 2010, 2012). In both experiments, subjects receive statistical and narrative information about the occurrence of vaccine-adverse events (VAE). They then judge the riskiness of the vaccination as well as the subjective probability of VAE.

Experiment 1

In this experiment, we compared the effect of the narrative bias on three related measures. We asked for a percent estimate of the likelihood of VAE. Only numeric measures allow for a meaningful quantification of the narrative bias; and this format has been shown to be the least context dependent and less error-prone than judgments of relative frequency (Haase, Renkewitz, & Betsch, 2013; Weinstein & Diefenbach, 1997). As a second measure of subjective probability, we included a verbally labeled 7-point rating scale in order to retain comparability with previous studies (Betsch et al., 2011). Further, this 7-point rating scale has been shown to be superior in behavior prediction as compared to a percent measure (Weinstein et al., 2007). We will therefore explore whether we find comparable results regarding vaccination intentions. Finally, we assessed perceived risk by means of a visual analog scale. Since subjective probability and perceived risk are distinct constructs, we assessed all dependent variables for every subject and varied the order of assessment between subjects. In our analyses of the narrative bias, we examined only the first measure completed by each subject in order to exclude carry-over effects.

Because we differentiate between subjective probability and perceived risk, and definitions of risk typically include a value dimension, we additionally assessed the perceived severity of VAE. We also assessed the intention to get vaccinated in order to compare the different measures in terms of behavior prediction.

Method

Subjects and design. A total of 290 students at the University of Erfurt participated in this lab-experiment in exchange for a small gift and the chance to win one of ten €50 notes (approx. US$67.50). Thirteen subjects were excluded because they had either taken part in a similar experiment before or reported in a post-experimental interview that they were unsure about the handling of the scales. Thus, the final sample includes $N = 277$ subjects, with *n*s for individual analyses ranging from 22 to 27. Sixty-eight subjects (24.5%) were male and the mean age (*SD*) was 22.13 (3.16) years.

Each subject was randomly assigned to one of 12 conditions, resulting from a $2 \times 2 \times 3$ between-subjects design with the relative number of narratives reporting VAE as the independent variable (1 or 8 narratives of 20, equaling 5% and 40%), the statistical probability of VAE (5% or 40%) as a second factor, and the first dependent variable as a third factor (7-point rating scale, percent estimate or visual analog scale). In addition, we assessed subjects' numeracy, as previous work suggests that individuals with low numeracy may be especially prone to biases due to narrative information (Dieckmann, Slovic, & Peters, 2009; Peters, 2008).

Procedure. All materials were presented on a computer screen. Subjects were provided with information about a fictitious severe disease (*dysomeria*) and the recommended vaccination. This was accompanied by a statistic reporting the likelihood of VAE occurring. Subsequently, subjects were asked to imagine that they found additional information about experiences with the vaccination on an Internet bulletin board. The narratives there reported either the occurrence (positive) or non-occurrence (negative) of VAE. Afterwards, subjects completed the dependent variable measures.

Manipulation of the statistical probability of adverse events. The statistical probability of VAE was explicitly expressed in percent together with a pictograph, that is, a matrix of 100 elements colored in one of two ways which indicated the presence or absence of VAE (created with http://www.iconarray.com, last accessed on October 24, 2014). Pictographs have been shown to reduce the effect of narrative information (Fagerlin et al., 2005). We manipulated the statistical probability of VAE (5% vs. 40%) between conditions.

Manipulation of relative frequency of narratives reporting adverse events. The narratives reported either the occurrence or non-occurrence of adverse events, with the number of narratives reporting VAE depending on condition (1 vs. 8 of 20 reports, resulting in 5% and 40%, respectively). The narratives were approximately equal in length (mean number of words = 57.5 and 52.2 for positive and negative narratives, respectively). In addition, positive narratives were pretested on 9-point rating scales concerning the severity of reported VAE, emotionality of content, and credibility. We selected narratives with moderate severity and emotionality, that is, ratings did not differ from a midpoint rating of 5 (severity: all ts ≤ 1.66, emotionality: all ts ≤ 1.98). The narratives were rated as equally credible (mean ratings did not differ from a rating of 6, all ts ≤ 1.48). The fictional authors' first names for all narratives were balanced for gender. The narratives were displayed as single cases, with one narrative per page. The pages were displayed in random order. In order to minimize any systematic influence due to additional information in the text, for example, concerning the vaccination procedure, narratives were elected at random when the whole sample was not needed. For example, in the 5% condition, one positive narrative out of a total of eight positive narratives (that were used in the 40% condition) was drawn for each subject. Appendix A presents four example narratives.

Dependent variables. Table 1 provides an overview of all dependent variables. Subjects completed all measures; however, the order of the following measures was varied between subjects: the subjective probability of the occurrence of adverse events (measured via two measures: numeric estimate in percent and probability rating on a 7-point rating scale) and the perceived risk of the vaccination (visual analog scale). For perceived risk, we used a non-numeric format so as to avoid making the probability dimension especially salient. However, to allow for comparisons with the subjective probability judgments and a quantification of the narrative bias, the visual analog scale provided scores between 0 and 100. No numeric feedback was provided to subjects.

In the subjective probability conditions, subjects provided their ratings on the specific subjective probability measure followed by the respective other measure and the visual analog scale to assess perceived risk. In the perceived risk condition, risk was assessed on the visual analog scale followed by the subjective probability measures in counterbalanced order. After all three measures were completed we assessed the severity of the possible adverse events as well as subjects' intentions to get vaccinated.

Manipulation check. After completing the dependent measures, subjects were asked to reproduce the stated statistical probability (5% or 40% depending on condition) and report the number of cases that reported VAE on the bulletin board (1 or 8).

Table 1
Overview of dependent variables

Construct	Scale type	Wording
Subjective probability	*Percent estimate*	What is the probability of experiencing vaccine-adverse events if you get vaccinated? (You will experience adverse events with a probability of__ %)
Subjective probability	*7-point rating scale*	What is the probability of experiencing vaccine-adverse events if you get vaccinated? (1 = almost zero, 2 = very small, 3 = small, 4 = moderate, 5 = large, 6 = very large, 7 = almost certain).
Risk	*Visual analog scale*	How risky do you judge the vaccination to be? (0 = not risky at all, 100 = very risky)
Severity	*7-point rating scale*	How severe do you judge the possible adverse events of the vaccination to be? (1 = not severe, 7 = very severe).
Intention	*7-point rating scale*	If you had the possibility to get vaccinated in the next week, what would you do? (1 = I would definitely not get vaccinated, 7 = I would definitely get vaccinated)

Note: No numeric anchors were provided to the subjects. In Experiment 1 the materials were in German.

Subjective numeracy. Subjective numeracy was assessed with a German translation of the Subjective Numeracy Scale (Fagerlin et al., 2007; German translation by Keller, Siegrist, & Visschers, 2009). The eight items were answered on a 6-point scale, where higher ratings indicate greater subjective numeracy (e.g., How good are you at working with fractions?).

Results

Manipulation check. In both conditions, roughly 96% of subjects were able to reproduce the given statistical probability (5% or 40%). We assumed a correct recall of the number of narratives if the recalled number was plus/minus one. For the condition in which 1 narrative reported VAE, 94.9% correctly recalled the absolute number (M_1 = 1.40, SD_1 = 1.75). In the 8 cases condition, 51% (M_8 = 8.33, SD_8 = 2.29) correctly recalled the absolute frequency of narratives reporting VAE. As the results did not change after eliminating subjects who did not correctly recall the encoded information, we used the full sample in our analyses.

Subjective numeracy. The obtained internal consistency of all items was sufficient, Cronbach's α = .77. The mean score of answers constitutes the subjective numeracy score (potential range 1–6). The mean subjective numeracy score (4.21, *SD* = 0.76) did not differ across conditions (all η_p^2s in a 2 × 2 × 3 ANOVA were ≤ .01, all *p*s ≥ .09).

Subjective probability and risk perception. The first goal of this experiment was to compare the effects of statistical and narrative information on different measures of perceived risk and subjective probability. Therefore, we calculated regression analyses for the measures using only the subsamples that responded to the respective construct first in the order of dependent variables. We excluded the samples in which other measures were completed prior to the dependent variable of interest to exclude carry-over effects (e.g., the influence of the numerically recalled probability on the general judgment of risk). Thus, we calculated three linear regressions predicting subjective probability (percent estimate and rating scale) and risk (visual analog scale), respectively. For all analyses we used standardized, continuous predictors. Interactions were calculated as the mathematical products of the factors (Cohen, Cohen, West, & Aiken, 2003).

In a first regression, we entered the manipulated factors and their interaction. In a second regression, we added subjective numeracy and the interactions of the factors with subjective numeracy.[1] The main results of the separate regressions are displayed in Table 2. The narrative biases are displayed in Figure 1. Both the statistical base-rate (β = .81) and the narratives (β = .15) significantly influenced the subjective probability of experiencing adverse events assessed as percent estimate. The statistical information had a stronger influence than the narrative information. A similar pattern of effects occurred when subjective probability was assessed by means of a 7-point rating scale (statistical base-rate: β = .56, narratives: β = .16), although the narratives' influence was not significant. For perceived risk assessed with the visual analog scale, however, the influence of the statistical base-rate information was lower than that of the narratives (statistical base-rate: β = .29, narratives: β = .43), indicating a stronger narrative bias.

In order to assess the different effects of narratives and statistics on the three dependent measures we also correlated each of the three dependent measures with the difference between the statistical and narrative information. This was done only for those subjects for whom narrative information differed from the statistical base-rate. Differences

[1] As multiple regression can obscure relationships between variables Appendix B presents a full correlation matrix of the independent and dependent variables as well as numeracy.

Table 2

Subjective probability (percent estimate, 7-point rating scale) and perceived risk as a function of the statistical base-rate, relative frequency of narratives reporting VAE, and subjective numeracy (Experiment 1)

Subjective probability (percent estimate) $n = 89$	β	*p*	*β*	*p*
Statistical base-rate (5% vs. 40%)	**.81**	**< .001**	**.83**	**< .001**
Narratives: frequency of VAE (5% vs. 40%)	**.15**	**.02**	**.14**	**.03**
Statistical base-rate × narratives	.00	.97	−.02	.79
Subj. numeracy			−.01	.84
Subj. numeracy × statistical base-rate			.10	.12
Subj. numeracy × narratives			**−.14**	**.02**
Subj. numeracy × statistical base-rate × narratives			−.03	.58
R^2	.68		.71	
Subjective probability (7-point rating scale) $n = 94$	**β**	***p***	***β***	***p***
Statistical base-rate (5% vs. 40%)	**.56**	**< .001**	**.55**	**< .001**
Narratives: frequency of VAE (5% vs. 40%)	.16	.07	.17	.06
Statistical base-rate × narratives	.07	.43	.03	.71
Subj. numeracy			.03	.75
Subj. numeracy × statistical base-rate			**.22**	**.02**
Subj. numeracy × narratives			.09	.30
Subj. numeracy × statistical base-rate × narratives			.07	.42
R^2	.35		.39	
Perceived risk (visual analog scale) $n = 94$	**β**	***p***	***β***	***p***
Statistical base-rate (5% vs. 40%)	**.29**	**.002**	**.29**	**.002**
Narratives: frequency of VAE (5% vs. 40%)	**.43**	**< .001**	**.42**	**< .001**
Statistical base-rate × narratives	−.13	.14	−.14	.14
Subj. numeracy			.01	.93
Subj. numeracy × statistical base-rate			.02	.87
Subj. numeracy × narratives			−.03	.74
Subj. numeracy × statistical base-rate × narratives			−.06	.51
R^2	.29		.30	

Note. Standardized betas (β) and respective *p*-values of significant effects are shown in boldface.

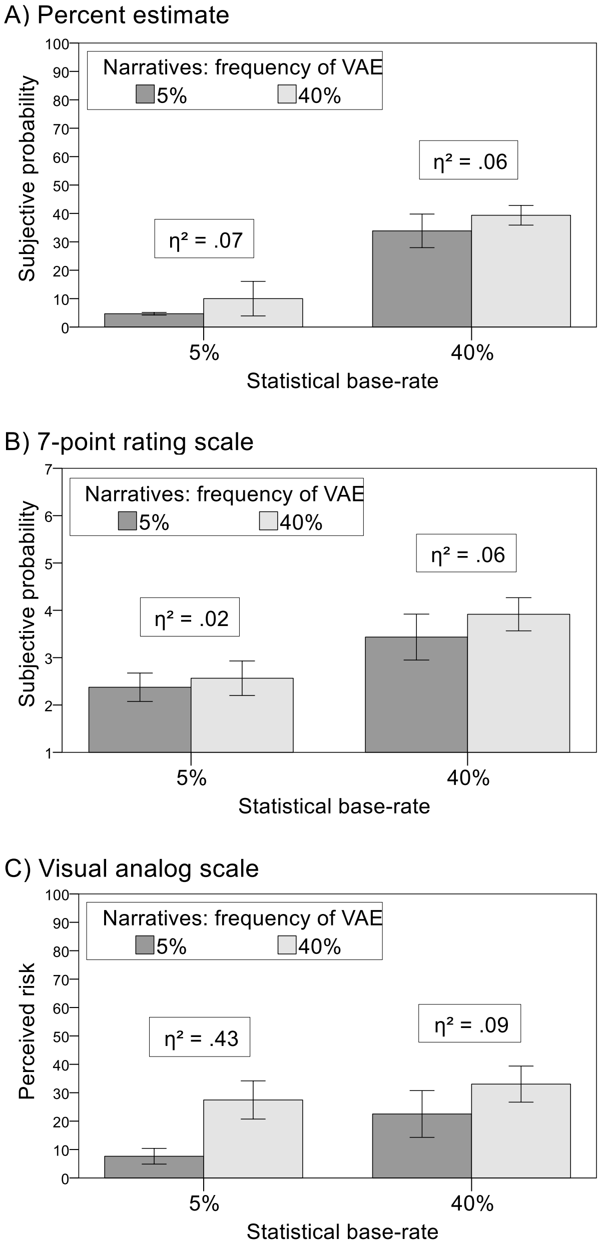

Figure 1. Subjective probability (A: n = 89, B: n = 94) and perceived risk (C: n = 94) as a function of statistical and narrative information. All factors were manipulated between subjects. Error bars = 95% CI.

were calculated by subtracting the frequency of VAE from the statistical base-rate, so a difference of −35% represents a low base-rate and a higher probability in the narratives, a difference of 35% represents the opposite. This way, a negative correlation indicates a stronger effect of narratives, a positive one a stronger effect of the statistical information. The correlations were $r = .67$, $p < .001$ for the percent estimate ($n = 44$), $r = .41$, $p = .005$ for the 7-point rating scale ($n = 46$), and $r = −.15$, $p = .343$ for the visual analog scale ($n = 45$). The last of these differed significantly from the other two (Percent estimate: Fisher's $z = 4.38$, $p < .001$, 7-point rating scale: Fisher's $z = 2.71$, $p = .007$).

When we entered subjective numeracy into the regression model, the percent estimate was a function of the number of narratives only when subjects had low subjective numeracy scores. Subjects high in subjective numeracy were unaffected by the number of narratives when judging the probability of VAE. This is evident in a significant interaction between subjective numeracy and the relative number of narratives ($\beta = -.14$).

When subjects judged the probability of VAE on the 7-point rating scale, the resulting judgments differed more strongly between the 5% and 40% statistical conditions for subjects high in subjective numeracy. Judgments by subjects low in subjective numeracy were more similar across statistical conditions. This was indicated by a significant interaction of subjective numeracy and statistical probability of VAE ($\beta = .22$).

Subjective numeracy did not affect ratings on the risk measure (all βs n.s.).

Carry-over effects. The results above suggest that risk judgments and subjective probability judgments are indeed very different—the probability judgments were less biased by narrative information than the risk judgment and depended more on the statistical base-rate. Judging probabilities before judging risk, therefore, may increase the saliency of the probability dimension of risk, resulting in a larger effect of the statistical base-rate on risk judgments. Conversely, judging risk before probability might increase the influence of narrative information. Both kinds of influence should manifest themselves in carry-over effects. To test this, we calculated three additional regression analyses with the respective

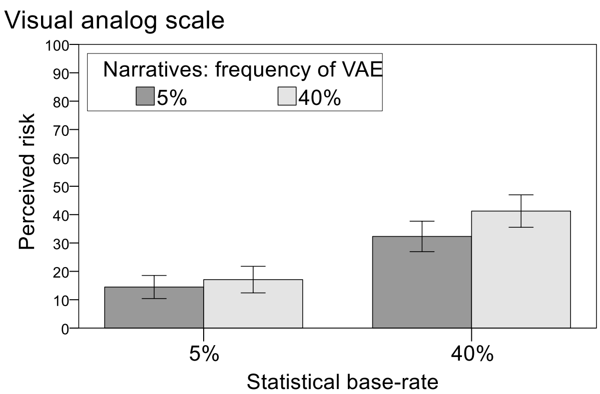

Figure 2. Perceived risk ($n = 183$) as a function of the statistical and narrative information. Subjective probability was assessed before the risk judgment. All factors were manipulated between subjects. Error bars = 95% CI.

other subsample, for example, we predicted ratings on the probability scales using only subjects that had first judged perceived risk and, vice versa, predicted risk judgments of subjects who had first estimated subjective probability of VAE.

Percent estimates were not affected by a prior risk judgment (statistical base-rate: $\beta = .81, p < .001$, narratives: $\beta = .19, p = .002$). Subjective numeracy had no influence, all other ts ≤ 1.45; $R^2 = .72$, $F(7, 86) = 31.62, p < .001$.

For the 7-point rating scale, on the other hand, we found a similar significant influence of statistical variation ($\beta = .56, p < .001$) but also a significant effect of the narratives ($\beta = .25$, $p = .003$). Again, there was no effect of subjective numeracy, all other ts ≤ 1.48; $R^2 = .41$, $F(7, 86) = 8.36, p < .001$. This indicates that considering the whole risk construct first renders a probability judgment more susceptible to the influence of irrelevant information.

For the risk measure, we found that judging probability first reversed the relative influence of both information types, thus resulting in a stronger influence of statistical information ($\beta = .51, p < .001$) than of the narratives ($\beta = .16, p = .012$) on subsequent risk judgments, all other ts ≤ 1.21; $R^2 = .32$, $F(7, 175) = 11.86, p < .001$. Note that increasing the salience of the probability dimension not only increased the effect of statistical information but also decreased the narratives' influence (see Figure 2 compared to Figure 1C).

Symmetry of the narrative bias. In order to assess whether the narrative bias is symmetric, we compared the effect sizes of the conditions when narratives were expected to increase vs. decrease the resulting judgments. The bars on the left in Figure 1 represent the case in which narratives should increase ratings of subjective probability and risk, because the relative frequency of narratives reporting VAE is equal to or greater than the statistical base-rate of 5%. The bars on the right represent the case in which narratives report an equal to or lower probability of VAE than the statistical information. If the narrative bias is symmetric, effect sizes displayed in Figure 1 should not differ between an expected increase vs. decrease (left vs. right).

In the case of the percent estimates, we observed a symmetric narrative bias in both directions. For the 7-point rating scale, the decreasing effect was slightly larger than the increasing effect, although both effects were rather small. For the risk measure, we observed a strong negativity bias, indicating that narratives increased rather than decreased risk perceptions.

Table 3
Correlations between subjective probability (percent estimate, 7-point rating scale), perceived risk, and intention to get vaccinated for the full sample in Experiment 1 (N = 277).

	Percent	7-point	Risk	Intention
Percent	—			
7-point	.70***	—		
Risk	.61***	.74***	—	
Intention	−.22***	−.34***	−.43***	—

****p* < .001.

Intention to get vaccinated. For each subsample, separate correlation analyses between the intention to get vaccinated and the respective dependent variable revealed virtually identical coefficients ($r_{PERC} = -.31$, $r_{7\text{-}POINT} = -.33$, $r_{RISK} = -.31$, all *p*s < .01). However, in an additional step-wise regression analysis across all subjects ($N = 277$) only perceived risk predicted the intention, whereas both other variables were excluded from the analysis ($\beta_{RISK} = -.43$, $p < .001$; $R^2 = .19$).

These results might indicate multicollinearity, that is, even though each measure predicts the intention on its own, they actually account for the same variance because they are correlated. In line with this, the correlation coefficients in Table 3 indicate that the percent estimate and the 7-point rating scale have some predictive power but that perceived risk accounts for the same as well as for additional and unique variance in vaccination intentions. Appendix C presents the same correlation matrices as Table 3 for each subsample. The absolute predictive power of all measures varies but the relation between measures remains stable, with perceived risk as the best predictor of behavioral intentions.

Summary. Narratives biased the perception of subjective probability and risk to different extents, depending on the measure with which the dependent variables were assessed. The relative effect of narratives was largest (and even a little larger than the effect of the statistical information) on perceived risk assessed with a visual analog scale. The narratives had a similar but smaller to negligible effect on both measures of subjective probability (RQ1). Variations in statistical information, on the other hand, had the greatest effect on subjective probability assessed as a percent estimate and the smallest effect on perceived risk assessed with a visual analog scale. These results underline the important conceptual distinction between subjective probability and perceived risk. Risk perception is often operationalized as a likelihood judgment. However, the manipulation of probabilistic information (all other variables were held constant) affected judgments of subjective probability and perceived risk differently. This is especially apparent when considering RQ2. The narrative bias was symmetric only when subjective probability was assessed in percent—that is, when narratives suggested a lower likelihood than the statistical base-rate, subjective probability decreased; it increased to the same extent when narratives suggested a higher likelihood compared to the statistical base-rate. Contrary to this finding, we observed a strong negativity bias on the risk measure—that is, a greater increase than decrease in risk perceptions due to narratives.

Previous research indicates that the 7-point rating scale is less sensitive to variations in objective probabilities (Betsch et al., 2011; Haase et al., 2013), which is supported by the present results: The 7-point rating scale was less able to map differences in the statistical base-rates than the percent estimates. In addition, the effect of narratives was smallest on this scale. If quantifying the effect of narratives on subjective probability is a goal, subjective probability should be assessed as percent estimates. Experiment 2 will therefore omit the 7-point rating scale.

In additional analyses, we found that individual differences in subjective numeracy play a differential role concerning the judgment of subjective probability both as a percent estimate and on a rating scale. Low subjective numeracy increased the influence of narratives when providing percent estimates, which matches previous findings (Dieckmann et al., 2009). For the 7-point rating scale, low subjective numeracy was related to less differentiation between statistical base-rates. Highly numerate subjects used the 7-point rating scale more broadly to differentiate between the 5% and 40% base-rate. In line with this, Peters and Bjalkebring (2014) found higher subjective numeracy to be related to better performance in a symbolic-number mapping task.

Experiment 2

Experiment 1 implicitly assumes that the relative frequency of narratives reporting the critical event influences risk perceptions. However, as stated in the introduction, encounter frequency theory and research on the ratio-bias suggest that the absolute number of narratives may drive this effect. In the current experimental paradigm, individuals would then perceive different risks when 8 of 20 narratives report VAE than when 4 of 10 do so. Thus, in this experiment we vary the absolute number of narratives while keeping the relative number of positive cases constant (RQ3).

In order to investigate whether the narrative bias is in part an experimental artifact, additional experimental conditions offer subjects the option to decide whether they want to view the narrative information in addition to the statistical information. This should communicate to subjects that the statistical information is sufficient to make a judgment (RQ4) and will also allow us to address the question whether narratives are an attractive source of information that are sought out even when statistical information is already available (RQ6). Further, in certain conditions we vary the sequence of statistical and narrative information to exclude recency as an alternative explanation (RQ5).

Method

The experimental set-up strongly resembled the first experiment.

Subjects and design. Subjects were recruited via Amazon Mechanical Turk and were paid US$1 (hourly wage: approx. US$4.14) through the Mechanical Turk payment system. Of the 515 individuals who clicked on the link to the survey, 479 completed the study. We excluded one individual who copied text from the page into a textbox, indicating that he or she did not read the instructions. In addition, we excluded three subjects who completed the survey in less than 5 minutes (M = 13 min 56 s, SD = 6 min), which falls below the minimum completion-time. Finally, 11 subjects indicated that they had previously participated in a similar study and were therefore excluded from the sample. Thus, analyses were calculated with a sample of N = 464 subjects, with *n*s for individual conditions ranging from 24 to 31. Nearly half of the sample identified as female (n = 230, 49.6%) and the mean age (SD) was 32.65 (10.85) years.

Subjects were randomly assigned to 16 conditions, resulting from a 2 × 2 × 3 between-subjects design plus four additional conditions described below (Figure 3). The main design is constituted by the following factors: 2 (sequence of dependent variables: risk perception followed by subjective probability and vice versa) × 2 (sample size: 10 vs. 20 cases) × 3 (relative frequency of narratives reporting adverse events: 10% vs. 20% vs. 40%). The statistical base-rate information was equal in all conditions (20%). Additionally, we assessed subjects' numeracy.

In order to test whether the narrative bias occurs due to a conversational norm indicating that all information provided must be relevant, we added two cells in which reading the narratives was optional, in contrast to all the above mentioned conditions in which reading the narratives was required. We did so in a 2 × 2 between-subjects subdesign, using two cells from the design reported above: 2 (relative frequency of adverse events: 10% vs. 40%) × 2

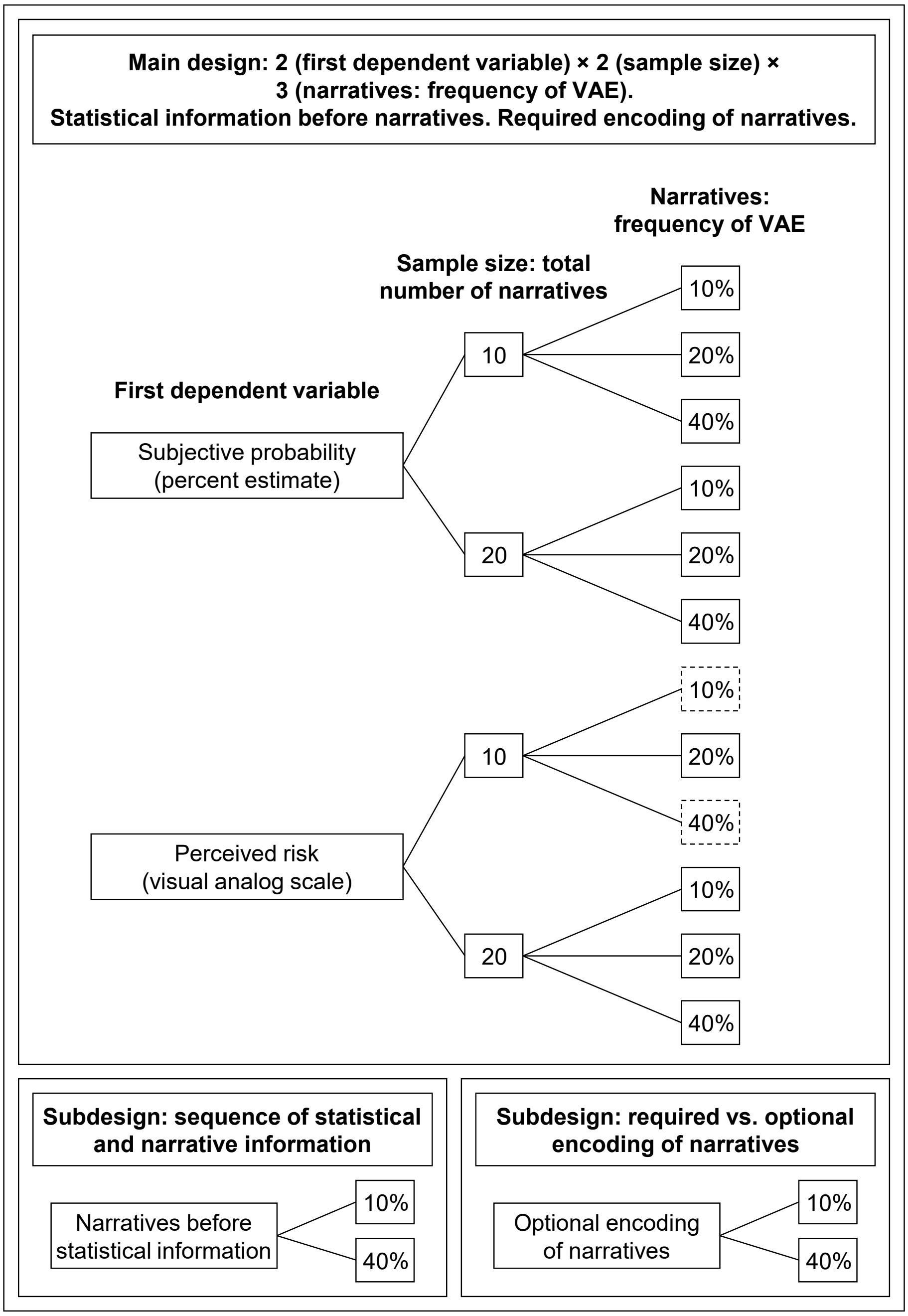

Figure 3. Design of Experiment 2. Top of the figure shows the main 2 × 2 × 3 design, the bottom shows both subdesigns. Dashed borders indicate cells from the main design used for comparison with the subdesigns.

(encoding of narrative information: required vs. optional). If an interaction suggests that the narrative bias disappears when encoding of the narratives is optional rather than required, we can assume that at least part of the narrative bias occurs due to the subjects' tendency to view all materials presented by the experimenter as relevant. For economic reasons, we decided to use only the risk measure as a dependent variable.

In order to test whether the narrative bias occurs due to a recency effect, we added two additional conditions in which we varied the order of the statistical and narrative information. Thus, the resulting 2 × 2 between-subjects subdesign was constituted by the following factors: 2 (relative frequency of adverse events: 10% vs. 40%) × 2 (sequence of information: statistic–narratives and vice versa). If an interaction suggests that the narrative bias disappears when the statistical information is presented after the narratives and before the dependent variables, we can assume that at least part of the narrative bias occurs due to a recency effect. Again, we used only the risk measure.

Procedure. As in Experiment 1, subjects read about the disease, the vaccine recommendation, and the statistical likelihood of VAE in written and graphic form. They were then presented with the narrative information within a simulated bulletin board. Finally, we collected dependent variables, manipulation checks, and control variables.

Statistical probability of adverse events. As in the first experiment, the statistical probability of VAE was stated explicitly in percent and displayed by means of a pictograph and was fixed at 20% in all conditions. The narratives either matched, exceeded, or fell below the statistical information.

Manipulation of relative frequency of narratives reporting adverse events. In all conditions, either 1, 2 or 4 of 10 or 2, 4 or 8 of 20 narratives reported VAE (resulting in relative frequencies of 10%, 20% and 40%, respectively). All reported adverse events were categorized as mild (e.g., insomnia, fever, rash; as identified in a pretest, see Experiment 1). The remaining cases reported unproblematic vaccination experiences. As in Experiment 1, the narratives were of equal length, randomized in their sequence, and displayed one at a time.

Required vs. optional reading of narratives. Two conditions offered subjects the choice to either view the narrative information or skip the simulated bulletin board. Subjects were asked: "Next, you have the opportunity to read a number of posts from an online message-board where people share their personal experiences with the vaccine. Would you like to read the posts?" (yes or no). In all other conditions, subjects were informed that on the subsequent pages they will see "a number of posts from an online message-board where people share their personal experiences with the vaccine". The instructions asked them to read all messages carefully.

Sequence of statistical and narrative information. All subjects learned that their doctor provided them with the statistical information. In two conditions, the statistical information appeared after the narrative information. In all other conditions, the statistical base-rate information was provided first.

Dependent variables. As dependent variables, we assessed perceived risk and subjective probability in the same manner as in Experiment 1 (Table 1). As a measure for subjective probability, we asked for percent estimates. Half of the subjects judged the subjective probability of VAE first and then rated their perceived risk and vice versa for the other half.

Manipulation checks. We asked subjects to recall the initially stated base-rate of VAE (20%) as well as the number of narratives that reported adverse events (1, 2, or 4 of 10 or 2, 4, or 8 of 20). We asked for the number of narratives only if the subjects had either seen them by default or if they had decided to read them.

Numeracy. In this experiment, we employed a more objective measure of numeracy—a combination of the 3-item scale by Schwartz, Woloshin, Black, and Welch (1997) and the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). The seven items involve short mathematical quizzes (e.g., "Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws, how many times would this five-sided die show an odd number (1, 3 or 5)?", correct answer = 30).

Results

Manipulation check. Ninety-two percent of subjects correctly recalled that the statistical base-rate was 20% (M_{20} = 21.16, SD_{20} = 9.12); 4% reported that it was below 20%, whereas 4% reported a base-rate greater than 20%.

We assumed a correct recall of the number of narratives if the recalled number was plus/minus one. For the condition in which 1 narrative reported VAE, 95.3% correctly recalled the absolute number (M_1 = 1.30, SD_1 = 1.32). In the 2 cases condition, 93.2% (M_2 = 2.45, SD_2 = 1.77), in the 4 cases condition 85.1% (M_4 = 4.71, SD_4 = 4.12), and in the 8 cases condition 41.4% (M_8 = 8.97, SD_8 = 11.78) correctly recalled the absolute frequency of narratives reporting VAE.

Numeracy. The sum score of all correctly solved numeracy items constitutes the numeracy score (potential range 0–7). The mean numeracy score (3.50, SD = 1.79) did not differ across conditions (all η_p^2s in a 3 × 2 × 2 ANOVA were ≤ .003, all ps ≥ .32).

Subjective probability and risk perception. The goal of this experiment was to assess whether the narrative bias occurs due to the relative or absolute frequency of narratives reporting VAE (RQ3). A main effect showing an increase with sample size (number of messages on the bulletin board) would indicate that the absolute number of narratives reporting VAE (possibly as well as the relative number) influences the dependent variables, because the absolute number of narratives is higher in the 20 cases condition (2, 4, 8) than in the 10 cases condition (1, 2, 4).

For all analyses, we calculated regression analyses with standardized, continuous predictors. Two separate linear regressions were calculated, predicting subjective probability (percent estimates) and perceived risk (visual analog scale). We again used only the subsamples in which the respective dependent variable was assessed first to exclude carry-over effects. Interactions were calculated as the mathematical products of the standardized predictors (Cohen et al., 2003). In a first regression, we entered the manipulated factors and their interaction. In a second regression, we added numeracy and the interactions of the factors with numeracy.[2]

Table 4 displays the results of the regression analyses. Subjective probability tended to be influenced by the relative frequency of positive narratives when the sample was small (10 cases) and was not influenced when it was large (20 cases), as indicated by an almost significant interaction of narratives and sample size ($\beta = -.14$). The effect was somewhat weaker when numeracy and the respective interactions were also entered into the regression. No other effects were significant.

Perceived risk was a function of the relative frequency of narratives reporting VAE ($\beta = .38$), with subjects in conditions with a higher relative frequency perceiving higher vaccination risks. The sample size did not affect perceived risk (Figure 4).[3]

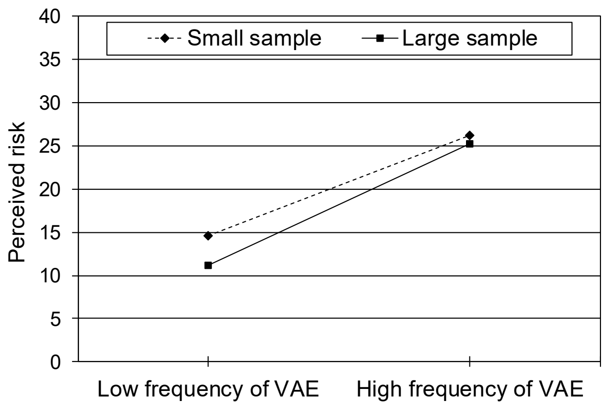

Figure 4. Unstandardized simple slopes of frequency of VAE predicting perceived risk for the small and the large sample.

---

[2] See Appendix B for a full correlation matrix.

[3] Simple slope figures were created with an Excel plotting sheet by Winnifred Louis, available at: http://www2.psy.uq.edu.au/ uqwloui1/, last accessed on March 3, 2015.

Table 4
Subjective probability and perceived risk as a function of sample size, frequency of VAE, and numeracy (Experiment 2).

Subjective probability (percent estimate) n = 181	β	p	β	p
Sample size (10 vs. 20%)	−.04	.61	−.03	.70
Narratives: frequency of VAE (10% vs. 20% vs. 40%)	.10	.16	.09	.26
Sample size × narratives	−.14	.06	−.13	.09
Numeracy			−.02	.80
Numeracy × sample size			.05	.50
Numeracy × narratives			−.10	.19
Numeracy × sample size × narratives			.11	.15
R^2	.03		.06	
Perceived risk (visual analog scale) n = 168	**β**	**p**	**β**	**p**
Sample size (10 vs. 20%)	−.07	.36	−.07	.35
Narratives: frequency of VAE (10% vs. 20% vs. 40%)	**.38**	**< .001**	**.37**	**< .001**
Sample size × narratives	.04	.60	.04	.62
Numeracy			**−.14**	**.05**
Numeracy × sample size			.02	.76
Numeracy × narratives			**−.14**	**.05**
Numeracy × sample size × narratives			.13	.06
R^2	.29		0.30	

Note. Standardized betas (β) and respective p-values of significant effects are shown in boldface.

Subjects with high numeracy generally perceived lower risk (β = −.14). Two interaction effects qualified this main effect. For highly numerate subjects, there was a weaker narrative bias, whereas the bias was stronger for subjects with low numeracy (Figure 5). The almost significant three-way interaction is displayed in Figures 6A and 6B, which show that there was no narrative bias for highly numerate subjects when the sample size was small.

In order to rule out the possibility that the lack of an effect of sample size was due to subjects not encoding the larger number of narratives as carefully as the small number, we analyzed reading times. We conducted two separate 3 × 2 × 2 ANOVAs with the total amount of time spent reading the narratives and the average amount of time per narrative as respective

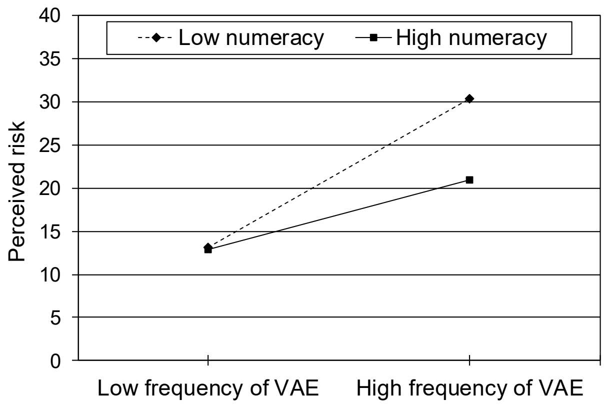

Figure 5. Unstandardized simple slopes of frequency of VAE predicting perceived risk for 1 *SD* below and 1 *SD* above the mean of numeracy.

dependent variables. Subjects took almost exactly twice as long to encode 20 narratives, M_{20} = 3 min 31 s, SD_{20} = 2 min 5 s, as compared to 10 narratives, M_{10} = 1 min 46 s, SD_{10} = 57 s; $F(1, 337) = 60.96$, $p < .001$, $\eta_p^2 = .15$). There were no other significant effects (Fs ≤ 1.11). Correspondingly, average reading times per narrative were virtually identical in the small and large sample conditions (M_{10} = 11 s, SD = 6 s; M_{20} = 11 s, SD = 8 s) and did not differ across any conditions (Fs < 1). Additionally, adding either time variable had no effect on the regression models.

Thus, the results show that subjects encoded small and large samples equally well. As there was no main effect of sample size on either dependent variable, this indicates that

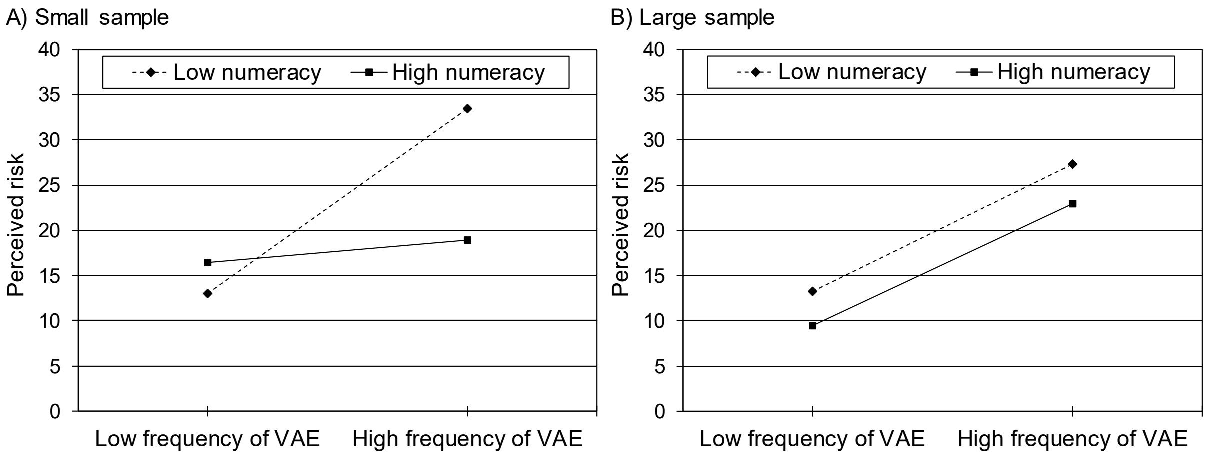

Figure 6. Unstandardized simple slopes of frequency of VAE predicting perceived risk for 1 *SD* below and 1 *SD* above the mean of numeracy, separate for the small and the large sample, illustrating the three-way interaction.

subjects extracted a relative frequency representation from the narratives. This relative frequency, rather than the absolute number of cases, drives the narrative bias (RQ3).

Carry-over effects. Again, to test for carry-over effects, we calculated the same two regression models for the respective other subsample. Considering the whole risk construct first led to a significant narrative bias on subjective probability estimates ($\beta = .26$, $p = .001$). An almost significant interaction with numeracy indicated that this bias was more pronounced for low numerate individuals ($\beta = -.14$, $p = .060$; all other ts < 1; $R^2 = .10$, $F(7, 160) = 2.48$, $p = .019$).

A prior probability estimate, on the other hand, decreased the influence of narrative variation on the risk judgment ($\beta = .20$, $p = .009$) and rendered all further effects non-significant (all other ts ≤ 1.81; $R^2 = .09$, $F(7, 173) = 2.34$, $p = .026$).

Required vs. optional encoding of narratives. Of the subjects who had a choice to read the narratives, 78.2% ($n = 43$) decided to do so. We conducted a 2 (frequency of VAE: 10% vs. 40%) × 2 (encoding of narrative information: required vs. optional) ANOVA with perceived risk as dependent variable. The analysis revealed only a strong narrative bias, $F(1, 92) = 19.11$, $p < .001$, $\eta_p^2 = .17$, all other Fs < 1. Consequently, we assume that narrative information affects risk perceptions irrespective of whether it had to be read or was freely chosen (RQ4).

Sequence of statistical and narrative information. In order to test for recency effects, we conducted a 2 (frequency of VAE: 10% vs. 40%) × 2 (sequence of statistical and narrative information) ANOVA with perceived risk as the dependent variable. We found a strong main effect for the relative frequency of VAE, $F(1, 109) = 16.07$, $p < .001$, $\eta_p^2 = .13$. All other effects were nonsignificant, Fs ≤ 1.7. Thus, the narrative bias occurred regardless of whether statistical base-rate information was provided before or after the narratives were encoded. This indicates that the narrative bias is not caused by a recency effect (RQ5).

Intention to get vaccinated. For all subjects ($N = 464$) the correlation between perceived risk and intention to get vaccinated is r = −.41 ($p < .001$). Subjective probability and intention are not correlated ($r = .01$, n.s.). Because we observed muticollinearity in Experiment 1, Appendix C presents correlations between both measures and intention. When perceived risk was assessed first, it correlates with subjective probability. However, in both subsamples only perceived risk predicts behavioral intentions.

Discussion

In two experiments we found that the biasing influence of narrative information on risk perception is in part a function of the dependent measure used to assess it. Narratives had the largest effect on a non-numerical risk measure, whereas two scales for subjective probability reflected mostly statistical variations. This stresses the importance of differentiating between the constructs risk and probability. Further, two-way carry-over effects between the respective measures indicated that the use of all scales was context dependent, for example, considering the risk construct first increased the influence of narrative information on subsequent probability judgments. Additionally, the risk measure was the best predictor of behavioral intentions, and only for the risk measure did we observe a negativity bias. Moreover, results indicate that subjects extracted a representation of relative frequency from the narratives, as changing the absolute number of single events while keeping their relative number constant did not change the narrative bias. Subjective and objective numeracy had opposing and somewhat weak effects on judgments. Finally, the option to freely choose whether to read the narrative information did not affect the narrative bias in any way. In addition, we found no indication of a recency effect as an explanation for the narratives' influence.

Issues of measurement. Regarding the task dependence of the narrative bias, three aspects of measuring risk perception must be considered: the representation on which a judgment is based, the scale used for assessment, and the context in which the scale is used.

Various theoretical approaches propose that risk judgments rely on two distinct representations or processes. These theories make diverse yet conceptually related differentiations between cognitive vs. affective risk evaluations, a belief in objective probabilities vs. an intuitive perception of risk, and verbatim vs. gist representations. The two respective components are understood to be distinct but may interact in the reasoning process, which moves along a continuum between them (Loewenstein et al., 2001; Reyna, 2008, 2012; Slovic et al., 2004; van Gelder, de Vries, & van der Pligt, 2009; Windschitl, Martin, & Flugstad, 2002).

The scales we used differ along at least two dimensions. First, whereas the rating scale offers only seven discrete categories for judgment, the percent format and the visual analog risk scale allow for quasi-continuous estimates, that is, 101 discrete categories, as responses were restricted to integers. This difference in resolution provides the latter scales with a natural advantage in terms of sensitivity to changes in subjective probability (Haase et al., 2013). Second, while the percent format is purely numeric, the rating scale and risk measure provide verbal labels. It has been argued that numeric probability measures induce rule-based reasoning in individuals and elicit beliefs in objective probability, whereas verbal scales lead to a more associative reasoning style and elicit rather intuitive thoughts about an uncertain prospect. These intuitive beliefs entail more than just a maximally accurate representation of likelihood. Rather, they also include notions of the value of a prospect, affective reactions to it, and its meaning in a given situation—all of which may make them more comparable to real-life situations. Accordingly, verbal probability scales have been shown to be more sensitive to context and framing effects as well as to be better predictors of preferences, behavioral intentions, and behavior than numeric scales. The risk measure extends this idea on an explicit conceptual level, as risk by definition encompasses more than mere probability. In addition, risk measures have been found to perform even better in predicting behavior (Baghal, 2011; Weinstein et al., 2007; Windschitl, 2002; Windschitl & Wells, 1996).

Finally, the interpretation of a question and the use of a response format have been shown to be affected by the context such as a preceding question (Schwarz, 1999, 2007). Building on these premises, we suggest that judgments in research on biased risk perception are in part task-dependent (RQ1). Subjects base their estimates on beliefs in objective likelihood and intuitive risk representations and engage in rule-based and associative reasoning styles. The degree to which these two representations inform the judgment and the manner in which they are weighed and processed are in part a function of the response scale provided, as well as prior elicitations of related constructs.

In line with this notion, narrative and statistical information affected the three dependent variables differently. Judgments on the 7-point rating scale were not influenced by variations in the narrative information (Experiment 1), which can partly be explained by the scale's low sensitivity, as even the explicitly stated statistical probabilities of 5% and 40% were mapped very close to each other on the rating scale. However, the verbal qualifiers of this scale make judgments prone to reflecting not only a likelihood representation but also other aspects of the uncertain prospect, for example, the severity of VAE (Weber & Hilton, 1990), which may have masked the effect of the narrative manipulation. Indeed, controlling for perceived severity ($\beta = .19$, $p = .024$) in the regression model not only significantly increased the amount of explained variance, from $R^2 = .39$ to $R^2 = .43$, $F(1, 85) = 5.31$, $p = .024$, but also rendered the variation in narratives a significant predictor of subjective probability ($\beta = .17$, $p = .046$; all other effects unchanged). We assume that these subjects attempted to provide judgments which, for the most part, reflect their beliefs in objective probability, as this was the first question asked. In contrast to this, estimates by subjects who had first considered the whole risk construct showed a clear narrative bias, indicating that subjects' interpretation of the 7-point rating scale—as a pure probability measure vs. a general risk measure—varies as a function of contextual factors.

The percent format, in comparison, elicits responses that are almost exclusively expressions of rule-based reasoning processes concerning numeric probabilities. The effect of variations in narrative information was smaller (Experiment 1) or negligible (Experiment 2) and symmetric as compared to the effect on risk judgments. Further, adding severity to the regression model had no effect in either of the experiments. Subjects encoded the likelihood of VAE in percent and were later asked for an estimate in percent. Thus, the format might have cued the retrieval of this specific information rather than a subjective representation of probability. However, even the percent format is not fully resistant to context effects—in Experiment 2, asking for a general risk judgment beforehand led to a narrative bias.

Finally, we observed the strongest narrative bias on the visual analog risk scale. In Experiment 1, narratives had a stronger effect on risk perceptions than the statistic. In Experiment 2, only risk perceptions were affected by narrative information. Further, perceived severity proved a strong predictor of perceived risk and improved the model to a large degree in both experiments (Experiment 1: $\beta = .31$, $p = .001$; from $R^2 = .30$ to $R^2 = .39$, $F(1, 85) = 12.17$, $p = .001$; Experiment 2: $\beta = .48$, $p < .001$; from $R^2 = .20$ to $R^2 = .42$, $F(1, 159) = 61.36$, $p < .001$; all other effects unchanged). As individuals expressed more than just a likelihood representation in their risk judgments, these estimates might be especially susceptible to contextual factors. Accordingly, asking for a probability estimate first increased the influence of statistical information on perceived risk (Experiment 1) and decreased the effect of narrative variation (both experiments). Still, risk estimates did not represent merely an analytic integration of likelihood and value. Additionally controlling for subjective probability estimates in the regression models eliminated the effect of statistical variation (Experiment 1) but not the effect of narratives on risk judgments (both experiments).

In line with previous research and our reasoning thus far, we found that the risk measure predicted behavioral intentions best. Decisions and behavior under risk, of course, have more antecedents than just the likelihood of a given outcome. Thus, a measure that elicits more than this likelihood representation will consequently lead to superior predictions. Our findings regarding the symmetry of the narrative bias (RQ2) lend further support to this explanation. When asked to provide percent estimates, subjects engaged in rule-based integration akin to a calculation, which, since the presented frequencies were symmetric, resulted in a symmetric bias. A more intuitive risk measure, on the other hand, led to a clear negativity bias. One explanation for the stronger impact of negative information is that it possesses greater diagnostic value. Consider the potential cost of ignoring a danger versus mistakenly missing out on a benefit. If judgments of perceived risk are more relevant for actual behavior, it would make sense to assign negative information more weight. However, when the judgment process follows a normative understanding of mathematics, equal numbers will receive equal weights (Baumeister et al., 2001; Siegrist & Cvetkovich, 2001; Skowronski & Carlston, 1989).

Narratives as a source of probabilistic information. The narrative information provided subjects with exemplars of the occurrence and non-occurrence of an uncertain outcome, that is, VAE. The encoding of such event frequencies is a predominantly automatic and accurate process (Hasher & Zacks, 1979; Zacks & Hasher, 2002). Accordingly, the manipulation checks showed that subjects were able to track the number of narratives reporting VAE, although there was some decline in accuracy when this number was larger. Nonetheless, results indicate that individuals perceived the absolute frequency of an uncertain event, yet extracted a relative frequency representation for subsequent risk judgments, as a change in total sample size did not affect the biasing influence of narrative information (RQ3).

This finding stands in contrast to some existing literature. Research on the ratio bias, for instance, would have predicted that subjects perceive a higher likelihood or risk when 8 of 20 narratives report VAE rather than 4 of 10, as they concentrate on the absolute frequency of the focal event and fail to take into account the total number of events (Denes-Raj & Epstein, 1994; Reyna & Brainerd, 2008). However, the occurrence of the ratio-bias appears to depend on within-subjects comparisons (Lefebvre, Vieider, & Villeval, 2010), whereas the present study used a between-subjects design.

Similarly, Obrecht et al. (2009) employed a within-subjects design in their encounter frequency account. However, while 4 of 10 and 8 of 20 narratives would result in equal ratios of positive and negative encounters, their theory hinges on the idea that the statistic enters the judgment process as simply one more instance indicating either the occurrence or non-occurrence of an event. This extra piece of information leads to differing ratios and, subsequently, differing predictions of perceived probability. There may be some merit to this theory if the statistic offers clear-cut evidence, that is, the likelihood is extremely high or low. However, a probability of 20% clearly indicates a certain amount of risk; one would be hard-pressed to simply interpret it as a non-occurrence of an event because it is numerically below 50%.

Taken together, our findings indicate that subjects interpreted the narratives as representing one sample of events conveying a probability and the statistic as another such sample. We did observe two almost significant interaction effects of sample size: First, probability estimates were biased by the narrative information only when the sample was small. This might be due to a more accurate tracking of event frequencies when only ten exemplars were presented. Second, individuals high in numeracy showed no narrative bias on the risk measure in a small sample as compared to a large one. This might indicate that these subjects did in fact consider sample size in their judgments, as a larger sample of 20 events does have a higher diagnostic value than a smaller sample of 10 cases. As both effects were barely significant and rather small, we do not believe that they impede our previous reasoning.

Numeracy. We observed opposing effects of numeracy on the respective dependent measures in the two experiments. It is important to note, however, that we employed two different instruments to assess numeracy. The Subjective Numeracy Scale in Experiment 1 measures self-assessed ability and preferences to understand and apply numbers, whereas the combined test in Experiment 2 objectively assesses the ability to perform mathematical operations with percentages and proportions. Even though the former measure was developed to serve as a proxy for objective performance tests, inconsistent results have been observed previously. In addition, it has been shown recently that subjective and objective numeracy scales share only a limited amount of variance and differ in their predictions of various biases (Hess, Visschers & Siegrist, 2011; Liberali, Reyna, Furlan, Stein, & Pardo, 2012; Peters & Bjalkebring, 2014).

In Experiment 1, only the measures of subjective probability were affected by subjective numeracy, although in different ways. The statistical variation led to more extreme values on the 7-point rating scale for individuals higher in subjective numeracy, which implies a relation to the way individuals use a given scale for mapping likelihood. In contrast, higher subjective numeracy resulted in percent estimates that were less biased by narrative information, indicating significance for the process of integrating probabilities. In Experiment 2, we observed that objective numeracy decreased risk perceptions in general and moderated the narrative bias. Percent estimates of subjective probability were moderated only by objective numeracy when risk was assessed first, leading to a narrative bias in the first place.

The question of whether a unitary construct underlies the various observed effects or lack thereof has been an ongoing debate (e.g., Nelson, Moser, & Han, 2013; Reyna, Nelson, Han, & Dieckmann, 2009). Subjective numeracy appears to reflect motivation and confidence regarding the use of numerical information, that is, it relates to how people approach a task.

Thus, the lack of an effect on the risk measure in Experiment 1 is in line with our thinking that the risk judgments were approached in a more intuitive rather than rule-based manner. Of course, the processing of proportions is still relevant in the actual formation of a risk judgment, as the results regarding the sample size in Experiment 2 clearly demonstrate. Thus, the effect of objective numeracy, which appears to drive actual number operations, does not contradict the findings from Experiment 1. These findings support the notion of two related but not identical numeracy constructs (Peters & Bjalkebring, 2014).

Experimental artifact and the motivation to understand risky outcomes. We addressed two potential experimental artifacts as alternative explanations for the narrative bias that both relate to conversational norms. First, encoding of the individuating information is typically mandatory in experiments like this. Thus, subjects might assume that it is relevant to the task at hand and thus only use it for this reason (Grice, 1975). We offered subjects a choice and found that, when reading the narratives was optional, the narrative bias occurred just as strongly ($r = .39$, $p = .009$) as when it was required ($r = .44$, $p = .001$; Fisher's $z = -0.27$, $p = .787$; RQ4).

Second, it has been argued that, when two conflicting pieces of information are presented, conversational norms indicate that the more informative and thus more important information is typically placed last (Krosnick at al., 1990). As the individuating information is generally presented after a base-rate in most related research, we varied the sequence of narrative and statistical information to exclude the possibility that the narrative bias is driven by recency. The results show that the narrative bias occurred independently of the narratives' position in the sequence of information (RQ5).

Thus, we found no evidence in favor of the hypothesis that the narrative bias is based on adherence to conversational norms. On the other hand, we also cannot rule out their significance. Even in the optional encoding conditions, the narratives were still provided by the experimenter. Therefore, it is still possible that the subjects assumed that they are relevant for their judgments. The fact that nearly 80% of subjects chose to read the narratives in addition to the statistical information might support this interpretation.

Ultimately, the importance of conversational norms in research on cognitive biases due to irrelevant information cannot be ascertained conclusively in an experiment, as all information is always provided by the experimenter. However, the motivation of the majority of subjects to read the narratives in this experiment might be explained differently. Huber, Wider, and Huber (1997) found that individuals are often more interested in the outcomes of risky situations rather than the likelihood of negative outcomes. This behavior appears to have biological roots, as fish and birds behave in a similar manner. The *costly information hypothesis* states that when information (e.g., concerning vaccine safety) is too costly to be acquired personally (because it might harm the organism), animals will take advantage of the relatively low-cost information provided by others (Boyd & Richerson, 1985; Webster & Laland, 2008).

Taken together, narratives appear to represent an attractive source of information, as they deliver details on specific outcomes of risky situations (RQ6). Whether people deem them to be relevant for their risk assessments due to their specificity (see for example Bar-Hillel, 1980) or simply because they have been presented cannot be answered decisively. However, the fact that in the present research both procedural manipulations had no effect at all on risk perceptions renders experimental artifacts as sole drivers of the narrative bias less likely.

Limitations. One goal of this study was to investigate the biasing effect of narrative information as a function of different self-report measures. However, a general concern in this line of research is a potential lack of external validity due to the use of hypothetical scenarios, meaning that presenting a bias incurs no cost to the subjects. Indeed, some biases have been found to decrease or disappear when payment is dependent on performance (Lefebvre et al., 2010). Future research should strive to substantiate the present findings by incentivizing non-biased judgments (Hertwig & Ortmann, 2001).

The data in Experiment 2 were collected through Amazon Mechanical Turk. The data quality may be affected by increased error variance if a greater number of subjects did not take their participation seriously. For this reason, we eliminated subjects whose time to complete the study suggested non-serious participation. Further, previous research has shown an advantage of Mechanical Turk samples in heterogeneity compared to the standard student sample as well as sufficient quality according to the psychometric publication standards (Buhrmester, Kwang & Gosling, 2011; Mason & Suri, 2012; Paolacci, Chandler & Ipeirotis, 2010).

Conclusion. The biasing effect of a small sample of single-case narratives is in part dependent on the measure used to assess it. Scales that gauge the likelihood dimension of an uncertain prospect are least affected. Narratives have the strongest effect on measures of a more extensive risk representation, which may entail a value dimension as well as other aspects, for example, an affective appraisal of the uncertain event. This more comprehensive idea of a risk appears to be of greater importance in guiding decisions and behavior than a strict likelihood representation. On the other hand, judgments of subjective probability as well as of perceived risk appear to be ad hoc constructions and are therefore susceptible to wording and framing effects. Attempts to predict preferences and behavior should therefore be viewed cautiously, as risk perception might change from the time of assessment to the time of action due to a change of context. Further, systematic review or overview articles need to not only specify the exact wording and scale format of the instruments that were used in the original research but also take concomitant assessments of related constructs into account. Individuals do extract a representation of likelihood from single-case exemplars. This representation drives the narrative bias. However, the effect is much smaller on scales that assess only the perceived likelihood as compared to a measure of a broader and more intuitive concept of risk.

Taken together, these results underline the important conceptual distinction between judgments of subjective probability and perceived risk. When individuals judge a risk, they take many other aspects of the risky prospect into account than merely its probability of occurrence. Next to its severity, characteristics such as the voluntariness of, knowledge about, and control over risk play a role in risk perception. Affective reactions, personal susceptibility, and the source of a risk are additional potentially relevant factors. The measures we investigated reflect these different concepts to differing degrees. For instance, the perceived severity of VAE had no effect on percent estimates of probability, a small effect on a verbal probability measure, and a strong effect on a measure of risk.

However, the narrative bias we observed cannot be attributed to any of these additional aspects of risk, as they were either held constant across subjects, for example, the emotionality of narratives, or controlled for through randomization. Further, Experiment 2 clearly indicates that the bias was driven by a representation of relative frequency, that is, probability. This representation had small effects on measures of probability, that is, instruments that are designed to solely assess this very representation. On the other hand, it had large effects on an inherently multidimensional measure of risk. Thus, we conclude that the relationship between representations of subjective probability and perceived risk is not yet fully understood. Future research should strive to understand what role likelihood representations play in the formation of risk perceptions.

References

Baghal, T. (2011). The measurement of risk perceptions: The case of smoking. *Journal of Risk Research, 14*(3), 351–364. https://doi.org/10.1080/13669877.2010.541559

Ball, D. J., & Watt, J. (2013). Further thoughts on the utility of risk matrices. *Risk Analysis, 33(11)*, 2068–2078. https://doi.org/10.1111/risa.12057

Bar-Hillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica, 44*(3), 211–233. https://doi.org/10.1016/0001-6918(80)90046-3

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology, 5*(4), 323–370. https://doi.org/10.1037//1089-2680.5.4.323

Beach, L. R., & Phillips, L. D. (1967). Subjective probabilities inferred from estimates and bets. *Journal of Experimental Psychology, 75*(3), 354–359. https://doi.org/10.1037/h0025061

Betsch, C., Brewer, N. T., Brocard, P., Davies, P., Gaissmaier, W., Haase, N., . . . Stryk, M. (2012). Opportunities and challenges of Web 2.0 for vaccination decisions. *Vaccine, 30*(25), 3727–3733. https://doi.org/10.1016/j.vaccine.2012.02.025

Betsch, C., Renkewitz, F., & Haase, N. (2013). Effect of narrative reports about vaccine adverse events and bias-awareness disclaimers on vaccine decisions: A simulation of an online patient social network. *Medical Decision Making, 33*(1), 14–25. https://doi.org/10.1177/0272989X12452342

Betsch, C., Ulshöfer, C., Renkewitz, F., & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making, 31*(5), 742–753. https://doi.org/10.1177/0272989X11400419

Borgida, E., & Nisbett, R. (1977). The differential impact of abstract vs. concrete information on decisions. *Journal of Applied Social Psychology, 7*(3), 258–271. https://doi.org/10.1111/j.1559-1816.1977.tb00750.x

Boyd, R., & Richerson, P. J. (1985). *Culture and the evolutionary process.* Chicago: University of Chicago Press.

Brewer, N. T., Chapman, G. B., Gibbons, F. X., Gerrard, M., McCaul, K. D., & Weinstein, N. D. (2007). Meta-analysis of the relationship between risk perception and health behavior: The example of vaccination. *Health Psychology, 26*(2), 136–45. https://doi.org/10.1037/0278-6133.26.2.136

Brun, W. (1994). Risk perception: Main issues, approaches and findings. In G. Wright & P. Ayton (Eds.), *Subjective probability* (pp. 295–320). Chichester, England: Jon Wiley & Sons.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk: A new source of inexpensive, yet high-quality, data? Perspectives on *Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making, 7*(1), 25–47. Retrieved from http://journal.sjdm.org/11/11808/jdm11808.pdf

de Wit, J. B. F., Das, E., & Vet, R. (2008). What works best: Objective statistics or a personal testimonial? An assessment of the persuasive effects of different types of message evidence on risk perception. *Health Psychology, 27*(1), 110–115. https://doi.org/10.1037/0278-6133.27.1.110

Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment. *Journal of Personality and Social Psychology, 66*(5), 819–829. https://doi.org/10.1037/0022-3514.66.5.819

Dieckmann, N. F., Slovic, P., & Peters, E. M. (2009). The use of narrative evidence and explicit likelihood by decisionmakers varying in numeracy. *Risk Analysis, 29*(10), 1473–1488. https://doi.org/10.1111/j.1539-6924.2009.01279.x

Eiser, J. R. (1994). Risk judgements reflect belief strength, not bias. *Psychology and Health, 9*(3), 197–199. https://doi.org/10.1080/08870449408407479

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making, 25*(4), 398–405. https://doi.org/10.1177/0272989X05278931

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the Subjective Numeracy Scale. *Medical Decision Making, 27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Fox, S., & Duggan, M. (2013). *Health online 2013*. Retrieved from PEW Internet & American Life Project website: http://www.pewinternet.org/2013/01/15/health-online-2013

Gardoni, P., & Murphy, C. (2013). A scale of risk. *Risk Analysis, 34*(7), 1208–1227. https://doi.org/10.1111/risa.12150

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and biases: The psychology of intuitive judgment*. Cambridge, England: Cambridge University Press.

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making, 9*(1), 1–14. Retrieved from http://journal.sjdm.org/13/131029/jdm131029.pdf

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.) *Syntax and semantics, 3: Speech acts* (pp. 41–58). New York: Academic Press.

Haase, N. & Betsch, C. (2012). Parents trust other parents: Lay vaccination narratives on the web may create doubt about vaccination safety. *Medical Decision Making, 32*(4), 645. https://doi.org/10.1177/0272989X12445286.

Haase, N., Renkewitz, F., & Betsch, C. (2013). The measurement of subjective probability: Evaluating the sensitivity and accuracy of various scales. *Risk Analysis, 33*(10), 1812–1828. https://doi.org/10.1111/risa.12025

Hasher, L., & Zacks, R. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*(3), 356–388. https://doi.org/10.1037/0096-3445.108.3.356

Hertwig, R., & Ortmann, A. (2001). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences, 24*(03), 383–403.

Hess, R., Visschers, V. H. M. & Siegrist, M. (2011). Risk communication with pictographs: The role of numeracy and graph processing. *Judgment and Decision Making, 6*(3), 263–274. Retrieved from http://journal.sjdm.org/11/10630/jdm10630.pdf

Hinyard, L. J., & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior, 34*(5), 777–792. https://doi.org/10.1177/1090198106291963

Huber, O., Wider, R., & Huber, O.W. (1997), Active information search and complete information presentation in naturalistic risky decision tasks. *Acta Psychologica, 95*(1), 15–29. https://doi.org/10.1016/S0001-6918(96)00028-5

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine, 28*(7), 1709–1716. https://doi.org/10.1016/j.vaccine.2009.12.022

Kata, A. (2012). Anti-vaccine activists, Web 2.0, and the postmodern paradigm – An overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine, 30*(25), 3778–3789. https://doi.org/10.1016/j.vaccine.2011.11.112

Keller, C., Siegrist, M., & Visschers, V. (2009). Effect of risk ladder format on risk perception in high- and low-numerate individuals. *Risk Analysis, 29*(9), 1255–1264. https://doi.org/10.1111/j.1539-6924.2009.01261.x

Knapp, P., Gardner, P. H., Raynor, D. K., Woolf, E., & McMillan, B. (2010). Perceived risk of tamoxifen side effects: A study of the use of absolute frequencies or frequency bands, with or without verbal descriptors. *Patient Education and Counseling, 79*(2), 267– 271. https://doi.org/10.1016/j.pec.2009.10.002

Krosnick, J. A., Li, F., & Lehman, D. R. (1990). Conversational conventions, order of information acquisition, and the effect of base rates and individuating information on social judgments. *Journal of Personality and Social Psychology, 59*(6), 1140–1152. https://doi.org/10.1037/0022-3514.59.6.1140

Lee, S. W. S., Schwarz, N., Taubman, D., & Hou, M. (2010). Sneezing in times of a flu pandemic: Public sneezing increases perception of unrelated risks and shifts preferences for federal spending. *Psychological Science, 21*(3), 375–377. https://doi.org/10.1177/0956797609359876

Lefebvre, M., Vieider, F. M., & Villeval, M. C. (2010). The ratio bias phenomenon: Fact or artifact? *Theory and Decision, 71*(4), 615–641. https://doi.org/10.1007/s11238-010-9212-9

Liberali, J. M., Reyna, V. F., Furlan, S., Stein, L. M., & Pardo, S. T. (2012). Individual differences in numeracy and cognitive reflection, with implications for biases and fallacies in probability judgment. *Journal of Behavioral Decision Making, 25*(4), 361–381. https://doi.org/10.1002/bdm.752

Loewenstein, G. F., Weber, E. U., Hsee, C. K., & Welch, N. (2001). Risk as feelings. *Psychological Bulletin, 127*(2), 267–286. https://doi.org/10.1037//0033-2909.127.2.267

Mason, W., & Suri, S. (2012). Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods, 44*(1), 1–23. https://doi.org/10.3758/s13428-011-0124-6

Nelson, W. L., Moser, R. P., & Han, P. K. J. (2013). Exploring objective and subjective numeracy at a population level: Findings from the 2007 Health Information National Trends Survey (HINTS). *Journal of Health Communication, 18*(2), 192–205. https://doi.org/10.1080/10810730.2012.688450

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition, 37*(5), 632–643. https://doi.org/10.3758/MC.37.5.632

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making, 5*(5), 411–419. Retrieved from http://journal.sjdm.org/10/10630a/jdm10630a.pdf

Peters, E. (2008). Numeracy and the perception and communication of risk. *Annals of the New York Academy of Sciences, 1128*, 1–7. https://doi.org/10.1196/annals.1399.001

Peters, E., & Bjalkebring, P. (2015). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology, 10*8(5), 802–822. https://doi.org/10.1037/pspp0000019

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making, 28*(6), 850–865. https://doi.org/10.1177/0272989X08327066

Reyna, V. F. (2012). Risk perception and communication in vaccination decisions: A fuzzy-trace theory approach. *Vaccine, 30*(25), 3790–3797. https://doi.org/10.1016/j.vaccine.2011.11.070

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences, 18*(1), 89–107. https://doi.org/10.1016/j.lindif.2007.03.011

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin, 135*(6), 943–973. https://doi.org/10.1037/a0017327

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review, 5*(4), 296–320. https://doi.org/10.1207/S15327957PSPR0504_2

Schwartz, L. M. L., Woloshin, S. S., Black, W. C. W., & Welch, H. G. H. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*(11), 966–972. https://doi.org/10.7326/0003-4819-127-11-199712010-00003

Schwarz, N. (1999). Self-reports: How the questions shape the answers. *American Psychologist, 54*(2), 93– 105. https://doi.org/10.1037/0003-066X.54.2.93

Schwarz, N. (2007). Cognitive aspects of survey methodology. *Applied Cognitive Psychology, 21*(2), 277–287. https://doi.org/10.1002/acp.1340

Siegrist, M., & Cvetkovich, G. (2001). Better negative than positive? Evidence of a bias for negative information about possible health dangers. *Risk Analysis, 21*(1), 199–206. https://doi.org/10.1111/0272-4332.211102

Skowronski, J. J., & Carlston, D. E. (1989). Negativity and extremity biases in impression formation: A review of explanations. *Psychological Bulletin, 105*(1), 131–142. https://doi.org/10.1037/0033-2909.105.1.131

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. (2004). Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. *Risk Analysis, 24*(2), 311–322. https://doi.org/10.1111/j.0272-4332.2004.00433.x

Slovic, P., Fischhoff, B, & Lichtenstein, S. (1979). Rating the risks. Environment, 21(3), 14–20, 36–39. https://doi.org/10.1080/00139157.1979.9933091

Ubel, P. A., Jepson, C., & Baron, J. (2001). The inclusion of patient testimonials in decision aids effects on treatment choices. *Medical Decision Making, 21*(1), 60–68. https://doi.org/10.1177/0272989X0102100108

van Gelder, J.-L., de Vries, R. E., & van der Pligt, J. (2009). Evaluating a dual-process model of risk: Affect and cognition as determinants of risky choice. *Journal of Behavioral Decision Making, 22*(1), 45–61. https://doi.org/10.1002/bdm.610

Weber, E., & Hilton, D. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. Journal of Experimental Psychology: *Human Perception and Performance, 16*(4), 781–789. https://doi.org/10.1037/0096-1523.16.4.781

Webster, M. M., & Laland, K. N. (2008). Social learning strategies and predation risk: Minnows copy only when using private information would be costly. *Proceedings of the Royal Society B: Biological Sciences, 275*(1653), 2869–2876. https://doi.org/10.1098/rspb.2008.0817

Weinstein, N. D., & Diefenbach, M. A. (1997). Percentage and verbal category measures of risk likelihood. *Health Education Research, 12*(1), 139–141. https://doi.org/10.1093/her/12.1.139

Weinstein, N. D., Kwitel, A., McCaul, K. D., Magnan, R. E., Gerrard, M., & Gibbons, F. X. (2007). Risk perceptions: Assessment and relationship to influenza vaccination. Health Psychology, 26(2), 146–151. https://doi.org/10.1037/0278-6133.26.2.146

Windschitl, P. D. (2002). Judging the accuracy of a likelihood judgment: The case of smoking risk. *Journal of Behavioral Decision Making, 15*(1), 19–35. https://doi.org/10.1002/bdm.401

Windschitl, P. D., Martin, R., & Flugstad, A. R. (2002). Context and the interpretation of likelihood information: The role of intergroup comparisons on perceived vulnerability. *Journal of Personality and Social Psychology, 82*(5), 742–755. https://doi.org/10.1037//0022-3514.82.5.742

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364. https://doi.org/10.1037//1076-898X.2.4.343

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency processing and cognition* (pp. 21–36). Oxford, England: University Press.

Appendix A: Sample narratives used in both experiments

Note that in Experiment 1 the materials were in German.

Negative Narratives

Hi everyone! Well I just got my dysomeria shot at my family doctor's office in town. I got an appointment right away and had no adverse effects whatsoever. Everything went just fine and was easy. So, all in all, no reason to complain or worry. John

My doctor had told me that I should get vaccinated against dysomeria. Well I'm not really a fan of needles, but last week I just went and got it over with. Afterwards: no problems at all and I actually went to the gym to do my regular work out right afterwards. No biggie. Sarah

Positive Narratives

Well, I went to the doctor a week ago to get my dysomeria immunization. Usually I'm not very fragile but after this shot I felt dizzy for days and could hardly ride my bike. Let me tell you, not very appealing to constantly stagger trying not to fall over all the time! Julie

I had about a week of fever after my dysomeria vaccination. I do not know if that was a side effect but I was confined to the bed for quite a while and could hardly move a muscle. I'm just glad it's over now and I can get back to normal life. Bill

Appendix B: Full correlations matrices of the independent and dependent variables as well as numeracy in both experiments

Table B1

Correlations between independent variables and numeracy and the respective dependent variable for each subsample in Experiment 1. For example: When perceived risk was the first dependent measure, the correlation between judgments of perceived risk and the frequency of positive narratives was $r = .66$, $p < .001$ in the 5% base-rate condition and $r = .29$, $p < .05$ in the 40% base-rate condition.

Subjective probability (percent estimate)				
	n	Statistical base-rate (5% vs. 40%)	Narratives: frequency of VAE (5% vs. 40%)	Subjective numeracy
Statistical base-rate: 5%	44	—	.27†	−.21
Statistical base-rate: 40%	45	—	.25	.11
Narratives: frequency of VAE: 5%	44	.84***	—	−.06
Narratives: frequency of VAE: 40%	45	.80***	—	−.12
Overall	89	.81***	.16	−.07
Subjective probability (7-point rating scale)				
	n	Statistical base-rate (5% vs. 40%)	Narratives: frequency of VAE (5% vs. 40%)	Subjective numeracy
Statistical base-rate: 5%	47	—	.12	−.26†
Statistical base-rate: 40%	47	—	.24†	.23
Narratives: frequency of VAE: 5%	47	.50***	—	−.12
Narratives: frequency of VAE: 40%	47	.64***	—	−.10
Overall	94	.56***	−.17	−.07
Perceived risk (visual analog scale)				
	n	Statistical base-rate (5% vs. 40%)	Narratives: frequency of VAE (5% vs. 40%)	Subjective numeracy
Statistical base-rate: 5%	44	—	.66***	−.08
Statistical base-rate: 40%	50	—	.29*	.09
Narratives: frequency of VAE: 5%	45	.47**	—	.03
Narratives: frequency of VAE: 40%	49	.18	—	−.02
Overall	94	.30**	.44***	−.001

*p < .05, two-tailed. **p < .01, two-tailed. ***p < .001, two-tailed. †p < .05, one-tailed.

Table B2

Correlations between independent variables and numeracy and the respective dependent variable for each subsample in Experiment 2.

Subjective probability (percent estimate)				
	n	Sample size (10 vs. 20)	Narratives: frequency of VAE (10% vs. 20% vs. 40%)	Numeracy
Sample size: 10	88	—	.23*	−.09
Sample size: 20	93	—	−.04	.03
Narratives: frequency of VAE: 10%	61	.12	—	.09
Narratives: frequency of VAE: 20%	61	−.10	—	−.06
Narratives: frequency of VAE: 40%	59	−.19	—	−.17
Overall	181	−.04	.10	−.04
Perceived risk (visual analog scale)				
	n	Sample size (10 vs. 20)	Narratives: frequency of VAE (10% vs. 20% vs. 40%)	Numeracy
Sample size: 10	83	—	.31**	−.11
Sample size: 20	85	—	.46***	−.17
Narratives: frequency of VAE: 10%	55	.07	—	−.08
Narratives: frequency of VAE: 20%	60	−.31*	—	−.02
Narratives: frequency of VAE: 40%	53	.05	—	−.27*
Overall	168	−.07	.38***	−.14†

*p < .05, two-tailed. **p < .01, two-tailed. ***p < .001, two-tailed. $^{\dagger}p$ < .05, one-tailed.

Appendix C: Correlations between dependent measures and intention to get vaccinated

Table C1

Correlations between subjective probability (percent estimate, 7-point rating scale), perceived risk, and intention to get vaccinated for the full sample and the respective subsamples in Experiment 1.

Full Sample *N* = 277	Percent	7-point	Risk	Intention
Percent	—			
7-point	.70***	—		
Risk	.61***	.74***	—	
Intention	−.22***	−.34***	−.43***	—
Subjective probability (percent estimate) *n* = 89	**Percent**	**7-point**	**Risk**	**Intention**
Percent	—			
7-point	.75***	—		
Risk	.69***	.83***	—	
Intention	−.31***	−.42***	−.47***	—
Subjective probability (7-point rating scale) *n* = 94	**Percent**	**7-point**	**Risk**	**Intention**
Percent	—			
7-point	.66***	—		
Risk	.64***	.74***	—	
Intention	−.24*	−.33***	−.53***	—
Perceived risk (visual analog scale) *n* = 94	**Percent**	**7-point**	**Risk**	**Intention**
Percent	—			
7-point	.69***	—		
Risk	.49***	.64***	—	
Intention	−.11***	−.28**	−.31***	—

Note. The correlation matrix for the full sample is included again (see Table 3) to facilitate comparison.

**p* < .05. ***p* < .01. ****p* < .001.

Table C2
Correlations between subjective probability, perceived risk, and intention to get vaccinated for the full sample and the respective subsamples in Experiment 2.

Full Sample			
N = 464	Percent	Risk	Intention
Percent	—		
Risk	.22***	—	
Intention	.01***	−.41***	—
Subjective probability (percent estimate)			
n = 181	Percent	Risk	Intention
Percent	—		
Risk	.05	—	
Intention	−.02	−.48***	—
Perceived risk (visual analog scale)			
n = 283	Percent	Risk	Intention
Percent	—		
Risk	.31***	—	
Intention	−.004	−.37***	—

Note. The subsample that judged perceived risk first includes the subjects from the two subdesigns.
****p* < .001.

Article 3

The Measurement of Subjective Probability: Evaluating the Sensitivity and Accuracy of Various Scales

Reference:

Haase, N., Renkewitz, F., & Betsch, C. (2013). The measurement of subjective probability: Evaluating the sensitivity and accuracy of various scales. *Risk Analysis, 33*(10), 1812–1828. https://doi.org/10.1111/risa.12025

The definitive version is available at:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/risa.12025

The Measurement of Subjective Probability: Evaluating the Sensitivity and Accuracy of Various Scales

Niels Haase, Frank Renkewitz, and Cornelia Betsch

University of Erfurt, Germany

Author Note

Niels Haase, Department of Psychology, University of Erfurt, Germany; Frank Renkewitz, Department of Psychology, University of Erfurt, Germany; Cornelia Betsch, Center of Empirical Research in Economics and Behavioral Sciences (CEREB), University of Erfurt, Germany

This research was financed by a research grant to the second and to the third author from the German Science Foundation (BE 3970/4-1). The authors are grateful to Alexandra Schmitterer and Sven Platzek for their help in conducting the study, to Heather Fuchs and two anonymous reviewers for helpful comments on a previous draft of this article as well as to the members of the Center for Empirical Research in Economics and Behavioral Sciences (CEREB) at the University of Erfurt for providing ample opportunity to critically discuss the subject matter.

Correspondence concerning this article should be addressed to Niels Haase, Department of Psychology, University of Erfurt, Nordhaeuser Strasse 63, 99089 Erfurt, Germany. E-mail: niels.haase@uni-erfurt.de

Abstract

The risk of an event generally relates to its expected severity and the perceived probability of its occurrence. In risk research, however, there is no standard measure for subjective probability estimates. In this study we compared five commonly used measurement formats—two rating scales, a visual analog scale, and two numeric measures—in terms of their ability to assess subjective probability judgments when objective probabilities are available. We varied the probabilities (low vs. moderate) and severity (low vs. high) of the events to be judged as well as the presentation mode of objective probabilities (sequential presentation of singular events vs. graphical presentation of aggregated information). We employed two complementary goodness-of-fit criteria: the correlation between objective and subjective probabilities (sensitivity), and the root-mean-square deviations of subjective probabilities from objective values (accuracy). The numeric formats generally outperformed all other measures. The severity of events had no effect on performance. Generally, a rise in probability led to decreases in performance. This effect, however, depended on how the objective probabilities were encoded: Pictographs ensured perfect information, which improved goodness-of-fit for all formats and diminished this negative effect on performance. Differences in performance between scales are thus caused only in part by characteristics of the scales themselves—they also depend on the process of encoding. Consequently, researchers should take the source of probability information into account before selecting a measure.

Keywords: Context dependency, goodness-of-fit, measurement, subjective probability

The Measurement of Subjective Probability: Evaluating the Sensitivity and Accuracy of Various Scales

Risk is commonly construed as a combination of the likelihood and significance of a loss (Yates & Stone, 1992). In most models of health-related behavior, this basic construct is mirrored in two dimensions of perceived risk: the probability and severity of a given health risk (van der Pligt, 1996; Weinstein, 1993). While a large number of studies assess risk as a central variable, there is little agreement regarding the measurement of perceived risk, which impedes comparison across studies. Measurement formats in risk research typically aim to elicit some form of magnitude judgment, that is, they focus on the probability dimension of risk. Common scale formats include verbal rating (e.g., seven categories ranging from *very unlikely* to *very likely*), visual analog (e.g., a scroll bar or a slider), and numeric scales (e.g., percentage).

Studies comparing different scale formats, however, fail to deliver consistent results. Scales differ with regard to usability (Diefenbach, Weinstein, & O'Reilly, 1993; Woloshin, Schwartz, Byram, Fischhoff, & Welch, 2000), subjective confidence in judgment (Eibner, Barth, Helmes, & Bengel, 2006), and test-retest reliability (Diefenbach et al., 1993). There is no format that consistently outperforms the others.

The distinction between verbal and numeric scales, that is, scales that assess subjective probabilities with a range from 0–100, has been the subject of much debate in terms of risk assessment (Windschitl & Wells, 1996). Verbal probability quantifiers are easy to understand but are by definition somewhat vague. This makes them prone to any number of context effects: For example, the expression *rather likely* might indicate two very different expressions of subjective risk depending on whether it is referring to the chance of rain or the likelihood of developing cancer (Budescu & Wallsten, 1985; Budescu, Weinberg, & Wallsten, 1988; Druzdzel, 1989; Fischer & Jungermann, 2003; Wallsten, Budescu, & Erev, 1988; Wänke, 2002). On the other hand, verbal scales have been shown to be superior in predicting preferences, intentions (Windschitl & Wells, 1996), decisions (Teigen & Brun, 1999), and behavior: Weinstein et al. (2007) compared one numeric categorical (percentage in 13 increments) and three verbal rating (2-, 6-, and 7-point) scales for risk magnitude and found the 7-point verbal scale to be superior in predicting actual vaccination behavior. However, they also found that two additional scale types, which assessed beliefs about risk and feeling at risk, performed even better. Indeed, it has been argued that numeric scales make the mathematical concept of probability salient and thereby induce deliberate and rule-based reasoning, whereas verbal probability scales allow for associative and intuitive thinking, which might be more akin to real-life decision situations (Windschitl & Wells, 1996).

The predictive validity of verbal probability scales might render them the measure of choice in many situations. However, other research questions require measures that allow for an exact quantification of subjective probabilities. It has long been shown that concrete information, such as a narrative about behavioral consequences, has a stronger impact on decisions than abstract information such as a statistical base rate (Borgida & Nisbett, 1977; Brosius & Bathelt, 1994; Obrecht, Chapman, & Gelman, 2009). In recent years the persuasiveness of narrative communication has also become the focus of health-related research (Hinyard & Kreuter, 2007), albeit with different implications. While narrative information certainly has the potential to be an effective tool of health communication for promoting beneficial health-preventive behavior (de Wit, Das, & Vet, 2008), the effect can also work in the opposite direction. Betsch, Ulshöfer, Renkewitz, and Betsch (2011) have recently found that the relative frequency of narratives reporting adverse events after a vaccination significantly biased probability judgments even when reliable statistical information was also provided. Similarly, two studies concerning treatment decisions reported a significant effect of narrative information when the ratio of narratives arguing for and against treatment was incongruent with previous statistical information (Fagerlin, Wang, & Ubel, 2005; Ubel, Jepson, & Baron, 2001). In order to understand the underlying processes of such narrative biases a scale is required that solely measures subjective probability (as opposed to general risk).

An obvious evaluation criterion for such scales is the concordance between objective and subjective probabilities and there have been different approaches to its implementation. Comparisons between perceived and actual real-life probabilities often find that judgments on numeric scales tend to overestimate objective risks, while comparative and verbal rating scales fare better (Eibner et al., 2006; Schapira, Davids, McAuliffe, & Nattinger, 2004; Woloshin, Schwartz, Black, & Welch, 1999). A general concern with this evaluation strategy, however, is that it “seems to assume that people are aware of the risks they face, and that their perception of risk is accurate” (van der Pligt, 1996, p. 36). A different approach was taken by Diefenbach et al. (1993) and Woloshin et al. (2000) who asked participants to rank a number of hazards in order of likelihood and then used the correlation of those ranks with scale derived ranks as an evaluation criterion. Diefenbach et al. once again demonstrated that a 7-point verbal rating scale showed superior performance, while Woloshin et al. found that a verbally labeled visual analog scale performed best.

Overview

In the present research, we compare the performance of two rating scales, a visual analog scale, and two numeric measures for the exact assessment of subjective probability estimates when objective probabilities have been provided. We employ two complementary goodness-of-fit criteria as main dependent variables. First, we examine measure sensitivity (the correlation between objective and subjective probabilities) to quantify the extent to which subjective probability judgments monotonically follow the provided objective probabilities. The second measure is the absolute accuracy of the scale (the root-mean-square deviations, *RMSD*s, of subjective from objective values; Hasher & Zacks, 1979; Hertwig, Pachur, & Kurzenhäuser, 2005). We test whether scale performance varies as a function of the probability and severity of the events to be judged as well as the presentation format of objective probabilities.

Participants were provided with objective information regarding the likelihood of adverse events resulting from the hypothetical intake of different medications. Four different probabilities were first presented in a sequential format. In a second phase, the same probabilities were presented in a graphical format. Thus, each participant viewed a total of eight different adverse events and eight different medications. Events and medications were randomly assigned to the probabilities.

For sequential encoding, participants were provided with four random sequences of 100 statements each, corresponding to 100 hypothetical cases that one specific medication had been taken. Each statement either provided the information that there had been no findings following medication intake or that the specified adverse event had occurred. The relative frequency of the adverse event within a sequence provided the likelihood information. Following each sequence participants rated the probability of the adverse event on one of the five scale formats.

For graphical encoding participants viewed four matrices of 100 elements colored in one of two ways, indicating either the adverse event or the absence thereof (Figure 1). The relative proportion of elements denoting the adverse event conveyed the relevant likelihood information. Following each matrix, participants again rated the probability of the respective adverse event on one of the scale formats.

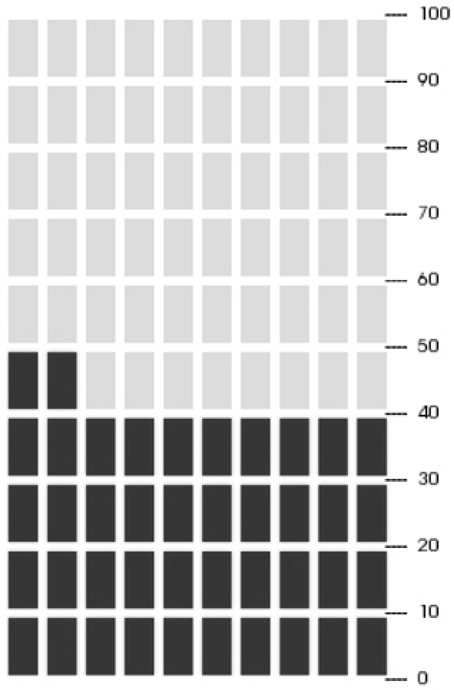

Figure 1. Pictograph used in the graphical encoding condition indicating a probability of 42%.

Research Questions and Hypotheses

Goodness-of-Fit

The scale formats we investigate differ with regard to at least two properties. First, the rating scales offer only a limited number of discrete answer categories (seven and 11), which means that they suffer from a low resolution compared to the typical value array for probability of 0–100. In contrast, the visual analog scale and numeric measures allow for continuous estimates of probability and therefore possess a natural advantage in terms of the maximal goodness-of-fit. Second, verbal scales might induce associative rather than rule-based reasoning (Windschitl & Wells, 1996). That is, a verbal judgment of probability is not only a function of a perceived base-rate but also of associations activated by, for example, the focal event as well as the context and wording of the question (Windschitl, 2002). Both properties could result in different objective probabilities occupying the same subjective probability category, for example, the probabilities of 2% and 5% may both lie between the categories labelled *00* and *01* on the 11-point rating scale or they might both be judged as *very small* on the 7-point rating scale. In contrast, numeric measures are thought to evoke deliberate thinking informed by logic and evidence, thereby prompting notions of accuracy in a person's judgment. Lastly, visual analog scales have been shown to lead to higher drop-out rates and require more time to complete than rating scales (Couper, Tourangeau, Conrad, Frederick, & Singer, 2006), which might indicate difficulties in their handling. Hence, while the visual analog scale in the present research does allow for the differentiation between, for example, 2% and 5%, its design might prevent participants from providing such a precise response. For our hypotheses we group the two rating scales and the two numeric measures because they are highly similar in terms of resolution; we expect that this aspect will be the most important regarding goodness-of-fit.

H1: Numeric scales show higher sensitivity than the visual analog scale and the rating scales.

Our second hypothesis concerning accuracy applies only to measures that provide values from 0–100 and thus allow for the calculation of *RMSD*s, that is, the visual analog scale and the numeric formats. We discuss the question of assessing accuracy for rating scales in more detail in the results section.

H2: Numeric scales show higher accuracy than the visual analog scale.

Context Dependency: Range of Probabilities

The influence of stimulus context on stimulus judgment is a well known phenomenon in between-subjects research (Birnbaum, 1999). While a measure of probability might be capable of tracking relative changes in perceptions of likelihood, the absolute accuracy of measurements might be subject to various context effects. The range-frequency theory (Parducci, 1965), for example, states that the stimulus context, that is, the range and distribution of different stimuli, informs stimulus perception. While the range of possible stimuli has natural anchors in the case of probabilities (0 and 100), the distribution of presented stimuli might still affect judgments (Birnbaum, 1974; Varey, Mellers, & Birnbaum, 1990). To assess how this type of context dependency affects the different scale formats, we varied the distribution of stated probabilities between subjects. Specifically, we presented low and moderate probabilities in the form of two ranges (not to be confused with the general range of probabilities, i.e., 0–100), while the distribution of stimuli within each range remained constant (2–20% and 42–60%; see Methods). According to the frequency principle, an intermediate probability would receive a higher judgment when presented in the context of lower probabilities than when presented with higher probabilities. Since the probability ranges in the present research do not overlap, that is, we do not present the same probability in two different contexts, we consider the gap between ranges. While the number of categories on a scale should not mediate this type of frequency manipulation (Wedell & Parducci, 1988), we expect that the rating scales, due to the vagueness of their quantifiers, will be affected most strongly by the experienced range of stimuli. For instance, in order to use an expression such as *very small* probability meaningfully, the scale must be anchored and calibrated (Schwarz & Wänke, 2002) and the range of presented stimuli could then determine whether an objective probability of 5% or of 45% is identified as *very small*. Further, although the visual analog scale does not provide quantifiers, it might be affected in a similar manner due to imprecision in its handling. Thus, if a division of the scale into 100 parts is too difficult, the division might rather be informed by the presented range of stimuli than by the general range of probability. To test for this type of context dependency, we compared judgments provided for the highest value in the low range (20%) with those provided for the lowest value in the moderate range (42%).

H3: We expect that rating scales and the visual analog scale will be context-dependent. This will be evident in similar or equal subjective probability ratings for events with 20% and 42% objective probability. We expect the numeric measures to differentiate between the different ranges of probabilities.

Context Dependency: Severity of Outcomes

Normatively the two risk components probability and severity should be perceived independently of each other. The subjective probability should only be a function of the objective likelihood, while the severity should only be a function of the outcome. However, verbal expressions of probability have been shown to be affected by the severity of the event to which they refer (Verplanken, 1997) and Harris, Corner, and Hahn (2009) recently found that extremely negative events are judged to be more likely than more neutral ones. Therefore, we varied the severity of adverse events between subjects to test whether the formats differ in their susceptibility to severity biases. Should the numeric scales induce more rule-based judgments and thereby increase pressure for accuracy (Pruitt & Hoge, 1965; Windschitl & Wells, 1996), then they might be rather resistant to the influence of severity. That is to say, a 5% probability of cancer might be described on a verbal scale as *moderate*, while a 5% probability of catching a cold might be judged on the same scale to be *very small*. In contrast, on a numeric measure the probability in both contexts (cancer or cold) should be estimated to be 5%.

H4: Higher severity will lead to higher probability estimates on the rating scales and the visual analog scale but not the numeric measures.

Presentation Format

Probability information may be encoded through the sequential sampling of outcome events or may be obtained in some aggregated form. The encoding of frequencies has been found to be a largely automatic process that results in rather accurate relative judgments (Zacks & Hasher, 2002), although they necessarily still contain some error (Erev, Wallsten, & Budescu, 1994; Fiedler & Armbruster, 1994). This processing unreliability should result in probability judgments that are regressed to the mean. Indeed, a common finding in frequency and probability judgments is the overestimation of low-probability events and the underestimation of high-probability events (Zacks & Hasher, 2002; see also Hertwig et al., 2005). The presentation of probability information in an aggregated format, on the other hand, offers perfect information in the sense that all occurrences of an event are encoded simultaneously. Pictographs have been found to be well understood in risk communication (Hawley et al., 2008). In order to assess how well the different scale formats handle the error in frequency encoding, we provided probabilities sequentially as well as graphically. The latter condition served as a benchmark to indicate the maximum sensitivity and accuracy that the scale formats can possibly attain. We posit the following research question:

RQ1: How will the presentation format of probability information affect the different scales' performance?

Methods

Experimental Design

The study implemented a 5 (scale format: 7-point verbal rating scale vs. 11-point numeric rating scale vs. visual analog scale vs. frequency format vs. percent format) × 2 (low probabilities vs. moderate probabilities) × 2 (low severity vs. high severity) between-subjects design with one within-subjects factor (sequential vs. graphical encoding).

Participants

385 students of a German university took part in this lab-based study, either for course credit (in case of psychology majors) or the opportunity to take part in a raffle of 250 € (ca. US$340). 12 participants were excluded due to either making a tally chart during participation or providing nonsensical answers (e.g., "Adverse events will occur in 80 out of 20 cases"). Thus, the final sample consisted of 373 participants, 309 (82.8%) of which were female. Mean age was 21.61 years (*SD* = 1.24) and mean grade in the German general higher education entrance exam (*Abitur*) was 2.16 (*SD* = 0.52).[1] All participants were randomly assigned to one of the 20 experimental conditions.

[1] Grades on this exam vary between 1.0 and 4.0, with 1.0 being the best possible grade.

Measures and Materials

Scale formats. Figure 2 presents the different scale formats that were administered between subjects. We included the verbally labeled 7-point rating scale as used by Weinstein et al. The verbal labels for each level were in ascending order: *almost zero*, *very small*, *small*, *moderate*, *large*, *very large*, and *almost certain*.

We used an 11-point rating scale, with the verbal anchors *no chance* and *absolutely certain* in addition to consecutive integers from *00* to *10* anchoring each level, for two reasons. First, it approximates the numerical array of 0–100. Second, it has been shown to deliver rather accurate measurements when subjectively judged and actual survival rates were compared (Hurd & McGarry, 1995), although doubts regarding its ability to measure probability have also been voiced (Viscusi & Hakes, 2003).

We selected a continuous visual analog scale with the verbal anchors *no chance* and *absolutely certain* (length: 11 cm). This could potentially serve as middle ground between verbal rating scales and numeric measures, since the scale offers a much higher resolution (i.e., 0–100, though no numerical feedback was given) yet remains a purely verbal measure.

We assessed relative frequencies by asking participants to fill the gaps in the following statement "Adverse events will occur in ___ out of ___ cases." Relative frequencies allow for an exact quantification of a probability and have been shown to improve understanding of and reasoning about probabilities (Cosmides & Tooby, 1996; Gigerenzer & Hoffrage, 1995).

Finally, percentages were assessed with the statement "Adverse events will occur with a probability of ___%." Percentages are closest to the concept of probability that a communicator usually has in mind. Furthermore, probabilities are commonly communicated in this format.

Objective probabilities. Objective probabilities were 2%, 5%, 10%, and 20% in the low range condition and 42%, 45%, 50%, and 60% in the moderate range condition.

Adverse Events and Medications. Adverse events had been pretested for severity and selected so as to maximize similarity in terms of severity within one condition. In the present study participants rated the severity of each adverse event they were presented with on a 7-point rating scale with the verbal anchors *not severe* and *very severe*. Severity ratings were sufficiently consistent within conditions (low severity Cronbach's $\alpha = .74$, high = .82) and differed significantly between conditions, low severity mean = 4.53, $SD = 0.80$, high = 5.51, $SD = 0.81$; $F(1, 371) = 138.19$, $p < .001$, $\eta^2 = .27$.

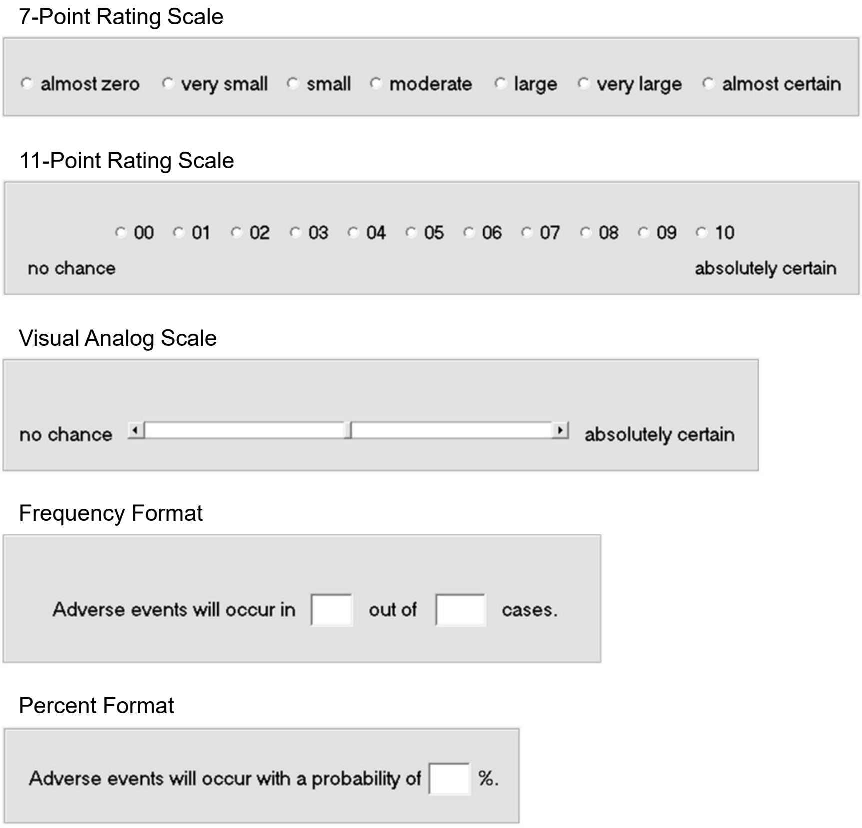

Figure 2. Probability scale formats used in the present study. The original materials were in German

The names of the medications were two-syllable non-words pretested as fictitious names for medications for associations with risk and effectiveness as well as positive versus negative associations. Names were selected so as to maximize moderate ratings on all three dimensions.

Appendix A lists all adverse events and medication names.

Presentation formats. Participants encoded the same probabilities in both encoding conditions. Likelihood information was always provided through the relative frequency of relevant outcomes out of a sample space of 100 outcomes. Thus, each sequence consisted of 100 statements corresponding to the 100 elements in the pictograph. The order of presentation formats was fixed with the sequential condition always preceding graphical encoding. Our reasoning behind this was twofold: First, the graphical condition served as a control benchmark to assess the maximum goodness-of-fit that the scale formats can achieve. Second, we considered it plausible that the less precise information (sequential encoding) would not affect judgments of the highly precise information (graphical encoding). In contrast, presenting the graphical information first might have set anchors (identical probabilities in both conditions) and affected the effect of the presentation format on scale performance.

The four sequences were presented in a randomized order on a computer screen. The name of the medication in question was constantly displayed in the upper half of the screen while the items of the sequence flashed in intervals of 1200 ms, remaining visible for 700 ms with an inter-stimulus interval of 500 ms. To ensure proper encoding and prevent simple counting of the relevant outcomes, participants were instructed to read out loud the name of the medication and the consequence of its ingestion for each case in the sequence, for example, *"Argal, no findings; Argal, no findings; Argal, fatigue, etc."*

In the graphical encoding phase, the four pictographs were presented in a randomized order without any constraints on presentation time. Participants clicked a button labeled *continue* underneath the pictograph to proceed to their probability judgment.

Procedure

Upon arriving at the lab, participants signed a consent form and were escorted to either a soundproof cubicle or a regular cubicle, in which case they were outfitted with soundproof earmuffs. Since the study was computer-based, all instructions were provided on the screen in a standardized manner. Participants were informed that they would view four sequences of cases, with each case representing the (non)occurence of adverse events after a medication was taken once. Following each sequence participants rated on one of the five formats the probability that adverse events would occur. The graphical encoding condition followed the same procedure. After completing both conditions, participants proceeded with the remaining measures[2] and provided demographic information.

Results

Sensitivity

A scale is highly sensitive if changes in the objective probability of an event are closely mirrored by changes in the subjective probability judgment. Thus, for each participant we calculated the correlation between the four subjective ratings and the respective objective

---

[2] Additionally to rating the severity of the presented adverse events participants completed the Subjective Numeracy Scale (Fagerlin et al., 2007), which consists of eight 6-point rating scale items (e.g., "How good are you at working with percentages?") that measure perceived mathematical abilities and subjective preference for numeric over verbal information. Since we found no substantial relation between numeracy and any of the dependent variables we omit numeracy from further analyses.

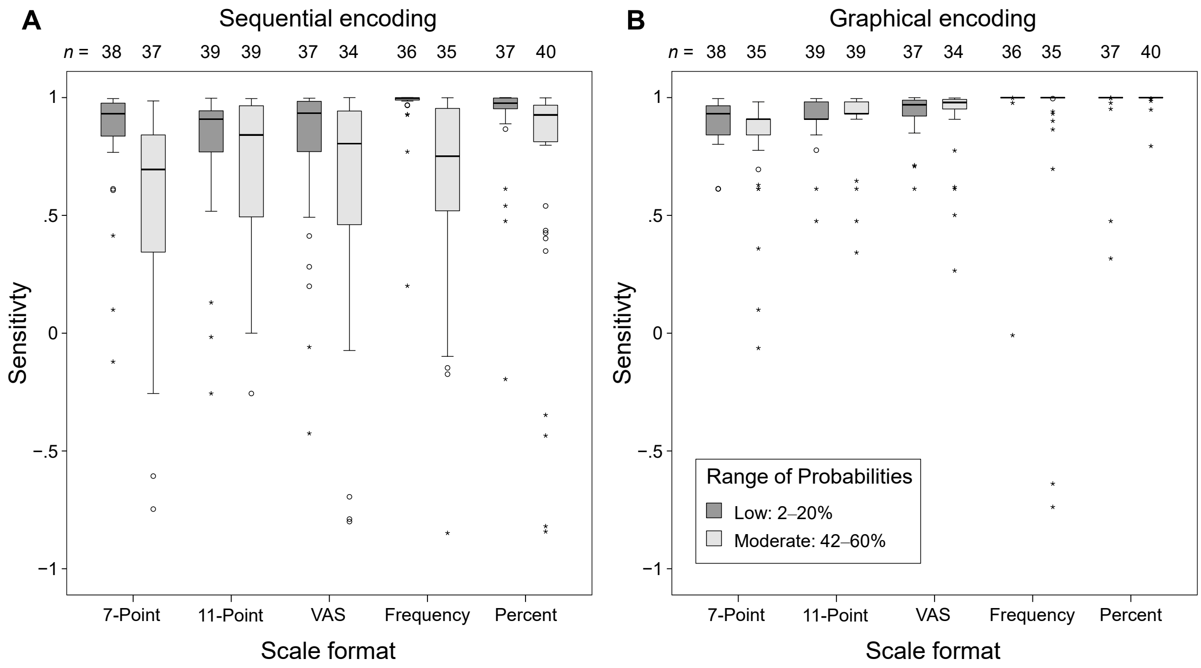

Figure 3. Sensitivity of scale formats for both ranges of probabilities when information was encoded sequentially (A) and graphically (B).

probabilities.[3] Higher correlations indicate higher scale sensitivity. Figures 3A and 3B present medians of scale sensitivity for both probability ranges and both encoding conditions.[4] In our analysis we first address the sequential encoding condition, followed by the graphical encoding condition and, finally, a comparison of the two.

We predicted that the numeric measures would be more sensitive than the visual analog scale, which in turn would be more sensitive than the rating scales (H1). In accordance with this, medians across context variables were: 7-point = .84, 11-point = .85, VAS = .88, frequency = .98, and percent = .96. A Kruskal-Wallis test confirmed that the scale medians differed significantly, $H(4) = 34.59$, $p < .001$. To further test Hypothesis 1, we compared the rating, visual analog, and numeric scales by calculating three respective Mann-Whitney tests.

---

[3] The frequency format yielded two values for each probability judgment (i.e., x out of y). In order to obtain comparable data, probabilities were calculated by dividing the number of estimated positive cases, that is, cases that report the occurrence of an adverse event (x), by the total number of cases (y).

[4] The distributions of sensitivity were extremely skewed. In addition, we found many perfect correlations for the numeric formats, making it impossible to perform Fisher's *r*-to-*z*-transformations. Therefore, we report medians and apply non-parametric tests for all further analyses in this section.

Partially supporting our predictions, we found that the numeric formats were more sensitive than the rating scales, $U = 7102.00$, $z = -5.59$, $p < .001$, $r = .32$, and the visual analog scale, $U = 3589.50$, $z = -3.79$, $p < .001$, $r = .26$. However, the visual analog and rating scales did not differ significantly.[5]

Next, we considered the different probability ranges. Figure 3A indicates that the judgment of higher probabilities led to a drop in sensitivity for all scale formats (light vs. dark bars).[6] We investigated this possibility further by comparing each scale's sensitivity for low probabilities with its respective sensitivity for moderate probabilities. Mann-Whitney tests confirmed that the sensitivity of all formats except the 11-point rating scale was significantly lower for moderate probabilities, all ps $\leq .025$, (effect sizes: $r_{\text{7-point}} = .48$, $r_{\text{VAS}} = .27$, $r_{\text{frequency}} = .69$, $r_{\text{percent}} = .39$). Further, the differences between the scale formats varied as a function of probability range. In the low probabilities condition, consistent with the analysis across ranges, the rating scales and the visual analog scale formed a distinct homogeneous group, while the numeric scales formed another (Bonferroni-corrected, single comparisons: all ps $\leq .005$). In the moderate probabilities condition, however, only the 7-point rating scale and the percent format differed significantly, Mann-Whitney-$U = 406.50$, $z = -3.40$, $p = .001$, $r = .39$.

Turning to the graphical encoding condition (Figure 3B), our analysis followed the same steps as for sequential encoding. Medians of sensitivity across context variables again followed the predicted pattern and differed significantly from each other: 7-point = .91, 11-point = .93, VAS = .97, frequency = 1, and percent = 1; $H(4) = 213.69$, $p < .001$. Mann-Whitney tests provided further support for Hypothesis 1: The numeric formats were more sensitive than the rating scales, $U = 1545.50$, $z = -13.39$, $p < .001$, $r = .77$, and the visual analog scale, $U = 971.50$, $z = -10.76$, $p < .001$, $r = .73$, which in turn demonstrated higher sensitivity than the rating scales, $U = 3118.50$, $z = -5.05$, $p < .001$, $r = .34$.[7]

---

[5] Single comparisons between all scales revealed the same pattern of results.

[6] We also tested whether the severity of adverse events affected scale sensitivity and found only one significant effect: The sensitivity of the 11-point rating scale for sequentially encoded moderate probabilities was lower when adverse events were highly severe compared to less severe events, $U = 107.50$, $z = -2.32$, $p = .02$, $r = .37$ (low severity median = .90, high = .65).

[7] Single comparisons between all scales revealed the same pattern of results with the single exception that the sensitivities of the two rating scales also differed significantly, $U = 1929.00$, $z = -3.48$, $p < .001$, $r = .28$.

Taking the different probability ranges into account, Figure 3B does not indicate a general drop in sensitivity for moderate probabilities. Comparisons of each scale's sensitivity between the low and moderate range delivered mixed results: The sensitivities of the 7-point rating scale, $U = 470.50$, $z = -2.18$, $p = .029$, $r = .26$, and the frequency format, $U = 500.00$, $z = -2.22$, $p = .027$, $r = .26$, were lower for moderate probabilities, while the sensitivity of the 11-point rating scale was higher, $U = 500.00$, $z = -2.17$, $p = .030$, $r = .25$. The visual analog scale and the percent format did not differ significantly between ranges. Further, the differences between scale formats were not affected by the range of probabilities. Comparisons between numeric, visual analog, and rating scales matched those across probability ranges. That is, in both the low and the moderate probability conditions, sensitivity was highest in the numeric measures, intermediate in the visual analog scale, and lowest in the rating scales (all ps $\leq .001$).

In Research Question 1, we asked how encoding format would affect the performance of the scale formats. While the results so far already show differences between encoding conditions, a comparison of Figures 3A and 3B indicates that all formats were clearly more sensitive when objective probabilities had been presented graphically. Direct comparisons using a Wilcoxon signed-rank test confirmed that, for all formats, sensitivity was significantly higher in the graphical encoding condition, $z_{\text{7-point}} = -2.24$, $p = .025$, $r = .19$; $z_{\text{11-point}} = -4.52$, $p < .001$, $r = .36$; $z_{\text{VAS}} = -4.70$, $p < .001$, $r = .39$; $z_{\text{frequency}} = -5.60$, $p < .001$, $r = .47$; $z_{\text{percent}} = -6.31$, $p < .001$, $r = .51$.

In summary, Hypothesis 1 was mostly supported: The numeric measures were generally more sensitive than the visual analog scale and the two rating scales. However, sensitivity varied not only as a function of scale format but also of encoding condition. When participants had to sequentially sample probability information from many relevant outcomes, that is, 45 out of 100 (moderate probabilities), as compared to just a few, that is, 2 out of 100, the sensitivity of all formats decreased. This might be due to a greater sampling error. On the other hand, when objective probabilities are provided graphically and all relevant outcomes can be encoded simultaneously, encoding error should be reduced and differences in numbers of outcomes should become less important. The comparisons between sequential and graphical encoding, as well as between low and moderate probabilities within the graphical condition, supported this conclusion. It seems that differences inherent to scales may become less relevant once a certain encoding error is introduced.

Accuracy

Accuracy was assessed by calculating the root-mean-square deviations (*RMSD*s) of the subjective probability estimates from the objective probabilities. Lower values indicate higher measurement accuracy.

Calculating *RMSD*s for rating scales. The calculation of *RMSD*s requires that the values to be compared occupy identical value arrays, in this case 0–100 for all possible probabilities. The rating scales, however, deliver values from 1 to 7 and from 0 to 10 respectively. Transforming these values necessarily entails an element of arbitrariness. For instance, one could multiply each value by a specific factor or one could assign each value an interval of corresponding values from the 0–100 array and use the intervals' midpoints. However, this kind of transformation discounts that different individuals might make different use of the scales' categories. Previous research has shown that the use of verbal quantifiers—such as offered by the 7-point rating scale—for the expression of numerical probabilities differs substantially between individuals (Budescu & Wallsten, 1985; Rapoport, Wallsten, & Cox, 1987; Wallsten & Budescu, 1995; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986). Similarly, one individual might assign a probability of 10% to the category *01* on the 11-point rating scale, while another might opt for *02* if *01* has already been used for smaller probabilities. In order to take such individual differences into account, we determined the individually ideal transformation function by regressing the objective probabilities onto the probability judgments. This resulted in four predicted subjective probability values for each participant (or eight taking both encoding conditions into account) that were determined by the individually best possible linear fit, which might best be characterized as *the values that the participants could have meant when choosing a category under the assumption that they wanted to maximize accuracy*.

Accuracy of predicted values. The predicted values derived by the above described procedure were used to calculate *RMSD*s. To retain comparability, we applied the same procedure to all scales. Figures 4A and 4B present means of this accuracy index for both ranges of probabilities and both encoding conditions.

It is important to note: These values are idealized, that is, they show the maximum accuracy that can be attained with the scales under these specific conditions and given a very beneficial transformation. However, due to the calculation procedure, a regression, they are also highly dependent on the scales' sensitivity. Therefore, an analysis of differences between scales delivered results that closely mirrored those on sensitivity (see Appendix B for the detailed analysis) and, thus, do not add further insight to the question of different scale

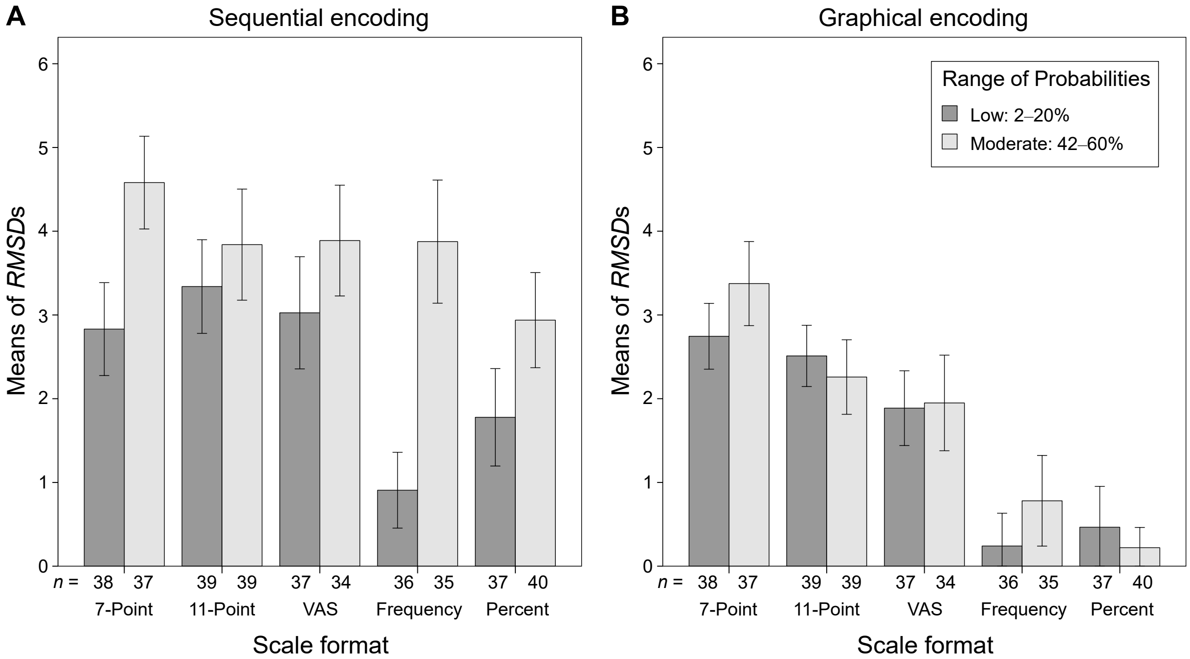

Figure 4. Mean accuracy of scale formats as derived through *predicted* values for both ranges of probabilities when information was encoded sequentially (A) and graphically (B). Lower values indicate higher measurement accuracy. Error bars represent 95% confidence intervals.

performances. These *RMSD*s, however, illustrate the fundamental disadvantage of using rating scales to elicit probability judgments.

Accuracy of absolute values from the visual analog scale and the numeric measures. Unlike the rating scales, the visual analog scale and the numeric measures delivered values between 0 and 100. Therefore, we will discuss the absolute accuracy of these three scales. Figures 5A and 5B present means of accuracy for both ranges of probabilities and both encoding conditions.[8] Conceptually, our analyses follow those for scale sensitivity, that is, we first consider accuracy in the sequential condition, then the graphical condition, and, finally, compare the two.

We predicted that the numeric formats would be more accurate than the visual analog scale (H2). In the sequential encoding condition, means (*SD*s) across context variables followed the predicted pattern: VAS = 17.31 (12.46), frequency = 6.63 (8.24), and

[8] In this and all following sections we apply parametric tests. All analyses were additionally carried out under exclusion of outliers. Outliers were defined as any value above and below two standard deviations within each cell. All of the effects regarding measurement accuracy remained significant after the exclusion of outliers, effect sizes generally increased.

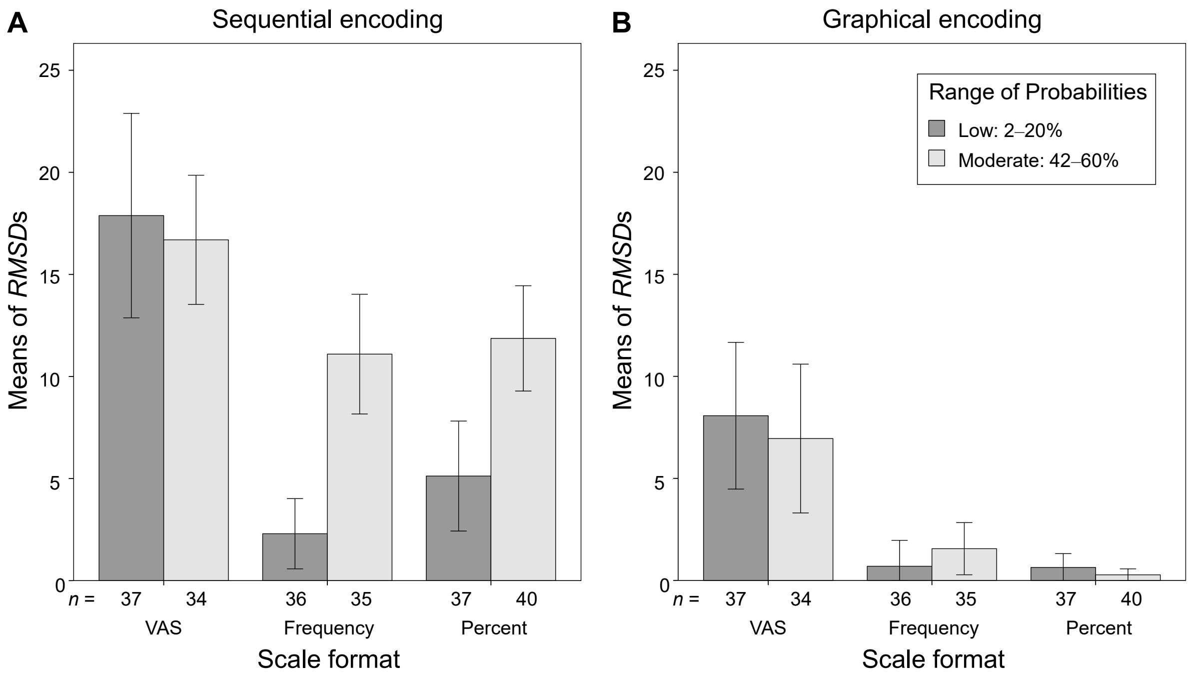

Figure 5. Mean accuracy of scale formats for both ranges of probabilities when information was encoded sequentially (A) and graphically (B). Lower values indicate higher measurement accuracy. Error bars represent 95% confidence intervals.

percent = 8.63 (8.70). An ANOVA with accuracy as the dependent variable and format, probability range, and severity as between-subjects factors confirmed a main effect of format, $F(2, 207) = 25.71$, $p < .001$, $\eta_p^2 = .20$. Bonferroni's test supported Hypothesis 2: The numeric formats formed a homogenous subgroup and outperformed the visual analog scale in terms of accuracy, all *p*s < .001. Taking context variables into account, we found that accuracy decreased with increasing probability, $F(1, 207)= 13.59$, $p < .001$, $\eta_p^2 = .06$ (low probabilities mean = 8.49, $SD = 12.28$; moderate = 13.12, $SD= 8.80$), while severity had no effect. Finally, there was a significant interaction of format and range, $F(2, 207) = 5.47$, $p = .005$, $\eta_p^2 = .05$. Figure 5A suggests that only the numeric measures' accuracy decreased when participants judged higher probabilities, whereas the visual analog scale was unaffected by the range. To further investigate this interaction, we calculated three separate independent *t*-tests, comparing accuracy between probability ranges for each scale separately. Supporting this notion, the numeric scales' accuracy was significantly lower (indicated by higher values) for moderate probabilities in comparison to low probabilities, $t(69)_{\text{frequency}} = -5.29$, $p < .001$, $r = .54$ (low probabilities mean = 2.30, $SD = 5.09$; moderate = 11.10, $SD = 8.54$); $t(75)_{\text{percent}} = -3.66$, $p < .001$, $r = .39$ (low probabilities mean = 5.12, $SD = 8.08$;

moderate = 11.87, *SD* = 8.06). The visual analog scale's accuracy did not differ between ranges, $p = .691$.

Turning next to the graphical encoding condition (Figure 5B), accuracy means (standard deviations) across context variables also decreased from the visual analog scale to the numeric measures: VAS = 7.54 (10.56), frequency = 1.12 (3.74), and percent = 0.45 (1.56). An ANOVA with accuracy as the dependent variable and format, probability range, and severity as between-subjects factors confirmed a significant effect of scale format, $F(2, 207) = 25.74$, $p < .001$, $\eta_p^2 = .20$. Again in line with expectations (H2), the numeric formats formed a homogenous subgroup (Bonferroni's test, all *p*s < .001). The context variables, range and severity, had no effect on scale accuracy in the graphical condition.

Finally, we compared the accuracy of the scale formats between encoding conditions (Figures 5A and 5B) using a mixed-design ANOVA with format as a between-subjects factor and encoding mode as a within-subjects factor. We found a significant effect of encoding condition, indicating greater accuracy of all formats when objective probabilities were encoded graphically in comparison to sequentially, $F(1, 216) = 162.92$, $p < .001$, $\eta_p^2 = .43$. We also observed a significant interaction between scale formats and encoding condition due to differences between encoding conditions varying in magnitude between scale formats, $F(2, 216) = 4.01$, $p = .02$, $\eta_p^2 = .04$.

In summary, Hypothesis 2 was supported: The numeric formats were generally more accurate than the visual analog scale. Further, we found additional support for the notion that, beyond characteristics of the measures, differences in goodness-of-fit between scales also vary as a function of encoding error. This was indicated by the drop in accuracy for moderate probabilities in the sequential condition as compared to the graphical encoding condition. However, we only observed this effect for the numeric formats, whereas the visual analog scale showed no such decrease.

Context Dependency

We manipulated the context of the objective probabilities along two dimensions: the range of probabilities and the severity of the events to which the probabilities pertain.

The range of probabilities. While the theoretical range of probabilities (0–100) is known beforehand and should therefore not affect likelihood judgments, we speculated that the distribution of actually presented probabilities would. To test this assumption, we varied the range of objective probabilities presented to the participants. Normatively, the difference between ranges should map onto a scale. That is, even though only probabilities of up to 20%

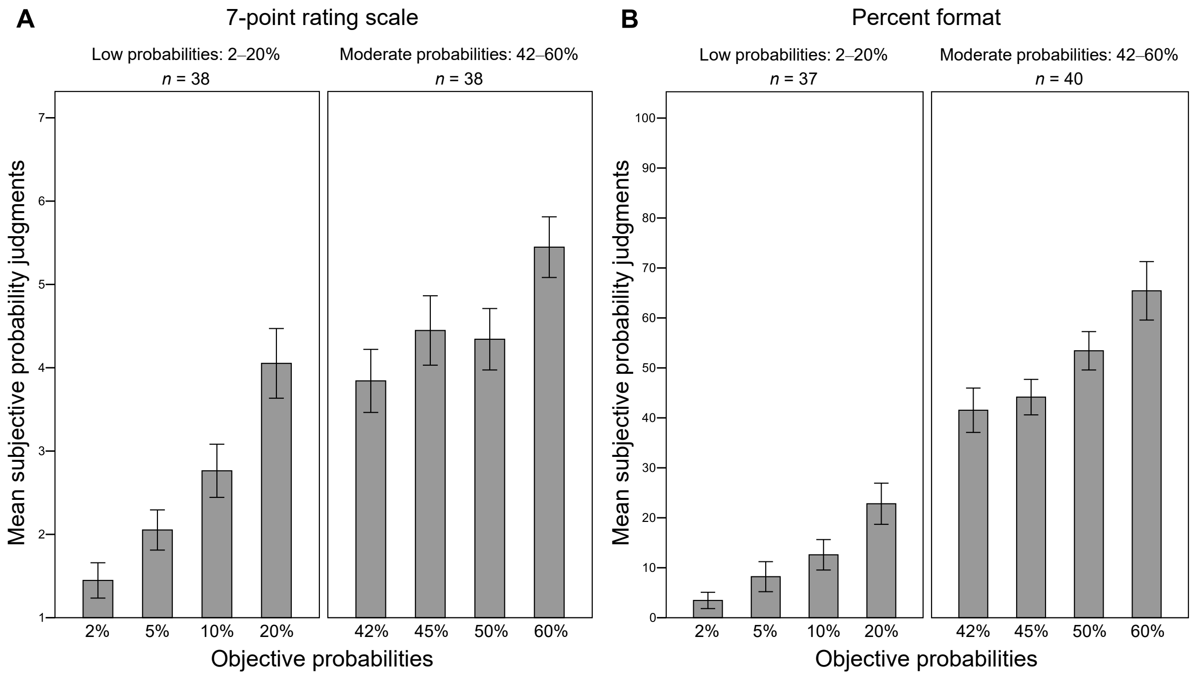

Figure 6. Mean subjective probability ratings on the 7-point rating scale (A) and on the percent format (B) in the sequential encoding condition. Error bars represent 95% confidence intervals.

were experienced, a probability of, for example, 42% should still receive a significantly higher value on that same scale when used by a different individual. We expected that the numeric measures would succeed in this because of their precisely defined categories. The rating scales and the visual analog scale, however, were expected to succumb to this context effect due to the vagueness of quantifiers and imprecision in handling, respectively (H3).

Beginning with the sequential encoding condition, Figures 6A and 6B exemplify differential mappings of probabilities on the 7-point verbal scale and the percent format, respectively. To assess the formats' performance, we consider the absolute judgments of the highest low probability (20%) and the lowest moderate probability (42%) that were given between subjects.[9] Mean ratings (standard deviations) of the respective 20%- and 42%-ratings were: 7-point = 4.05 (1.27) vs. 3.84 (1.15); 11-point = 3.87 (1.88) vs. 4.39 (1.91); VAS = 36.46 (22.41) vs. 44.09 (21.80); frequency = 0.19 (0.03) vs. 0.43 (0.13); percent = 22.81 (12.35) vs. 41.53 (13.88). In line with our prediction, these values suggest that neither the rating scales nor the visual analog scale differentiated between probability

[9] Because means did not vary as a function of severity, we collapsed the data over both severity conditions.

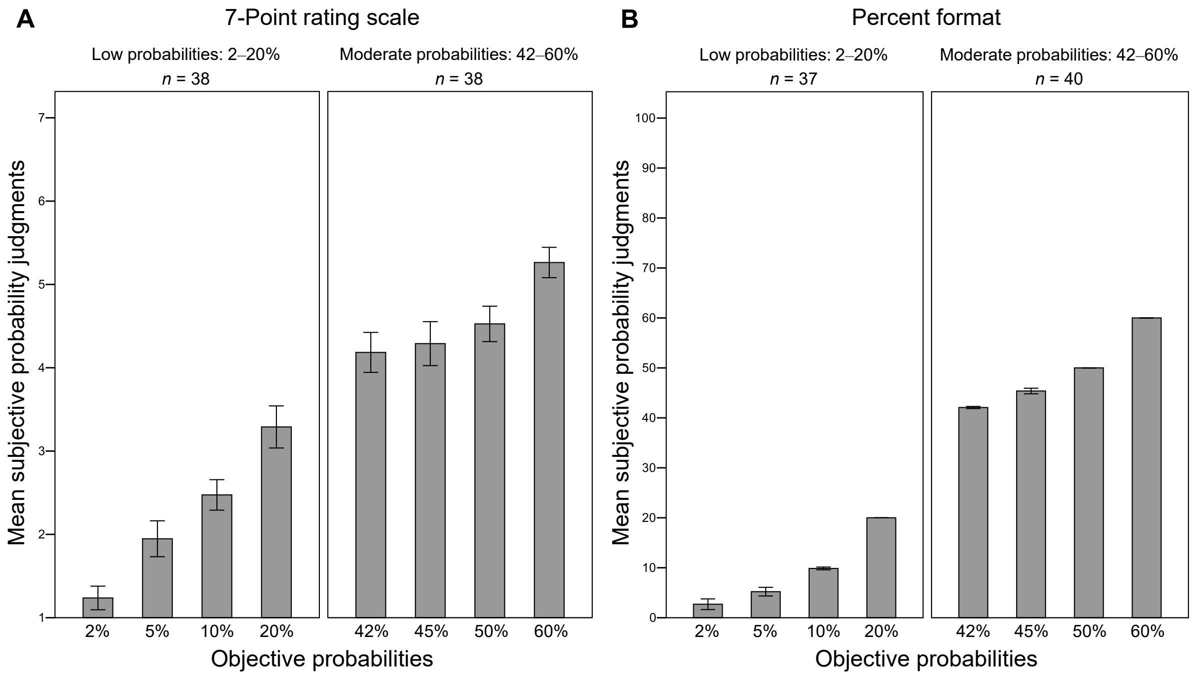

Figure 7. Mean subjective probability ratings on the 7-point rating scale (A) and on the percent format (B) in the graphical encoding condition. Error bars represent 95% confidence intervals.

ranges. This is most evident in the case of the 7-point verbal scale, which delivered an average rating for a probability of 42% that was actually lower than that for 20%. To test our hypothesis, we calculated independent *t*-tests for each of the value pairings and found that judgments made on the scales in question did not differ significantly, $t(74)_{\text{7-point}} = 0.76$, $p = .45$; $t(74)_{\text{11-point}} = -1.19$, $p = .24$; $t(74)_{\text{VAS}} = -1.45$, $p = .15$.[10]

In the graphical encoding condition, on the other hand, context dependency was markedly reduced. Figures 7A and 7B present means of the 7-point rating scale and the percent format. Means (standard deviations) of the respective 20%- and 42%-ratings were: 7-point = 3.29 (0.77) vs. 4.18 (0.73); 11-point = 2.62 (1.16) vs. 4.10 (0.50); VAS = 23.78 (14.86) vs. 46.88 (14.38); frequency = 0.20 (0.00) vs. 0.42 (0.03); percent = 20.00 (0.00) vs. 42.01 (0.66). *T*-tests confirmed that all scales differentiated significantly between the two ranges of probabilities (all *p*s < .001). These results indicate once more the role of encoding error with regard to differences in scale performance.

[10] The exclusion of outliers did not change the results of the *t*-tests in this section.

The severity of outcomes. We predicted that greater severity would lead to higher probability judgments on the rating scales and the visual analog scale (H4). To test this, we calculated mixed-design ANOVAs with the four respective probability judgments as a within-subjects factor and severity as a between-subjects factor. Analyses were conducted separately for each scale format (due to different scale units) as well as for both probability ranges (because ranges were manipulated between subjects and inevitably affected probability ratings). Contrary to our prediction, probability judgments did not differ systematically as a function of the severity of observed outcomes in either of the encoding conditions.[11]

Discussion

We evaluated five different scale formats for measuring subjective probability estimates with regard to two goodness-of-fit criteria pertaining to the concordance between objective and subjective probabilities. We varied the range of probabilities and the severity of outcomes and provided objective probabilities in both a sequential and an aggregated graphical format.

We observed that numeric scales for assessing subjective probability judgments generally fared better than rating scales or a visual analog scale in terms of sensitivity and showed higher accuracy than a visual analog scale. This superior performance is most likely due, in part, to the numeric scales' higher resolution in terms of possible categories, which readily offer one category for each possible probability. The two rating scales, on the other hand, specify categories in such a manner that more than one of the presented probabilities fall within the range of one category. That is, probabilities of 2% and 5% may both be described as *very small*; even more pronounced in the case of the 11-point rating scale, 2%, 5% and 10% fall between the categories *00* and *01*. Finally, the somewhat mixed performance of the visual analog scale might be explained by difficulties in its use (Couper et al., 2006). While the scale does assign every position a score between 0 and 100, thereby theoretically allowing for a perfect goodness-of-fit, it offers no numeric feedback.

[11] After the exclusion of outliers we found significantly higher probability estimates for sequentially encoded high severity adverse events in the low probabilities condition on the 11-point scale, $F(1, 29) = 6.64$, $p = .015$, $\eta_p^2 = .19$ In the graphical condition we found significantly higher probability estimates for low probability high severity events on the 7-point scale, $F(1, 30) = 4.95$, $p = .034$, $\eta_p^2 = .14$ as well as on the 11-point scale $F(1, 33) = 6.41$, $p = .016$, $\eta_p^2 = .16$.

However, goodness-of-fit was determined not only by the scales' resolution but also varied as a function of the encoding of objective probabilities. We observed consistently higher sensitivity and accuracy when probability information was presented simultaneously. The sampling of relative event frequencies from a sequence of events inevitably contains a certain amount of noise (Erev et al., 1994; Fiedler, 2000; Fiedler & Armbruster, 1994). Further, translating single occurrences of an outcome into a judgment of frequency involves a mathematical transformation. Both aspects of sequential encoding might introduce error into estimates. The graphical presentation employed in the present research, on the other hand, allowed participants to simply count the relevant outcomes and, at the same time, provided a numeric reference scale that quantified all possible outcomes. This type of presentation format has been shown to decrease base-rate neglect (Obrecht et al., 2009) and improve probability weighting in risky choices in comparison to sequential presentation (Hilbig & Glöckner, 2011).

The importance of error in available probability information becomes more apparent when considering the effect of the two probability ranges presented. Sampling the probability of an event that is almost as equally frequent as its opposite (moderate probabilities condition) should be more error-prone than spotting a few highly salient events. Accordingly, we observed generally lower sensitivity following the sequential encoding of higher probabilities. In addition, the context dependency we observed, that is, the rating scales' inability to differentiate between the two ranges of probabilities, disappeared in the graphical condition. Had this effect been caused only by inherent attributes of the scales, it would have appeared in both encoding conditions. Hence, it seems prudent to conclude that the precision of encoded information plays an important role in the goodness-of-fit of subjective probability measures.

Against our expectations and previous research (Harris et al., 2009), none of the scales were in any way affected by the severity of outcomes. This lack of effect might be explained by the design of the study. The presentation of relative frequencies of outcomes may have induced a general sense of rule-based reasoning and focused participants' attention on delivering accurate magnitude judgments. The scale-inherent performance differences in the present study are thus most likely caused by the resolution of measures rather than their verbal or numerical quality. Single comparisons of the two rating scales confirmed this notion, indicating no differences between the verbal 7-point and the numeric 11-point rating scale on any of the dependent variables. Hence, it seems likely that in the present context all scales were interpreted and used as measures of probability rather than measures of general risk or concern (cf. Borland, 1997).

An explanation for the notion that scale resolution becomes less relevant with less reliable available information might be that the differing judgments were based on different representations. Fuzzy-trace theory posits that individuals extract two independent kinds of representations from information: verbatim and gist. While the verbatim representations are for surface form and include, for example, exact numbers, gist representations capture the meaning of information (Reyna, 2012). In the present context, this distinction can be thought of as the difference between absolute probability judgments and the ordinal relations between probabilities (Reyna & Hamilton, 2001). Thus, in the graphical encoding condition, participants had verbatim knowledge of the presented probability readily available and could therefore base their judgment exclusively on this highly accurate representation. Differences in performance were thus due to the scales' resolution either allowing for or preventing an exact expression of this representation. In contrast, following sequential encoding, verbatim representations of the presented probabilities were less available or even absent, while gist representations, for example, the general range (i.e., all small probabilities or all moderate probabilities) and rank order mostly remained (Zacks & Hasher, 2002). Consequently, judgments in this condition were rather gist-based. The data support this interpretation. First, all scales were very sensitive in all conditions, indicating that the rank order of the presented probabilities was available at the time of judgment. Second, in the graphical condition the numeric measures showed higher sensitivity because their high resolution allowed for quantifying the ordinal differences between scales. On the other hand, in the sequential condition, scales were not used for exact quantification (because that knowledge was not available) but merely to express the ordinal relation of the probabilities. Hence, the differences in scale resolutions were less important and the differences in sensitivity between scales markedly reduced.

The distinction between verbatim and gist representations finds another more profound application in risk research. It is generally assumed that, in addition to a person's belief about an objective probability of an outcome, there exists a more intuitive representation of uncertainty with regard to the occurrence of that outcome (Reyna, 2004). The latter concept has been shown to play a more important role in guiding decisions and behavior (Windschitl, 2002), which is in line with the earlier reported results from Weinstein et al. (2007). However, when research addresses biases in the perception of objective probabilities, measures are needed that tap into the beliefs about these probabilities. The high sensitivity and independence of severity that we found in the present research indicates that all scales were understood and used in such a manner.

On the other hand, the scale formats differed greatly in their performance with regard to mapping objective probabilities when the spacing of presented probabilities was skewed, that is, low and moderate ranges. This effect has important implications for research. When the encoding of probability information is error-prone, rating and visual analog scales are not suitable for between-subjects designs due to the dependence of scores on the distribution of stimuli. Even under conditions of perfect information, this effect, though somewhat reduced, is still observable. While all scales differentiated between low and moderate probabilities in the graphical condition, the gap between ranges—a difference of 22%—was not adequately mapped onto scales when compared to the within-subjects differences, for example, the 10%- and 20%-rating (Figures 7A and 7B).

Thus, not every scale format we investigated can be equally recommended for use in research. All scales were high in sensitivity and should allow for a meaningful ordinal ranking of a limited number of different probabilities. For the rating scales, the number of categories is an obvious upper limit. However, if the differences between probabilities are very diverse (e.g., 5%, 10%, 30%, 80%), researchers must keep in mind that these differences will not be represented on the scales. The performance of the visual analog scale was inconsistent. Despite its high resolution, its sensitivity was on par with the rating scales. On the other hand, its accuracy was unaffected by encoding error. One speculative explanation for this could be that accuracy for low probabilities was decreased beyond the effect of encoding error (accuracy was very low in both conditions) because judgments had to be made at the extreme end of the scale, which may have increased difficulties in use. Finally, the numeric measures offered the most precise probability ratings and proved to be unaffected by the distribution of the stimuli. Previous research has shown that the use of natural frequencies, instead of percentages, can improve reasoning about probabilities and decrease a number of commonly found biases (Reyna & Brainerd, 2008). In the present context, however, the frequency format showed a much stronger drop in sensitivity for moderate probabilities in the sequential condition than the percentage format. The fact that participants have to provide two values instead of just one might make this measure simply more error-prone—out of the 12 participants that had to be dropped from analyses, eight were dropped due to meaningless answers on the frequency measure.

The encoding conditions that we implemented were rather artificial, which might impede the generalization of results. While a format such as the pictograph might be applied within the context of risk communication, it is unlikely that the sequential encoding of outcomes relevant to a question would ever happen in such a controlled and blocked fashion.

Further, as one reviewer noted, participants may simply have counted the relevant outcomes in the sequential condition, thereby effectively canceling the difference between encoding conditions. While we cannot rule out this explanation, we believe that a counting strategy would simply have led to a quantitative reduction of the effect of encoding conditions by providing more reliable probability information. In addition, two participants had to be dropped from analyses due to keeping a written tally of relevant outcomes, indicating that it was most likely not possible to read out loud and simultaneously count the stimuli.

In conclusion, the measurement formats we compared differ markedly in terms of sensitivity, accuracy, and dependency on stimulus distributions. These differences are caused only in part by characteristics inherent to the scales (i.e. resolution, verbal vs. numeric)—factors that become less relevant with increasing encoding error. Therefore, when deciding which scale to use in research, the source of the probability information that informs the subjective probability judgments must be taken into account. In order to understand the processes that lead to typical biases in probability judgments, a scale is needed that measures probability alone. In the present research, even the verbally labeled 7-point rating scale delivered judgments that closely followed the presented objective probabilities. However, a meaningful quantification of probability judgments, that is, one that allows between-subjects comparisons, requires a numeric measure.

References

Betsch, C., Ulshöfer, C., Renkewitz, F., & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making, 31*(5), 742–753. https://doi.org/10.1177/0272989X11400419

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics, 15*(1), 89–96. https://doi.org/10.3758/BF03205834

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods, 4*(3), 243–249. https://doi.org/10.1037/1082-989X.4.3.243

Borgida, E., & Nisbett, R. E. (1977). The differential impact of abstract vs. concrete information on decisions. *Journal of Applied Social Psychology, 7*(3), 258–271. https://doi.org/10.1111/j.1559-1816.1977.tb00750.x

Borland, R. (1997). What do people's estimates of smoking related risk mean? *Psychology and Health, 12*(4), 513–521. https://doi.org/10.1080/08870449708406727

Brosius, H.-B., & Bathelt, A. (1994). The utility of exemplars in persuasive communications. *Communication Research, 21*(1), 48–78. https://doi.org/10.1177/009365094021001004

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*(3), 391–405. https://doi.org/10.1016/0749-5978(85)90007-X

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance, 14*(2), 281–294. https://doi.org/10.1037//0096-1523.14.2.281

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition, 58*(1), 1–73. https://doi.org/10.1016/0010-0277(95)00664-8

Couper, M. P., Tourangeau, R., Conrad, Frederick, G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales. *Social Science Computer Review, 24*(2), 227–245. https://doi.org/10.1177/0894439305281503

de Wit, J. B. F., Das, E., & Vet, R. (2008). What works best: Objective statistics or a personal testimonial? An assessment of the persuasive effects of different types of message evidence on risk perception. *Health Psychology, 27*(1), 110–115. https://doi.org/10.1037/0278-6133.27.1.110

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research, 8*(2), 181–192. https://doi.org/10.1093/her/8.2.181

Druzdzel, M. J. (1989). *Verbal uncertainty expressions: Literature review* (Technical Report CMU-EPP-1990-03-02). Retrieved from University of Pittsburgh website: http://www.pitt.edu/~druzdzel/psfiles/verbal.pdf

Eibner, F., Barth, J., Helmes, A., & Bengel, J. (2006). Variations in subjective breast cancer risk estimations when using different measurements for assessing breast cancer risk perception. *Health, Risk & Society, 8*(2), 197–210. https://doi.org/10.1080/13698570600677407

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519–527. https://doi.org/10.1037//0033-295X.101.3.519

Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: Is a picture worth a thousand statistics? *Medical Decision Making, 25*(4), 398–405. https://doi.org/10.1177/0272989X05278931

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making, 27*(5), 672–680. https://doi.org/10.1177/0272989X07304449

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review, 107*(4), 659–676. https://doi.org/10.1037//0033-295X.107A659

Fiedler, K., & Armbruster, T. (1994). Two halfs may be more than one whole: Category-split effects on frequency illusions. *Journal of Personality and Social Psychology, 66*(4), 633–645. https://doi.org/10.1037//0022-3514.66.4.633

Fischer, K., & Jungermann, H. (2003). "Zu Risiken und Nebenwirkungen fragen Sie Ihren Arzt oder Apotheker": Kommunikation von Unsicherheit im medizinischen Kontext ["For risks and side-effects please ask your doctor or pharmacist": Communication of uncertrainty in the medical context]. *Zeitschrift für Gesundheitspsychologie, 11*(3), 87–98. https://doi.org/10.1026//0943-8149.11.3.87

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Harris, A. J. L., Corner, A., & Hahn, U. (2009). Estimating the probability of negative events. *Cognition, 110*(1), 51–64. https://doi.org/10.1016/j.cognition.2008.10.006

Hasher, L., & Zacks, R. T. (1979). Automatic and effortful processes in memory. *Journal of Experimental Psychology: General, 108*(3), 356–388. https://doi.org/10.1037/0096-3445.108.3.356

Hawley, S. T., Zikmund-Fisher, B., Ubel, P., Jancovic, A., Lucas, T., & Fagerlin, A. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices. *Patient Education and Counseling, 73*(3), 448–455. https://doi.org/10.1016/j.pec.2008.07.023

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(4), 621–642. https://doi.org/10.1037/0278-7393.31.4.621

Hilbig, B. E., & Glöckner, A. (2011). Yes, they can! Appropriate weighting of small probabilities as a function of information acquisition. *Acta Psychologica, 138*(3), 390–396. https://doi.org/10.1016/j.actpsy.2011.09.005

Hinyard, L. J., & Kreuter, M. W. (2007). Using narrative communication as a tool for health behavior change: A conceptual, theoretical, and empirical overview. *Health Education & Behavior, 34*(5), 777–792. https://doi.org/10.1177/1090198106291963

Hurd, M. D., & McGarry, K. (1995). Evaluation of the subjective probabilities of survival in the Health and Retirement Study. *The Journal of Human Resources, 30*, S268–S292. https://doi.org/10.2307/146285

Obrecht, N. A., Chapman, G. B., & Gelman, R. (2009). An encounter frequency account of how experience affects likelihood estimation. *Memory & Cognition, 37*(5), 632–643. https://doi.org/10.3758/MC.37.5.632

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407–418. https://doi.org/10.1037/h0022602

Pruitt, D. G., & Hoge, R. D. (1965). Strength of the relationship between the value of an event and its subjective probability as a function of method of measurement. *Journal of Experimental Psychology, 69*(5), 483–489. https://doi.org/10.1037/h0021721

Rapoport, A., Wallsten, T. S., & Cox, J. A. (1987). Direct and indirect scaling of membership functions of probability phrases. *Mathematical Modelling, 9*(6), 397–417. https://doi.org/10.1016/0270-0255(87)90506-9

Reyna, V. F. (2004). How people make decisions that involve risk: A dual-processes approach. *Current Directions in Psychological Science, 13*(2), 60–66. https://doi.org/10.1111/j.0963-7214.2004.00275.x

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in fuzzy-trace theory. *Judgment and Decision Making, 7*(3), 332–359. Retrieved from http://journal.sjdm.org/11/111031/jdm111031.pdf

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences, 18*(1), 89–107. https://doi.org/10.1016/j.lindif.2007.03.011

Reyna, V. F., & Hamilton, A. J. (2001). The importance of memory in informed consent for surgical risk. *Medical Decision Making, 21*(2), 152–155. https://doi.org/10.1177/0272989X0102100209

Schapira, M. M., Davids, S. L., McAuliffe, T. L., & Nattinger, A. B. (2004). Agreement between scales in the measurement of breast cancer risk perceptions. *Risk Analysis, 24*(3), 665–673. https://doi.org/10.1111/j.0272-4332.2004.00466.x

Schwarz, N., & Wänke, M. (2002). Experiential and contextual heuristics in frequency judgement: Ease of recall and response scales. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency processing and cognition* (pp. 89–108). Oxford, England: University Press. https://doi.org/10.1093/acprof:oso/9780198508632.003.0006

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes, 80*(2), 155–190. https://doi.org/10.1006/obhd.1999.2857

Ubel, P. A., Jepson, C., & Baron, J. (2001). The inclusion of patient testimonials in decision aids: Effects on treatment choices. *Medical Decision Making, 21*(1), 60–68. https://doi.org/10.1177/0272989X0102100108

van der Pligt, J. (1996). Risk perception and self-protective behavior. *European Psychologist, 1*(1), 34–43. https://doi.org/10.1027/1016-9040.1.1.34

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 613–625. https://doi.org/10.1037/0096-1523.16.3.613

Verplanken, B. (1997). The effect of catastrophe potential on the interpretation of numerical probabilities of the occurrence of hazards. *Journal of Applied Social Psychology, 27*(16), 1453–1467. https://doi.org/10.1111/j.1559-1816.1997.tb01608.x

Viscusi, W. K., & Hakes, J. (2003). Risk ratings that do not measure probabilities. *Journal of Risk Research, 6*(1), 23–43. https://doi.org/10.1080/1366987032000047789

Wallsten, T. S., Budescu, D., & Erev, I. (1988). Understanding and using linguistic uncertainties. *Acta Psychologica, 68*(1–3), 39–52. https://doi.org/10.1016/0001-6918(88)90044-3 Get

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review, 10*(1), 43–62. https://doi.org/10.1017/S0269888900007256

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*(4), 348–365. https://doi.org/10.1037//0096-3445.115.4.348

Wänke, M. (2002). Conversational norms and the interpretation of vague quantifiers. *Applied Cognitive Psychology, 16*(3), 301–307. https://doi.org/10.1002/acp.787

Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55*(3), 341–356. https://doi.org/10.1037/0022-3514.55.3.341

Weinstein, N. D. (1993). Testing four competing theories of health-protective behavior. *Health Psychology, 12*(4), 324–333. https://doi.org/10.1037/0278-6133.12.4.324

Weinstein, N. D., Kwitel, A., McCaul, K. D., Magnan, R. E., Gerrard, M., & Gibbons, F. X. (2007). Risk perceptions: Assessment and relationship to influenza vaccination. *Health Psychology, 26*(2), 146–151. https://doi.org/10.1037/0278-6133.26.2.146

Windschitl, P. D. (2002). Judging the accuracy of a likelihood judgment: The case of smoking risk. *Journal of Behavioral Decision Making, 15*(1), 19–35. https://doi.org/10.1002/bdm.401

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364. https://doi.org/10.1037//1076-898X.2.4.343

Woloshin, S., Schwartz, L. M., Black, W. C., & Welch, H. G. (1999). Women's perceptions of breast cancer risk: How you ask matters. *Medical Decision Making, 19*(3), 221–229. https://doi.org/10.1177/0272989X9901900301

Woloshin, S., Schwartz, L. M., Byram, S., Fischhoff, B., & Welch, H. G. (2000). A new scale for assessing perceptions of chance: A validation study. *Medical Decision Making, 20*(3), 298–307. https://doi.org/10.1177/0272989X0002000306

Yates, J. F., & Stone, E. R. (1992). The risk construct. In J. F. Yates (Ed.), *Wiley series in human performance and cognition. Risk-taking behavior* (pp. 1–25). Oxford England: John Wiley & Sons.

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency processing and cognition* (pp. 21–36). Oxford, England: University Press. https://doi.org/10.1093/acprof:oso/9780198508632.003.0002

Appendix A: Adverse Events and Names of Medications Used in the Experiment

Adverse Events		
Low Severity	High Severity	Medications
Fever	Asthma	Argal
Dizziness	Diabetes	Tirav
Rash	Meningitis	Tookl
Restlessness	Epilepsy	Endas
Vomiting	Cerebral Edema	Vahto
Headache	Blindness	Tigla
Cough	Autism	Tsovir
Muscle Aches	Neurodermatitis	Drigul

Appendix B: Analysis of Scales' Accuracy Calculated with Predicted Subjective Probability Values

All analyses were additionally carried out under exclusion of outliers. Outliers were defined as any value above and below two standard deviations within each cell. All of the effects remained significant after the exclusion of outliers, effect sizes generally increased.

Means (*SD*s) across context variables in the sequential encoding condition: 7-point = 3.69 (1.88), 11-point = 3.59 (1.90), VAS = 3.44 (1.99), frequency = 2.37 (2.31), and percent = 2.38 (1.84). An ANOVA with accuracy as the dependent variable and format, probability range, and severity as between-subjects factors found a main effect of format, $F(4, 352) = 10.51$, $p < .001$, $\eta_p^2 = .11$. Bonferroni's test revealed that the numeric formats formed one of two homogenous subgroups and outperformed the remaining three measures in terms of accuracy, all ps $\leq .005$. Further, accuracy decreased with increasing probability, $F(1, 352) = 59.60$, $p < .001$, $\eta_p^2 = .15$ (low probabilities mean = 2.40, $SD = 1.92$; moderate = 3.81, $SD = 1.96$), and increasing severity, although the latter effect was weak, $F(1, 352) = 4.60$, $p = .033$, $\eta_p^2 = .01$ (low severity mean = 2.92, $SD = 1.91$; high = 3.28, $SD = 2.19$). Finally, there was a significant interaction of format and range, $F(4, 352) = 4.95$, $p = .001$, $\eta_p^2 = .05$.

Two separate ANOVAs, comparing accuracy between scale formats for each range of probabilities separately (and collapsing data across levels of severity because of that manipulation's very small effect, $\eta_p^2 = .01$) found that, when probabilities were low, accuracy differed significantly between formats, $F(4, 182) = 12.78$, $p < .001$, $\eta^2 = .22$, with the numeric measures and the remaining three scales forming two separate subgroups. However, the difference between the 7-point rating scale and the percent format was only marginally significant (Bonferroni's test, $p = .086$). In contrast, the small, significant effect in the moderate probabilities condition, $F(4, 180) = 3.62$, $p = .007$, $\eta^2 = .07$, is explained by a single significant difference between the 7-point rating scale and percent format (Bonferroni's test, $p = .002$); all other comparisons were non-significant.

Means (standard deviations) across context variables in the graphical encoding condition: 7-point = 3.05 (1.36), 11-point = 2.39 (1.25), VAS = 1.92 (1.48), frequency = 0.51 (1.40), and percent = 0.33 (1.15). An ANOVA with accuracy as the dependent variable and format, probability range, and severity as between-subjects factors found a significant effect of scale format, $F(4, 350) = 59.27$, $p < .001$, $\eta_p^2 = .40$. The numeric formats formed one homogenous subgroup, the visual analog scale and the 11-point rating scale formed a separate subgroup that differed from the 7-point rating scale (Bonferroni's test, all *p*s $\leq .023$). The context variables, range and severity, had no effect on scale accuracy in the graphical condition.

Comparison of accuracy between encoding conditions, using a mixed-design ANOVA with format as a between-subjects factor and encoding mode as a within-subjects factor, found a significant effect of encoding condition, indicating greater accuracy of all formats when objective probabilities were encoded graphically in comparison to sequentially, $F(1, 364) = 149.17$, $p < .001$, $\eta_p^2 = .29$. A significant interaction between scale formats and encoding condition is due to differences between encoding conditions varying in magnitude between scale formats, $F(4, 364) = 4.73$, $p = .001$, $\eta_p^2 = .05$.

Article 4

Self-Report Measures of Subjective Probability: Error and Anchor Effects

Reference:

Haase, N., & Betsch, T. (2016). *Self-report measures of subjective probability: Error and anchor effects.* Manuscript in preparation

Self-Report Measures of Subjective Probability: Error and Anchor Effects

Niels Haase and Tilmann Betsch
University of Erfurt, Germany

Author Note

Niels Haase, Department of Psychology, University of Erfurt, Germany; Tilmann Betsch, Department of Psychology, University of Erfurt, Germany.

Niels Haase is now at Department of Psychology, University of Konstanz, Germany.

This research was financed, in part, by a research grant from the German Research Foundation (BE 3970/4-1) to Cornelia Betsch and Frank Renkewitz, both of the Department of Psychology, University of Erfurt and, in part, by a research grant for young researchers from the Faculty of Education at the University of Erfurt to the first author. The authors are grateful to Cornelia Betsch and Frank Renkewitz for providing some of the funding, to Philipp Schmid and Lina Gerold for their help in conducting the third experiment, to Johannes Ritter for many helpful conversations and comments, and to the members of the Center for Empirical Research in Economics and Behavioral Sciences (CEREB) at the University of Erfurt for providing ample opportunity to critically discuss the subject matter.

Correspondence concerning this article should be addressed to Niels Haase, Department of Psychology, Social Psychology and Decision Sciences, University of Konstanz, P.O. Box 43, 78457 Konstanz, Germany. E-mail: niels.haase@uni-konstanz.de

Abstract

Recent evidence indicates that instruments for the assessment of subjective probability do not just differ due to scale-inherent characteristics, but also as a function of the error in the underlying representations. In three experiments subjects encoded the same two ranges of objective probabilities as sequences of option-outcome pairs and judged subjective probability either on a verbally labeled 7-point rating scale or in the form of percent estimates. The two ranges shared one common stimulus and were presented within-subjects. In Experiment 1 we observed that under highly error-prone conditions the high resolution of the percent format offers no advantage in terms of sensitivity. Additionally, imprecise representations, rather than the stimulus distribution, can result in apparent context effects on the rating scale. In Experiment 2 we found that an anchor in the stimuli can tether judgments on the rating scale to the scale's midpoint but in turn result in inconsistent judgment functions on the percent format that are an expression of regression toward the mean. In Experiment 3 we discovered that imprecise representations change the way the rating scale is used while percent estimates remain consistent. We conclude that a verbal rating scale does not allow a meaningful quantification or meaningful comparisons between experimental conditions and should not be used in research on subjective probability. The percent format captures the underlying representations reliably and consistently but is very sensitive to noise and can lead to classic regression fallacies.

Keywords: Subjective probability, scale sensitivity, context effects, anchor effects, regression toward the mean

Self-Report Measures of Subjective Probability: Error and Anchor Effects

Subjective probability[1] is a central variable for a large number of research questions. However, comparisons across studies are hindered because there is no standard of measurement. In fact, the question of how to assess perceptions of likelihood has been the focus of research in its own right for at least 50 years. Elicitation approaches range from inferring probabilities from behavior, such as bets (Beach & Phillips, 1967; Beach & Wise, 1969) or choices between lotteries (Baillon, 2008; Edwards, 1962) to a great variety of self-report instruments. The latter include complex methods, such as eliciting fractiles or even whole distributions (bins-and-balls method) for multinomial or continuous variables (e.g., Delavande & Rohwedder, 2008; Goldstein & Rothschild, 2014) and a plethora of different formats for direct estimation, especially in case of binomial distributions, that ask for some kind of magnitude judgment (Bilgin, 2012; Diefenbach, Weinstein, & O'Reilly, 1993; Galanter, 1962; Woloshin, Schwartz, Black, & Welch, 1999; Woloshin, Schwartz, Byram, Fischhoff, & Welch, 2000).

Similarly, very different evaluative criteria have been employed to evaluate these instruments. Subjective probability scales have been studied in terms of usability, confidence in judgment, test-retest reliability, and with regard to behavior prediction (Diefenbach et al., 1993; Weinstein et al., 2007). These criteria are certainly informative. However, many research questions address biases in probability judgments and to assess these biases reliably, that is, to differentiate biases in the representation of probability from measurement biases, it is necessary to the study scale formats in terms of correspondence between objective and subjective probabilities.

Comparisons between probability judgments and real-life data, such as objective risk factors, have found numeric instruments to produce overestimations whereas verbal and comparative scales fare much better (Eibner, Barth, Helmes, & Bengel, 2006; Schapira, Davids, McAuliffe, & Nattinger, 2004; Woloshin et al., 1999). This approach to evaluate the performance of different scale formats, however, has one essential caveat. It assumes that the

[1] The term subjective probability is sometimes used to denote one specific interpretation of the concept of probability, that is, the subjectivist interpretation. This view understands probability as a personal degree of belief and denies that the idea of a true or objective probability can even be meaningful (de Finetti, 1970). In this work, however, we adhere to the frequentist position and interpret the true probability of an event as its relative occurrence in a reference set. We use the term subjective probability to refer to an estimate of objective probability.

subjects have an accurate perception of these probabilities, which cannot be known a priori. So, if smokers underestimate their survival chances it is impossible to tell how much of this apparent bias is caused by a biased perception and how much is caused by the scale format itself (e.g., Viscusi & Hakes, 2003). One way to circumvent this is to provide all subjects with the same information and to thus create a normative truth to which judgments can be compared.

Haase, Renkewitz, and Betsch (2013) employed such a paradigm to evaluate five self-report instruments for subjective probability in terms of sensitivity and vulnerability to context effects. Subjects encoded four different probabilities from two ranges (low and moderate, between-subjects) first as sequences of single events and second in aggregate form as pictographs. Following each presentation, subjects judged the encoded probability on one of the five formats. Haase et al. found two numeric scales to be more sensitive than two rating scales and a visual analog measure. Judgments on the numeric scales were also more consistent across the two probability ranges whereas the other formats each produced two very different judgment functions for the two stimulus contexts, for example, on a 7-point rating scale an objective probability of 20% (the highest probability in the low range) received a higher rating than an objective probability of 42% (the lowest probability in the moderate range).

However, the performance differences between scales differed themselves as a function of encoding error. When the available information became less precise the differences between scales in sensitivity decreased while differences in context dependency increased. The authors (Haase et al., 2013) interpreted these findings in terms of fuzzy-trace theory, which posits that when individuals encode information they create multiple representations along a continuum from verbatim and exact at one end to vague and gist-like at the other (Reyna & Brainerd, 1995). For numerical information this gist continuum can be thought of as a hierarchy of imprecision comparable to different levels of measurement, that is, from ratio to nominal (Reyna, 2012). Crucially, gist representations are not built upon verbatim representations but are encoded simultaneously and independently. Furthermore, the utilization of any one or more than one representation is task-dependent and individuals have a preference to rely on the most gist-like representation possible for any given problem (Corbin, Reyna, Weldon, & Brainerd, 2015; Reyna, 2012).

Within this theoretical framework Haase et al. (2013) argued that, when virtually error-free verbatim representations are available, a numeric scale with a high resolution allows for an exact quantification of differences between probabilities and as a consequence much more sensitive judgments than the few and by definition equidistant categories of a rating scale. At the same time, precise information prompts subjects who use a rating scale to place close probabilities, for example, 5% and 10%, in the same category to retain a coherent mapping of the theoretically defined 0–100 probability range onto the available categories and to thus reduce sensitivity further.

On the other hand, when representations are less precise due to error-prone encoding, subjects tend to rely more on gist representations, such as the general range and rank order of the presented probabilities. As a consequence, the advantage of a high-resolution scale is reduced and the differences in sensitivity between scale formats decrease. At the same time, in order to express ordinal relations, subjects using a rating scale tend to use more categories rather than preserve a judgment range consistent with the underlying probability range. For example, on the 7-point rating scale, due to its limited number of categories, this led to the apparent context effect of an objective probability of 20% receiving a higher judgment than one of 42%. A numeric format, on the other hand, offers enough (or virtually unlimited) categories and thus allows for a much more coherent mapping, even if judgments are mostly based on ordinal ranks. The authors concluded that researchers deciding on a scale format need to take the expected precision of the representations that inform subjective probability judgments into account as it affects how a scale will be used. In most applied contexts all tested formats will allow to assess perceived ordinal relations between probabilities but a meaningful quantification of probability judgments for between-subjects comparisons calls for a numeric instrument.

This research follows up on these results. In three experiments we compared the best and the worst scale formats from that study—a percent estimate and a verbally labeled 7-point rating scale—to further explore the effects of error-prone encoding of probability information on scale sensitivity and the occurrence of context effects. The sensitivity of a scale quantifies the degree to which changes in the objective probabilities are mirrored by changes in subjective judgments. The occurrence of context effects served as a coarse indicator of accuracy. Typically, accuracy is assessed as deviations in judgments from an objective norm. This kind of calculation, however, requires the compared values to occupy identical value arrays, that is, 0–100 in this case. We address the question of transforming rating scale values to a 0–100 array in more detail in an additional analysis after reporting the experiments. However, to anticipate our conclusion, we believe that such a transformation is not appropriate and therefore refrain from assessing scale accuracy this way. On the other hand, probabilities are bounded at 0 and 100% (or 1) and have a natural anchor point at 50% (or

0.5). Irrespective of any inter-context incoherence, if a scale does not coherently map these anchors, it cannot be interpreted as quantifying subjective probabilities in any meaningful way beyond the ordinal scale. Therefore, we created a research paradigm that would allow for a direct test of this assumption. Before we explain our design in more detail we provide a review of relevant issues and findings.

Scale Formats

The scale formats we investigated differ in at least two aspects that are relevant for our evaluative criteria. First and most important is the format's resolution: the 7-point rating scale offers the subject seven discrete categories to make a judgment while the percent format allows for quasi-continuous estimates (judgments were restricted to integers in the 0–100 range). The question of the optimal number of response categories for rating scales has been discussed for a century and there appears to be a tentative consensus, that more than seven categories (plus or minus two) provide little gain in discriminative power (Cox, 1980; Diefenbach et al., 1993; McKelvie, 1978; Preston & Colman, 2000). However, a scale for probability judgments must not only be able to differentiate between probabilities but also relate judgments meaningfully to the underlying probability continuum. Depending on the precision of the representations this may present a conflict of interest to a larger or smaller degree. For instance, given perfect knowledge a subject might reasonably rate the objective probabilities of 5% and of 8% both as *very small* on the 7-point scale even though she is aware of the difference.

On the other hand, it is as unrealistic to assume perfect knowledge in any applied context as it is to even ask for such a judgment when a known probability could simply be reproduced (unless one is interested in how one scale translates to the other). In fact, some researchers have argued that judgments of probability which do not involve any actual calculations are based on a coarse internal representation with a limited number of states or categories (Sun, Wang, Zhang, & Smith, 2008). In line with this, Diefenbach et al. (1993; Weinstein & Diefenbach, 1997) found that more than seven scale categories did neither improve the correlations between probability judgments and risk factors nor the agreement of ranks derived from probability judgments with the direct ranking of likelihoods. Furthermore, the 7-point rating scale was much preferred by subjects in terms of perceived accuracy and ease of use.

Assuming coarse representations, there is then an obvious downside to the high resolution of the percent format: The requirement to provide a precise estimate can add additional noise to subjective probability judgments. Imagine two subjects judging the same three probabilities. Both know the probabilities' rank order and that they are below 50%. If these subjects make their judgments on the 7-point rating scale they will both in all likelihood assign the categories 1, 2, and 3. If, on the other hand, they provide their judgments as percent estimates, one of them might judge the probabilities as 10%, 20%, and 30%, the other as 8%, 22%, and 36%. Thus, even though on the individual level both subjects' judgments are equivalent in terms of sensitivity to the objective stimuli, on the aggregate level judgments may contain additional noise due to the scale's high resolution.

The issue is further complicated when considering the distribution of the objective probabilities. Equidistant or approximately equidistant stimuli will benefit the rating scale when representations are imprecise while a high-resolution scale format facilitates unevenly spaced judgments, and thus additional noise. On the other hand, a highly skewed stimulus distribution can be easily mapped with percent estimates but creates the above described conflict for users of the rating scale which increases with an increasing precision of representations. Thus, a scale's resolution is relevant for its sensitivity to subjective probability representations as well as to its ability to meaningfully map the probability array. Further, it stands to reason that this relevancy changes not only as a function of the stimulus distribution but also of the accuracy of said representations.

Second, our scale formats differ in that the 7-point rating scale provides verbal quantifiers as category labels whereas the percent estimate is purely numeric. Verbal quantifiers are easy to understand but inevitably somewhat vague. There exists a large body of research on the relation between numerical and verbal probability expressions (e.g., Bocklisch, Bocklisch, & Krems 2010; Budescu & Wallsten, 1985; Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986) which we address at the end of the article. However, it has also been argued that numeric scales promote deliberate and rule-based thinking whereas verbal probability measures induce an associative and intuitive reasoning style. Thus elicited and sometimes called intuitive perceptions of certainty do not just comprise beliefs in objective probabilities but may also include notions of the value of a prospect, its meaning in a given situation, and affective reactions toward it (Windschitl, 2002; Windschitl & Wells, 1996). For instance, a 10% chance of rain might be judged as *small* while a 10% likelihood of dying during a surgical procedure might reasonably be judged as *large*. Percent estimates of both probabilities, on the other hand, would be expected to only reflect the actual 10%. In line with this reasoning, verbal rating scales of subjective probability as compared to numerical estimates have been found to be better predictors of behavioral intentions and actual behavior that, of course, have more antecedents than just the likelihood of an outcome (Baghal, 2011;

Betsch, Haase, Renkewitz, & Schmid, 2015; Weinstein et al., 2007; Windschitl & Wells, 1996). However, if a rating scale is used to express perceptions about aspects of an uncertain prospect other than its likelihood, then one can exclude this format a priori as a viable instrument to assess biases in subjective probability, unless all these aspects were to be controlled for. It appears, though, that the degree to which a rating scale goes beyond pure probability depends on the experimental setting (Betsch et al., 2015; Haase et al., 2013). We address this issue in Experiment 1.

Context Effects

The domain of an uncertain prospect, for example cancer vs. rain, is but one part of the context of the subjective probability judgment. Research into how the context can exert an unwanted (as in not being of interest to the person asking for the judgment) influence on a judgment has focused on many aspects, such as formal features of instruments and preceding questions (e.g., Belli, Conrad, & Wright, 2007; Schwarz & Sudman, 1992). When we address context effects in this article, however, we specifically refer to the influence of the context created by the stimulus distribution itself. Such effects have most extensively been studied within the framework of range-frequency theory (Parducci, 1965). Originally developed for categorical judgments of multiple stimuli, the theory makes two basic assumptions. First, subjects match the range of available categories to the range of presented stimuli with the most extreme stimuli occupying the two respective end-categories and the remaining categories assigned to equal subranges of the stimuli (the range principle). Second, subjects tend to assign the same number of stimuli to each available category (the frequency principle). If stimuli are distributed uniformly the frequency principle has no effect. However, if for example the distribution is positively skewed, that is, smaller stimuli are presented with greater frequency, then some of these smaller stimuli will have to be placed in larger categories leading to a steeper judgment function. As a result, an intermediate stimulus will receive a higher judgment than when the stimulus distribution is negatively skewed. The final judgment presents a compromise between both principles.

Range-frequency theory has been used to explain context effects on categorical ratings in a multitude of domains (e.g., size, happiness, attractiveness, strength of mental disorder, student evaluations) but also on magnitude estimations, salary allocations and other number estimates (Mellers & Birnbaum, 1982, 1983; Wedell & Parducci, 1988, 2000; Wedell, Parducci, & Lane, 1990). Additionally, it has been argued that a single stimulus can invoke its own context. For example, Birnbaum (1999) showed that in a between-subjects design the number 9 was judged to be subjectively larger than the number 221.

Since both principles of the theory describe a system of assigning categories to stimuli, a major focus of research and a point of contention (e.g., McKelvie, 2001; Wedell, 1990) has been the importance of the scale format used for judgments in creating context effects. In general, a larger number of judgment categories reduces the context effects through both principles (Parducci, 1982; Parducci & Wedell, 1986; Wedell & Parducci, 2000; Wedell et al., 1990). We address the specific findings that are relevant for our design below.

More importantly though is the question of whether range-frequency theory is applicable to probability judgments. Context effects through the frequency principle have been observed for judgment functions with up to four anchors (Birnbaum, 1974). However, besides providing anchors (i.e., 0, 50, and 100%), it has been argued that judgments of proportion and by extension probability represent a qualitative rather than a quantitative continuum (Stevens & Galanter, 1957) or an absolute scale and thus might not lend themselves to a relativistic interpretation. To our knowledge, only one study has addressed this issue. Varey, Mellers, and Birnbaum (1990) presented subjects with arrays of black and white dots and elicited percent estimates of the respective proportions. They found that skewed stimulus distributions changed the relation between judged and actual proportions and urged caution when interpreting between-subjects comparisons of subjective probability judgments.

Systematic Effects of Random Error

Interestingly, Varey et al. (1990) also noted that the judgments in their study were regressive (i.e., small proportions were overestimated, large ones underestimated) and pointed out that such a pattern could account for the typical finding of overconfidence in research on calibration if one were to plot subjective probabilities as a function of objective probabilities instead of the customary other way around. This idea was idea was implemented by Erev, Wallsten, and Budescu (1994) and independently developed by Pfeifer (1994) who argued that at least part of the often observed overconfidence could in fact be a failure to recognize and thus a misinterpretation of regression toward the mean. This approach sparked the development of numerous so-called stochastic or random-error models which share the assumption that various biases in probabilistic reasoning are due to effects of random variation or noise in the reasoning process but differ in the assumed error distributions (e.g., Budescu, Erev, & Wallsten, 1997; Juslin, Olsson, & Björkman, 1997; Wallsten & González-Vallejo, 1994). More recently, broader frameworks have been suggested to reconcile these different models (Costello & Watts, 2014; Hilbert, 2012).

Regression toward the mean has flummoxed social scientists since it was first observed by Galton in 1885 and has been subject to what some researchers have called a mythologization (Maraun, Gabriel, & Martin, 2011). Therefore, it is important to note that regression toward the mean does not affect or explain anything. It is simply a property of a bivariate distribution if the two variables are not perfectly correlated (see Campbell & Kenny, 1999 for an excellent introduction). The failure to recognize this, however, can lead to a misinterpretation of aggregate data in terms of psychologically meaningful concepts such as overconfidence. Such misinterpretations, or regression fallacies, have predominantly been observed in studies with a pre- and post-test (e.g., Kahneman & Tversky, 1973; Verkooijen, Stok, & Mollen, 2015) but also in research on the availability heuristic (Sedlmeier, Hertwig, & Gigerenzer, 1998), frequency judgments (Hertwig, Pachur, & Kurzenhäuser, 2005), and the description-experience gap in risky choice (Glöckner, Hilbig, Henninger, & Fiedler, 2016).

Overview

We conducted three experiments to further study how error in the sampling of event frequencies affects different scale formats for the assessment of event probabilities in terms of scale sensitivity and the occurrence of context effects. In each experiment we presented the same two ranges of probabilities as sequences of option-outcome pairs. These ranges shared one common stimulus and were presented within-subjects. Probability judgments were made either on the verbally labeled 7-point rating scale or as percent estimates. We assessed scale sensitivity—defined as the Pearson correlation between objective and subjective probabilities—at the individual as well as the aggregate score level. Additionally, we analyzed the relative variability of judgments in relation to encoding conditions and scale sensitivity. We tested for context effects by comparing judgments of the common stimulus from the two probability ranges.

Experiment 1 served to establish the specific research paradigm and to test the scales' performance under highly error-prone encoding conditions. In Experiment 2 we examined the special function of the 50% probability midpoint as an anchor in judgments. In Experiment 3 we additionally varied the error in encoding to illustrate the interaction effect of anchor points and error on scale sensitivity and the occurrence of context effects. As the three experiments were very similar and followed nearly identical procedures, we describe the method and data analysis in detail for Experiment 1, while we confine those sections in the descriptions of Experiments 2 and 3 to changes and specific features as appropriate.

Experiment 1

This experiment addressed three research questions. First, in Haase et al. (2013) only percent estimates were still significantly more sensitive than the 7-point rating scale when the encoding of event frequencies had been the most difficult. However, even under those conditions, scale sensitivity was still rather high. We were interested to see how the scales fare, when the sampling process is extremely error-prone. As delineated above, we expected the percent format to lose all advantage in terms of sensitivity and possibly even to suffer due to its high-resolution demands.

Second, proponents of range-frequency theory have called the validity of between-subjects research into question if differing judgment contexts are not taken into account (Birnbaum, 1999; Varey et al., 1990). Therefore, we presented the two judgment contexts within-subjects. We reasoned that if the previously observed context effects were at least in part a function of the encoding error and the scale format rather than of the distribution of stimuli (see also below), than they should also occur in a within-subjects design. We expected higher ratings for the common stimulus in the low-range context as compared to the high-range context but percent estimates to be consistent across contexts.

Third, Haase et al. (2013) found no effects of events' severity on probability judgments and concluded that all scale formats had been used to express beliefs in objective probabilities rather than broader perceptions of certainty. We varied the judgment domain between-subjects and hoped to replicate this finding as this would allow us to interpret differences in scale performance to be due to characteristics of the scale formats rather than due to subjects expressing different concepts.

Method

The experiment implemented a 2 (scale format: 7-point rating scale vs. percent estimate) × 2 (domain: adverse drug reaction vs. crime) between-subjects design with one within-subjects factor (range of probabilities: low vs. high).

Procedure. Upon arriving in the lab subjects signed a consent form and were escorted to a soundproof cubicle. They were instructed to turn off all communication devices and to not use any kind of aid while participating, for example, taking notes. All additional instructions were provided on the screen in a standardized manner.

In the adverse drug reaction conditions subjects were asked to imagine that they wanted to take a medication and that five different drugs were available. The same adverse reaction was known to occur at different rates with each drug. They would see a sequence of instances that these drugs had been taken. In each case either the adverse reaction had occurred or nothing had happened.

Analogously, in the crime conditions subjects were asked to imagine that were on their way home late at night and that they could choose one of five different streets. They would then see a sequence of instances that people had chosen to walk in these streets at night. In each case either a specific crime had occurred or nothing had happened.

Afterwards they were to judge the probability of the adverse reaction or the crime for each medication or street. Subjects were instructed to read out loud every name and every outcome as they appeared in order to ensure full encoding of all information and that they were not to try to simply count the presented events. They were also informed that an audio recording would be produced in order to check the compliance with the encoding instructions. After a practice trial with a shortened sequence of 10 items that did not appear later in the experiment, participants completed six trials—three low range and three high range sequences—in random order. Subsequently, they proceeded with additional measures, provided demographic information, and were debriefed.

Subjects. A total of 104 students at a German university took part in this lab-based study, either for a payment of €3 (approximately US$3.75) or for course credit. Seventeen subjects had to be excluded because they did not read out loud the stimuli during encoding ($n = 6$) or because their reading could not be confirmed due to technical difficulties with the recording equipment ($n = 11$). Thus, the final sample included $N = 87$ subjects, 68 (78.2%) of whom were female, with *n*s for individual analyses ranging from 19 to 25. Mean age was 22.41 years ($SD = 2.31$) and mean grade in the Abitur (German general higher education entrance exam) was 2.19 ($SD = 0.55$).[2] Subjects were randomly assigned to one of the four between-subjects conditions.

Objective probabilities. The presented probabilities were 10%, 20%, 30%, 40%, and 50% in the low range and 50%, 60%, 70%, 80% and 90% in the high range. We chose this distribution for three reasons. First, equidistant spacing theoretically allows for perfect sensitivity on the rating scale. Second, presenting more stimuli than available categories creates the above described conflict for subjects using the rating-scale and might lead to apparent context effects, which can be easily tested for by comparing the common stimulus. Third, across contexts the stimuli cover nearly the whole continuum of probability which

[2] Grades on this exam vary between 1.0 and 4.0 with 1.0 being the best possible grade.

allows for more general conclusions than the previous study by Haase et al. (2013) and has implications for predictions according to range-frequency theory.

The range principle indicates that subjects match the subjective judgment range to the presented stimulus range. However, this matching is not absolute but the result of an inferential process and therefore malleable. It varies as a function of anchors, background, and familiarity with the stimulus domain (Parducci & Perrett, 1971; Sarris & Parducci, 1978; Wedell et al., 1990). It seems reasonable to assume that due to the three absolute anchors of the probability range the subjective range is fixed irrespective of context. Nonetheless, presenting the whole range within-subjects ensures this.

Context effects would then be expected to be caused by a skewed stimulus distribution through the frequency principle. This skewing can be achieved by presenting the same stimuli with different frequencies or by varying the spacing of the stimuli within a fixed range as we do in these experiments. The number of available judgment categories only affects context effects in the former case (Parducci & Wedell, 1986). Thus, if our design manipulated the context between-subjects, range-frequency theory would predict higher judgments for the low range common stimulus on both scale formats. For a within-subjects design the prediction would be to find severely reduced (Haubensak, 1992) or no context effects.

Stimulus material. Subjects encoded different probabilities as sets of option-outcome pairs with each outcome being the occurrence or non-occurrence of a focal event. The relative frequency of the focal event within one set conveyed the likelihood information. Sets consisted of 20 items each, which referred to 20 instances the respective option hat been chosen. For example, in order to present a 40% likelihood of experiencing dizziness after taking a dose of *Peter Pharma*, the name *Peter* coincided eight times with the word *Dizziness* and 12 times with the word *Nothing*.

Options. For six trials we needed 30 different options. We used 15 male and 15 female German first names adopted from Betsch, Glauer, Renkewitz, Winkler, and Sedlmeier (2010) followed either by the word *Pharma* or the word *Street*, for example, *Lisa Pharma*. Appendix A lists the names.

Outcome domain. The presented option-outcome pairs stood either for hypothetical medications and the likelihood of adverse reactions following their ingestion or for hypothetical streets and the likelihood of a crime occurring there. For a total of six trials we needed six different adverse reactions and six different crimes. For the former we chose six out of the eight low severity adverse drug reactions from Haase et al. (2013). For the latter we chose six crimes from Mannhaupt (1983). Appendix A lists all outcomes.

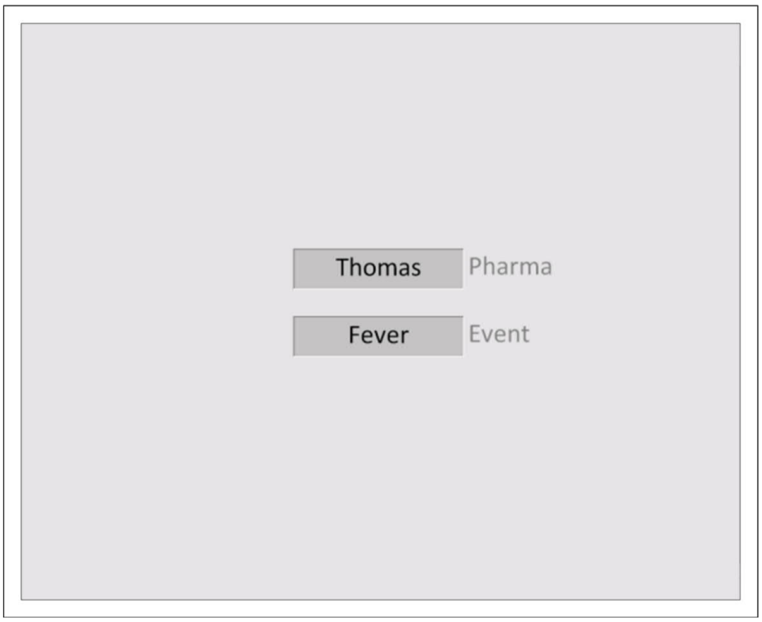

Figure 1. Presentation of stimuli in the encoding conditions without an anchor in Experiment 1 and 2. Original materials were in German.

Creation of stimuli. For each subject six sequences of stimuli were created at the beginning of the experimental session. The six different outcomes were distributed randomly over the six sequences of stimuli. Likewise, the 30 names were distributed randomly over the altogether 30 options with the constraints that all five names within one sequence were of the same gender and that neither small nor large probabilities were exclusively associated with female or male names, for example, small probabilities: two sets with female names, one with male names; large probabilities: two sets with male names, one with female names; and vice versa. The variant of these range-gender pairings was randomly selected for each subject.

Encoding of probabilities. Option-outcome pairs were presented as a fast sequence. This presentation format has been used in comparable research (Goldstein & Rothschild, 2014; Haase et al., 2013) and communicates probabilities in an easily understandable way. A large body of research indicates that the encoding of frequencies is a mostly automatic and accurate process (Hasher & Zacks, 1984; Zacks & Hasher, 2002). However, one goal of this experiment was to increase encoding error and thus add noise to the representations. Therefore, all five sets within one trial and hence five probabilities were presented within one sequence of items. The 100 items in one sequence were presented at random. The name of the option and the outcome were presented in two centered grey boxes over a light grey background. The boxes measured 20.7 mm high by 86.5 mm wide and were labeled *Pharma* or *Street* and *Event* in a medium grey. The font used for the labels as well as for the items was

36 point Calibri (Figure 1). The option-outcome pairs flashed in 1300 ms intervals with the stimulus remaining visible for 1000 ms and a 300 ms inter-stimulus interval.

To ensure full encoding of all stimuli and to prevent simple counting of the focal event, subjects were instructed to read out loud the first name (omitting the word *Pharma*/*Street*) and the associated outcome upon appearance, for example, *"Maria, Nothing; Laura, Nothing, Maria, Headache; Anna, Headache; Nina, Nothing; etc."* We produced audio recordings of each subject to check the compliance with this instruction.

Subjective probability scale format. After encoding one sequence and thus five sets of option-outcome pairs, subjects judged the probability of the outcome occurring for each option in random order, for example, "What is the probability of experiencing the adverse event dizziness with Peter Pharma?" We manipulated the scale format used for subjective probability judgments between subjects.

7-point rating scale. The 7-point rating scale offered verbal labels for each category. Those were in ascending order: *almost zero*, *very small*, *small*, *moderate*, *large*, *very large*, and *almost certain*.

Percent estimate. In the percent estimate condition subjects were asked to fill in the statement "The event will occur with a probability of ___%."

Perceived severity of outcomes. Following the six trials of probability judgments subjects judged the severity of each presented outcome on a visual analog scale (scroll bar) with the verbal anchors not *severe* (score = 0) and *very severe* (= 100). The scale was anchored at the midpoint (= 50) and no numeric feedback was provided. The severity ratings served as a check for the domain manipulation and as control variable for the probability judgments.

Subjective numeracy. Subjects also completed the Subjective Numeracy Scale that measures self-assessed ability to perform mathematical operations with proportions and preferences for numerical or verbal likelihood information (Fagerlin et al., 2007; German translation by Keller, Siegrist, & Visschers, 2009). It consists of eight six-point rating scale items, for example, "How good are you at working with percentages?" Mean subjective numeracy was 4.20 (SD = 0.77) and did not differ across conditions, Fs ≤ 1.98, ps ≥ .163. Since we found no relation between subjective numeracy and any of the dependent variables, we omit this variable from all further analyses.

Data analysis. We wanted to ensure that all judgments were based on identical information and thus excluded all subjects who did not read out loud the stimuli during encoding or whose reading could not be confirmed due to technical problems. Next we checked for outliers at the trial level. We excluded a trial if the subject's five respective judgments were on average more than three median absolute deviations (*MAD*) above the respective median. The *MAD* is a robust measure of dispersion, not impacted itself by outliers and sample size and has been recommended for detecting outliers. The criterion of three *MAD*s is considered conservative (Leys, Ley, Klein, Bernard, & Licata, 2013).

Subjects completed three trials under identical conditions and we aggregated at the individual level first and then performed parametric testing with these individual mean scores. We assessed scale sensitivity as the Pearson correlation between objective and subjective probabilities at the level of the individual as well as based on aggregate scores. Note that correlation coefficients based on means can differ drastically from those based on raw scores (Nickerson, 1995). Thus, to compute individual sensitivity we calculated the correlation for each trial and averaged these three coefficients to form the individual score. Correspondingly, to calculate the aggregate sensitivity we used all judgments, that is, three per subject, rather than the individual mean judgments. We applied Fisher's r-to-z transformation[3] to average correlations, to calculate means and confidence intervals, and to test for differences (Gorsuch & Lehmann, 2010). In the rare instances of a perfect correlation we recoded the sensitivity score to 0.9999. We report mean sensitivity scores transformed back to r.[4]

We reasoned that the high resolution of the percent format might lead to greater variability and thus more noise in the judgments resulting in reduced sensitivity. We tested this assumption with the coefficient of variation (*CV*). The *CV* is the standard deviation normalized by the mean and thus represents a dimensionless measure of variability that allows for the comparison between different scale formats. *CV*s were also calculated using all judgments rather than individual means.

Finally, we compared the two respective judgments of the common stimulus with dependent t-tests based on the individual means.

Results

There were no outliers in this experiment. Figure 2 presents mean subjective probability judgments as a function of presented probabilities for both scale formats, both

[3] $z = 0.5\log_e\left[\frac{(1+r)}{(1-r)}\right]$.

[4] $r = \frac{(e^{2z}-1)}{(e^{2z}+1)}$.

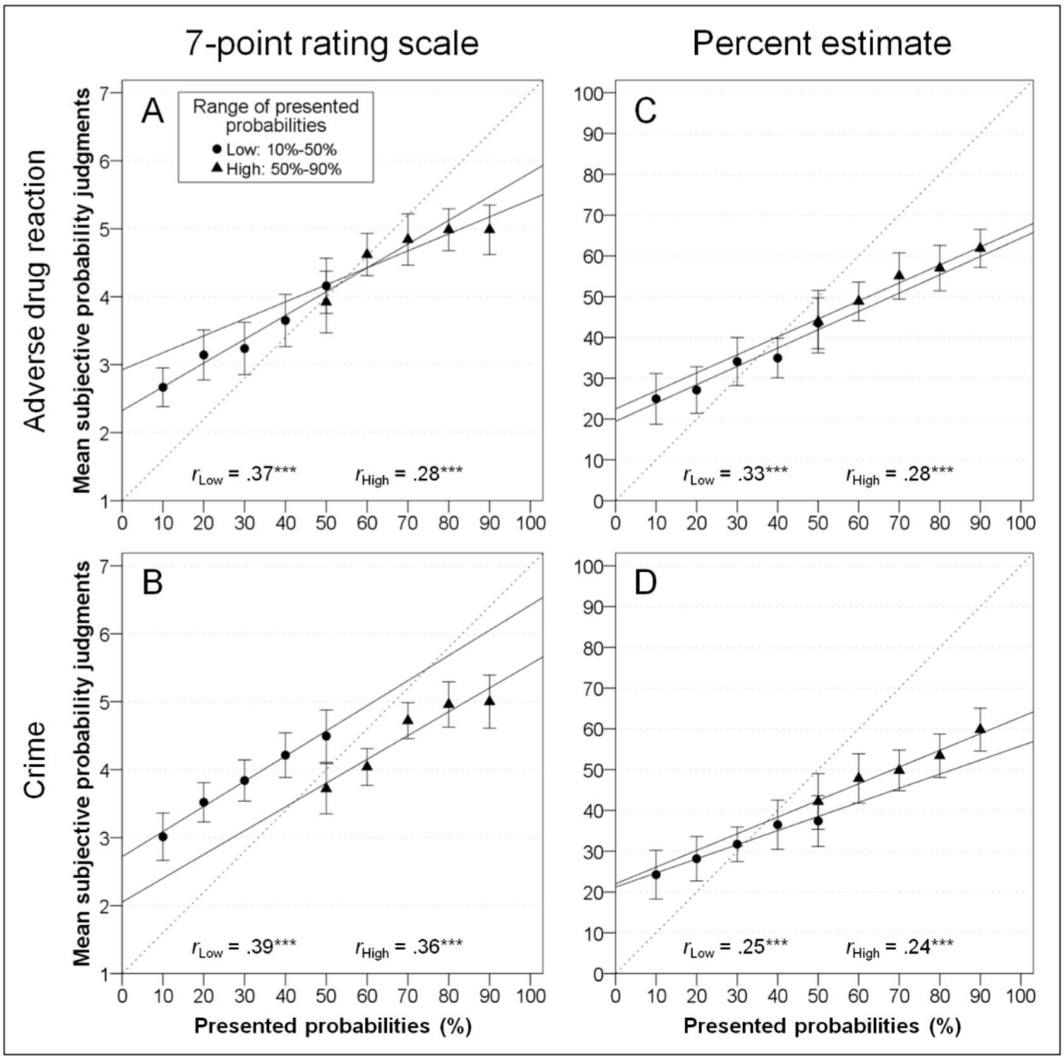

Figure 2. Mean subjective probability judgments as a function of presented probabilities (A: $n = 21$, B: $n = 25$, C: $n = 19$, D: $n = 22$). Solid lines represent the best linear fit. Correlation coefficients indicate scale sensitivity at the aggregate score level. Error bars = 95% within-subjects CIs.[5] $***p < .001$.

judgment domains, and both probability ranges. Visual inspection reveals strong regression toward the mean in all conditions (see also the analysis of scale sensitivity below). Note that for the rating scale the dashed identity line serves only to illustrate the relation between the category range and the full probability continuum. For slopes of 1, two even steeper identity lines, one for each judgment context, would be necessary.

---

[5] To calculate within-subjects CIs we applied the method suggested by Cousineau (2005) with the corrections suggested by Morey (2008) and Baguley (2012). CIs are based on participant-mean centered scores and calibrated so that nonoverlapping CIs correspond to a confidence of the difference between the two related means that does not include zero.

Perceived severity of outcomes. Severity ratings were sufficiently consistent within domain conditions and thus averaged to form one severity score, adverse drug reaction Cronbach's alpha = .78; crime = .84. We conducted a 2 (scale format: 7-point rating scale vs. percent estimate) × 2 (domain: adverse drug reaction vs. crime) ANOVA with perceived severity as dependent variable as a check of the domain manipulation. The outcomes in the crime condition were judged to be more severe, $M = 75.40$, $SD = 14.52$, than the adverse reactions, $M = 62.83$, $SD = 18.15$; $F(1, 83) = 13.91$, $p < .001$, $\eta_p^2 = .14$. However, we also found a small main effect of the scale format on the severity ratings indicating that subjects in the percent conditions perceived the outcomes to be slightly more severe, $M = 73.50$, $SD = 16.30$, than those who provided judgments on the rating scale, $M = 66.16$, $SD = 17.74$; $F(1, 83) = 4.34$, $p = .040$, $\eta_p^2 = .05$. The interaction was not significant, $F < 1$.

One goal of this experiment was to establish that in our paradigm subjects used the two scale formats solely as measures of beliefs in objective probabilities, that is, independently of the respective outcomes. As the main effect of the scale format on perceived severity, though small, might suggest a relation, we calculated correlations between perceived severity and probability judgments. We found no discernible pattern and, out of altogether 40 pairings (10 judgments per condition and four conditions), only three significant (one-sided) positive correlations between small range percent estimates and perceived severity of drug adverse reactions: $r_{10\%} = .56$, $p = .006$; $r_{30\%} = .47$, $p = .021$; $r_{50\%} = .45$, $p = .027$.

Subjective probability as a function of the content domain. If the subjects used the two scale formats only to express beliefs in objective probabilities, then their judgments should not vary as a function of the judgment domain. In order to test this assumption, we conducted two mixed design ANOVAs—one for each scale format—with the domain as between-subjects factor and the ten objective probabilities as within-subjects factor. In both cases Mauchly's test indicated a violation of the assumption of sphericity, percent: $\chi^2(44) = 145.60$, $p < .001$; 7-point: $\chi^2(44) = 74.04$, $p = .003$. Therefore, we report Greenhouse-Geisser corrected tests, percent: $\varepsilon = .40$; 7-point: $\varepsilon = .74$.

For the percent estimates we found a significant effect of the different probabilities indicating, of course, only that higher presented probabilities were judged to be higher, $F(3.62, 141.25) = 38.43$, $p < .001$, $\eta_p^2 = .50$. Crucially, there was neither a main effect of the domain on the subjective probability judgments nor an interaction, Fs < 1.

For the 7-point rating scale we found the same effect of presented probabilities on subjective probability judgments, $F(6.65, 292.50) = 37.00$, $p < .001$, $\eta_p^2 = .46$. However, we also found a small, though not significant, main effect of the domain, $F(1, 44) = 2.29$, $p = .137$, $\eta^2_p = .05$, and a small significant interaction effect, $F(6.65, 292.50) = 2.49$, $p = .019$, $\eta^2_p = .05$. In order to understand these mixed results we calculated independent *t*-tests for each level of objective probability.[6] Two probabilities in the low range were judged to be significantly higher and one probability in the high range was judged to be significantly lower when they referred to crimes as compared to drug adverse reactions, $t(44)_{30\%} = 2.64$, $p = .011$, $r = .37$; $t(44)_{40\%} = 2.27$, $p = .028$, $r = .32$; $t(44)_{60\%} = -2.70$, $p = .010$, $r = .38$. Thus, there was generally no systematic effect of the judgment domain on subjective probability ratings.

Sensitivity. The sensitivity of a scale quantifies the extent to which subjective probability ratings monotonically follow the objective probabilities. For the analysis at the individual score level we recoded perfect sensitivity in one out of 505 (0.20%) cases.

Figure 3 presents mean sensitivity. All scores (range = .42–.58) were markedly lower than in the study by Haase et al. (2013) reflecting the higher error in encoding conditions. A mixed-design ANOVA with sensitivity as dependent variable indicated no main effects of

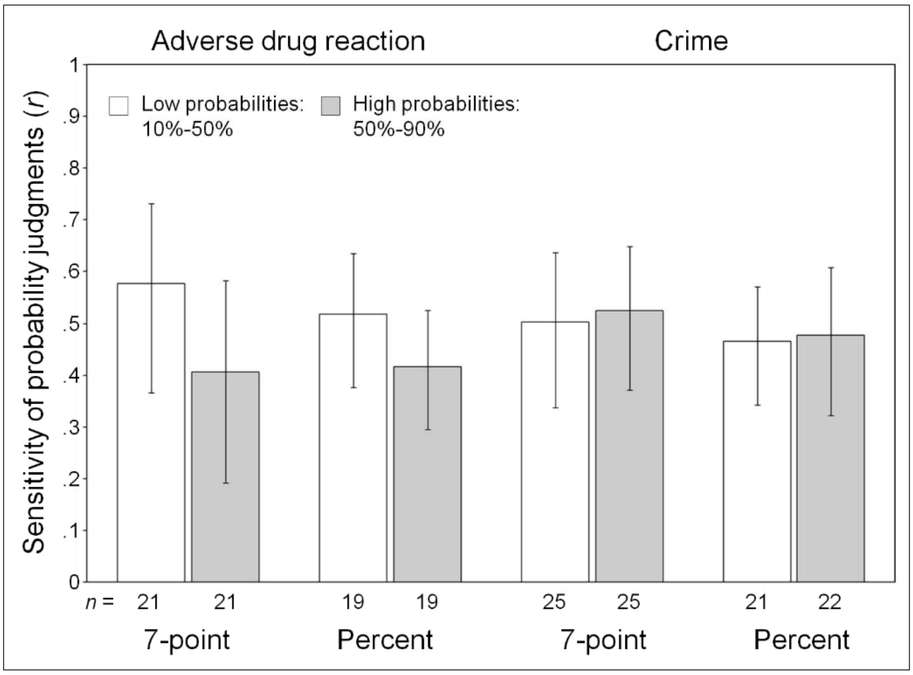

Figure 3. Sensitivity at the individual score level of both scale formats for both ranges of probabilities and in both judgment domains. Error bars = 95% CIs. CIs were calculated by transforming Fisher's *z*-values back to *r* and are therefore asymmetrical.

[6] As we are essentially testing for the null hypothesis we refrain from a Bonferroni correction of significance levels.

content domain, scale format, or probability range, all Fs ≤ 1.09, all ps ≥ .301. The apparent interaction between range and content domain indicating an effect of the probability range in the adverse drug reaction condition but not the crime condition was not significant, $F(1, 82) = 2.97$, $p = .089$, $\eta^2_p = .04$, nor were the other interaction terms, Fs < 1.

At the aggregate level, scale sensitivity was lower than at the individual level (range = .24–.39, Figure 2). To test for differences between conditions we employed a test of heterogeneity described by Fleiss (1993), using SPSS syntax provided by Weaver and Wuensch (2013). When the null hypothesis is true, that is, when all correlations are equivalent, the test statistic Q has an approximate chi-square distribution with $df = k - 1$ where k denotes the number of independent correlations.[7] Although correlation coefficients were generally higher for the 7-point rating scale, the differences in aggregate sensitivity as a function of the scale format or of the domain were not significant in either probability range, low: $Q = 5.08$, $df = 3$, $p = .166$; high: $Q = 3.16$, $df = 3$, $p = .367$.

For the within-subjects comparisons between probability ranges we employed a t-test developed by Williams (1959) for two dependent correlations with one common variable, that is, r_{12} vs. r_{13}.[8] The test statistic follows approximately a t-distribution with $df = n - 3$.[9] Aggregate sensitivity did not differ between probability ranges in any of the between-subjects conditions, all ts ≤ |1.34|, all ps ≥ .180.

[7] $Q = \sum_{i=1}^{k} W_i(z_i - \bar{z})^2$ where k denotes the number of independent correlations, z_i is the Fisher's r-to-z transformed value of the ith correlation, W_i is the reciprocal of its variance, that is, $n_i - 3$, and $\bar{z}$ is a weighted average of the k correlations: $\bar{z} = \frac{\sum W_i z_i}{\sum W_i}$.

[8] Strictly speaking, our test has the structure r_{12} vs. r_{34} because we are comparing the correlation between low presented probabilities and low range probability judgments with that between high presented probabilities and high range judgments. However, as the two presented ranges, that is, 10%–50% in 10% increments and 50%–90% in 10% increments, are equivalent in this context we can use the t-test by Williams rather than a less parsimonious test such as the *ZPF* test statistic by Steiger (1980) for two dependent nonoverlapping correlations.

[9] $$t(n-3) = (r_{12} - r_{13})\sqrt{\frac{(n-1)(1+r_{23})}{2\left(\frac{n-1}{n-3}\right)|R| + \frac{(r_{12}+r_{13})^2}{4}(1-r_{23})^3}}$$

where $|R| = 1 - r_{12}^2 - r_{13}^2 - r_{23}^2 + 2r_{12}r_{13}r_{23}$.

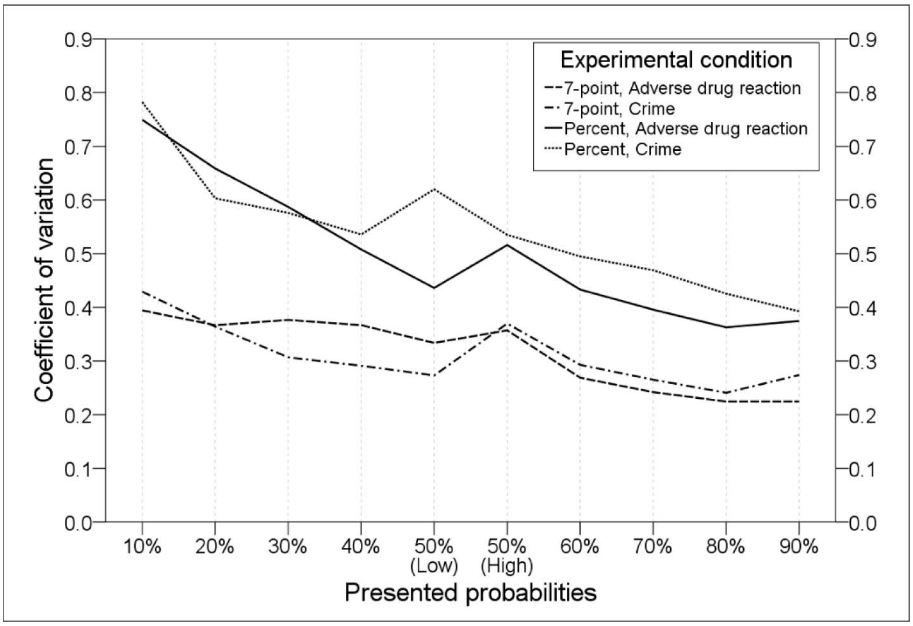

Figure 4. Coefficients of variation as variability profiles over all probability judgments for all between-subjects conditions.

Error in judgments. We expected percent estimates to show a higher variation relative to the judgments on the 7-point rating scale. To compare the relative variation between scale formats we calculated the coefficient of variation (*CV*) for every judgment in every condition.[10] Figure 4 presents *CV*s for each condition in the form of variability profiles, that is, an array of *CV*s from the same sample is depicted as a line over a meaningful order of measurements, in this case ascending presented probabilities.

To our knowledge there exists neither a multivariate test to compare variability profiles with each other (taking the dependence of *CV*s into account) nor to test multiple *CV*s against each other in one procedure. As a proxy we calculated the mean *CV* for each between-subjects condition across all judgments and compared these by conducting single comparisons using the *t*-test for independent *CV*s[11] by Sokal and Braumann (1980). Mean *CV*s for

---

[10] We applied the following correction to the calculation of the *CV* as suggested by Sokal and Braumann (1980): $\left(1 + \frac{1}{4n}\right)\frac{s}{\bar{x}}$. This correction only makes an appreciable difference when samples are very small and affected our data only in the third decimal. We applied it nonetheless because the tests we calculated to compare *CV*s were designed and evaluated for thusly corrected *CV*s.

[11] $t = \frac{CV_1 - CV_2}{\sqrt{s^2_{CV_1} + s^2_{CV_2}}}$ with $df = n_1 + n_2 - 2$ where $s_{CV} = \sqrt{\frac{CV^2}{2n}\left(\frac{n}{n-1} + 2CV^2\right)\left(1 + \frac{1}{4n}\right)^2}$.

judgments on the 7-point rating scale were: adverse drug reaction = 0.315; crime = 0.311; and for percent estimates: adverse drug reaction = 0.502; crime = 0.543. Judgments did not differ in variation between content domains on either scale format, *t*s < |1|. Comparisons between scale formats within and across domains, and assuming a Bonferroni corrected significance criterion of .008, on the other hand, revealed in line with our hypothesis a significantly higher variation in percent estimates as compared to judgments on the rating scale, *t*s ≥ |2.83|. *p*s ≤ .004, *r*s between .25 and .29.

Common stimulus. Subjects encoded an objective probability of 50% in both probability ranges. Judgments of this common stimulus allow for a direct test of context effects. We expected the common stimulus to receive higher judgments on the rating scale in the low-range context than in the high-range context but percent estimates to be consistent across contexts. As the points and CIs in Figure 2 overlap, Figure 5 presents only the judgments of the common stimulus separate for both contexts.

Percent estimates were consistent across ranges, *t*s < 1, *p*s ≥ .378. For the 7-point rating scale we found the expected relation between scores only in the crime domain, $t(24) = 2.65$, $p = .014$, $r = .48$. Ratings of adverse reactions showed the expected trend but did not differ significantly, $t < 1$, $p = .494$.

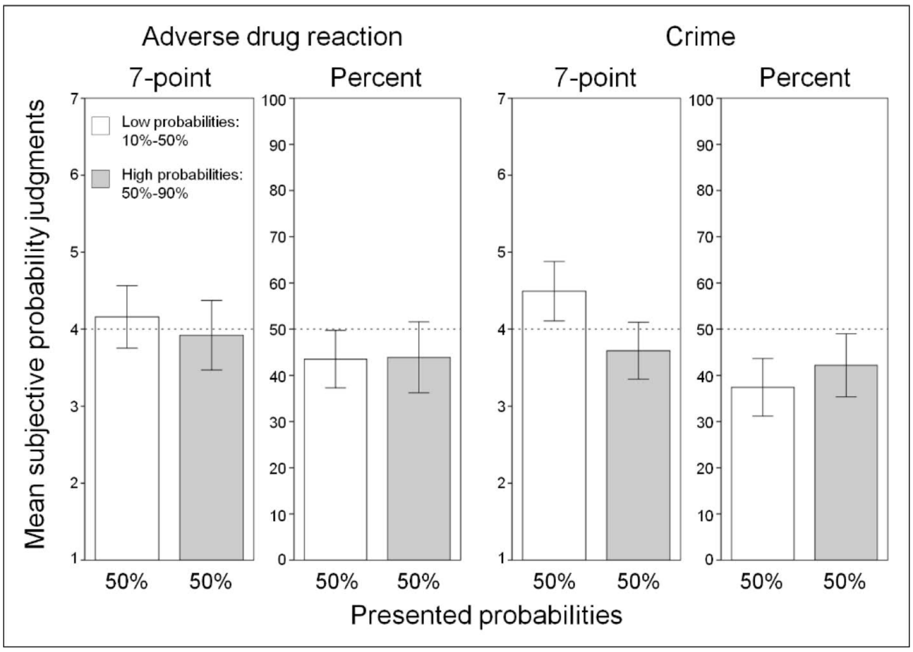

Figure 5. Subjective probability judgments of 50% probabilities in their respective range context for each between-subjects condition. The dotted line represents the scale's respective midpoint. Error bars = 95% within-subjects CIs.

Discussion

We observed no systematic relation between subjective probability judgments and the content domain or the perceived severity of outcomes. Thus, we conclude that subjects used both scale formats to express beliefs in objective probabilities (i.e., to estimate rather than to evaluate) and that differences between formats are caused directly by overt characteristics of the scales rather than by the elicitation of different concepts.

In line with our expectations, scale sensitivity was generally reduced and did not differ between formats at the individual or the aggregate score level. For the latter we observed generally lower scores for the percent format but differences were not significant. Given our relatively small sample size, however, the test for differences (an extension of the common z-test) was likely underpowered (Kenny, 1987). The high error in the sampling of event frequencies had a stronger effect on percent estimates than on category ratings as indicated by significantly higher relative variation in the former as compared to the latter. However, beyond these scale-specific factors, sensitivity was generally reduced by an inability to clearly differentiate five levels of probability with either format (e.g., overlapping CIs in Figure 2).

Percent estimates were consistent across judgment contexts, that is, irrespective of the stimulus distribution, objective probabilities were mapped in the same way onto the judgment categories (i.e., integers). Results regarding the rating scale, however, were mixed. While subjects in the adverse drug reactions condition produced two similar judgment functions, subjects in the crime condition clearly used the rating scale incoherently and judged the common stimulus differently across contexts. It is unlikely that this discrepancy is a systematic effect of the content domain as judgments were otherwise not systematically affected by it. Moreover, if subjects did not try to express five distinct probabilities, users of the rating scale were not faced with the above described conflict between probability mapping and stimulus discrimination anyway.

We believe that this discrepancy is the result of the generally large error in representations and reveals a certain amount of arbitrariness in the positioning of judgments along the category range. Keep in mind that subjects encoded all five probabilities and thus a representation of the context, that is, the general range of stimuli, before providing their judgments. Additionally, note that judgments on the rating scale seem to be loosely anchored around the scale midpoint whereas the linear judgment functions of the percent estimates cross the identity line at around 30%. The latter might be explained by the mental representation of integers, which we return to in the general discussion. The former, however, might indicate that subjects did not place their judgments in relation to the scale's endpoints, that is, to a coherent mapping of the probability continuum on to the category range, but rather in relation to the scale's midpoint (Marsh & Parducci, 1978; Marsh, 1983). However, unreliable representations of context, that is, of the range of presented probabilities in relation to 50%, together with the inherent vagueness of the rating scale's categories allowed for some fluctuation around the midpoint resulting in apparent context effects in one condition but not in the other. Thus, in Experiment 2 we investigated whether an anchor in the sample of events can fasten category ratings to the midpoint.

Experiment 2

The probability continuum has two clearly defined bounds at 0 and 100%. Assuming sampling error, subjective probability estimates might deviate from objective probabilities in the positive as well as the negative direction. However, the more extreme an objective probability is, the less symmetric this deviation will be. For instance, the estimate of an objective probability of 10% can only deviate by 10% into the negative but by 90% into the positive. Thus, even if we assume that the sampling error is symmetric (as most error-models do) these bounds will lead to asymmetric error near the endpoints of the probability continuum and increase regression toward the mean. Likewise, if we introduce an anchor indicating that all probabilities in a given trial are below or above a certain reference point, we would expect to constrain the possible range of estimates and to observe judgments that are regressed in relation to this bound (Hollands & Dyre, 2000).

In Experiment 1 we observed that judgments on the rating scale appeared to be anchored roughly around the midpoint. We assumed that subjects used the midpoint as a reference to position their five respective judgments within one context and that due to imprecise representations of said context these five judgments fluctuated around this reference point. Thus, in this experiment we introduced an anchor into the encoding of the likelihood information indicating whether the presented probabilities were equal to and above or equal to and below 50%. We expected percent estimates to be regressed within context, that is, with relation to this new bound resulting in a step-like pattern of judgments across contexts with the 50% in the low range to be estimated as lower than the same probability in the high range. If judgments on the rating scale are anchored at the midpoint, however, then the common stimulus of 50% should receive the same rating across contexts. In addition to probability judgments we collected judgments of the absolute frequency of the target events. These judgments served as a standard of comparison for judgments of the common stimulus. We expected to find the same pattern of within-context regression as for the percent estimates irrespective of the used scale format. Besides this manipulation we aimed to make the paradigm as similar to Experiment 1 as possible and our expectations regarding scale sensitivity were unchanged.

Method

The experiment implemented a 2 (scale format: 7-point rating scale vs. percent estimate) × 2 (encoding mode: anchor vs. no anchor) between-subjects design with one within-subjects factor (low range of probabilities vs. high range).

Procedure. The procedure was identical to Experiment 1 with the exception that after encoding the likelihood information subjects first judged the absolute frequency of all target events in random order and immediately afterwards their respective probabilities. Also, we collected different additional measures as explained below.

Subjects. A total of 116 students at a German university took part in this lab-based study, either for a payment of €3 (approximately US$3.75) or for course credit. Eleven subjects had to be excluded because they did not read out loud the stimuli during encoding ($n = 7$) or because their reading could not be confirmed due to technical difficulties with the recording equipment ($n = 4$). Thus, the final sample included $N = 105$ subjects, 82 (78.1%) of whom were female, with *n*s for individual analyses ranging from 25 to 28. Mean age was 22.55 years ($SD = 3.15$) and mean grade in the Abitur exam was 2.09 ($SD = 0.47$). Subjects were randomly assigned to one of the four between-subjects conditions.

Encoding of probabilities. The event frequencies in Experiment 1 were presented in what could be called a purely sequential manner, that is, each option-outcome pair appeared for 1000 ms and disappeared afterwards. In order to provide an anchor, we introduced a minimum of aggregate information while otherwise keeping the presentation identical. Thus, the no anchor condition was identical to Experiment 1 while in the anchor condition the items of a sequence were listed row-wise on the screen and remained visible until the end of the sequence. Each item was presented in a frame with two labeled boxes as in the sequential encoding condition. The frame measured 11.3 mm by 59 mm and each box 5.3 mm by 37.7 mm (Figure 6). The items were presented at the same speed as in the no anchor condition with a new frame appearing every 1300 ms. Thus, the encoding process was virtually identical in both conditions with the exception that the anchored presentation allowed for extracting one additional piece of information, namely whether the target event (e.g., Headache) appeared in the majority or minority of cases.

Figure 6. Presentation of stimuli in the encoding conditions with an anchor (and high error) in Experiment 2 and 3. Original materials were in German.

Frequency judgments. Subjects judged the absolute frequency of the adverse reaction for each medication in random order, for example, "How often did the adverse event dizziness occur with Peter Pharma?" Participants filled in the statement "The adverse event occurred ___ times."

Verbal-numeric scale mapping. After the six trials subjects were asked to assign numeric percent values to the verbal labels of the 7-point rating scale. The labels were presented in random order and participants could either assign a point value or an interval of values along the percent array of 0–100, for example, "The term 'very small' corresponds to a probability of ___% to ___%." For point values subjects only filled in the left blank. We address the results of this task after reporting the experiments.

Numeracy. In Experiment 1 we observed no relation between subjective numeracy and any of the dependent variables. As subjective and objective numeracy appear to be related but not identical (Peters & Bjalkebring, 2014), in this experiment subjects completed a combination of the 3-item numeracy scale by Schwartz, Woloshin, Black, and Welch (1997) and the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). The altogether seven items are comprised of actual mathematical quizzes, for example, "Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?" The sum score of all correct answers (possible range: 0–7) constitutes the numeracy score.

Mean numeracy was 3.40 (SD = 1.27) and did not differ across conditions, Fs < 1. Again, we found no relation between numeracy and any of the dependent variables and thus omit this variable from all further analyses.

Results

There were no outliers in the probability judgments. Figure 7 presents mean subjective probability judgments. Visual inspection reveals generally similar judgment functions to those in Experiment 1. Judgments on the rating scale were consistent across contexts and anchored at the scale's midpoint whereas percent estimates only followed the same function when no anchor was provided. In the anchor condition estimates were regressed within context resulting in two different judgment functions.

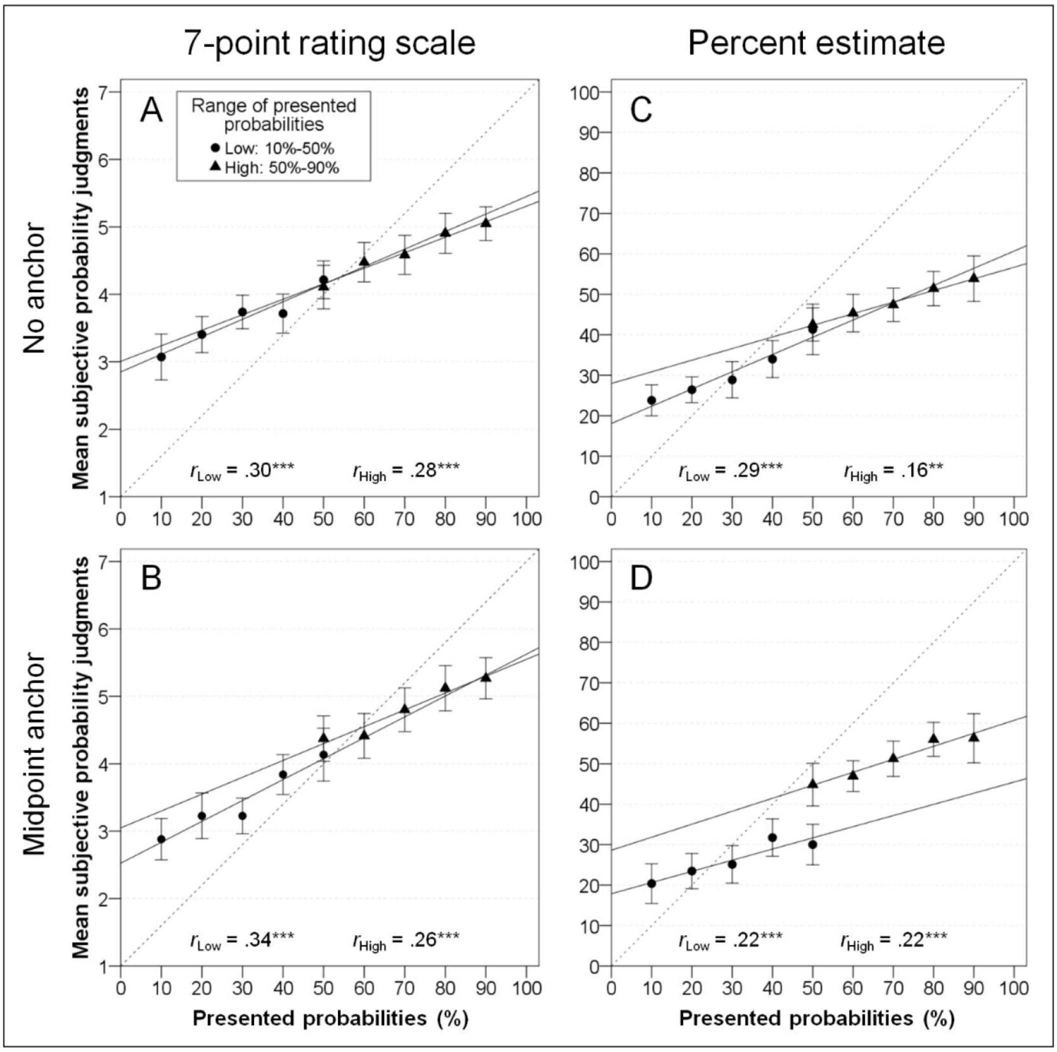

Figure 7. Mean subjective probability judgments as a function of presented probabilities (A: n = 28, B: n = 25, C: n = 27, D: n = 25). Solid lines represent the best linear fit. Correlation coefficients indicate scale sensitivity at the aggregate score level. Error bars = 95% within-subjects CIs. **p < .005. ***p < .001.

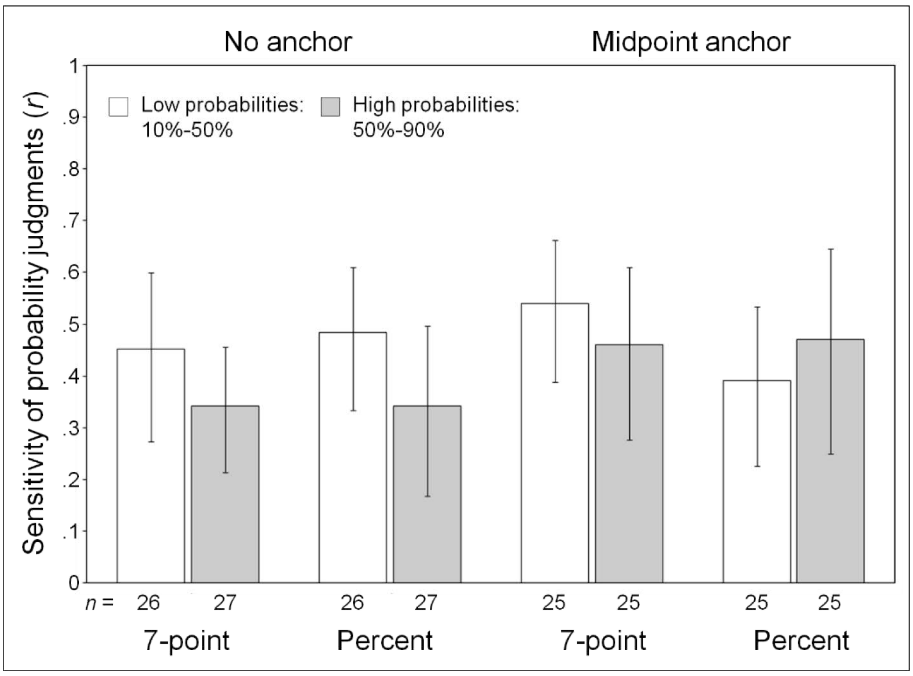

Figure 8. Sensitivity at the individual score level of both scale formats for both ranges of probabilities and both encoding conditions. Error bars = 95% CIs.

Sensitivity. We recoded perfect sensitivity in one out of 600 (0.17%) correlations. Figure 8 presents mean sensitivity. The scores (range = .34–.54) were comparable to Experiment 1. A mixed-design ANOVA indicated no effects of scale format, encoding mode, or probability range, all Fs ≤ 1.90, all ps $\geq .172$.

Likewise at the aggregate score level, sensitivity did not differ as a function of encoding mode or scale format for either range of probabilities, low: $Q = 3.68$, $df = 3$, $p = .299$; high: $Q = 2.85$, $df = 3$, $p = .415$, but note that aggregate sensitivity was again consistently lower for percent estimates when compared to the rating scale. Comparisons between ranges revealed that percent estimates in the no anchor condition were slightly less sensitive to high as compared to low probabilities, $t(402) = 2.37$, $p = .018$, $r = .12$. In all other conditions sensitivity did not vary as a function of the probability range, all ts ≤ 1.25, all ps $\geq .212$.

Error in judgments. The pattern of variation was consistent with Experiment 1 (Figure 9). Mean CVs for judgments on the rating scale were: no anchor = .298; anchor = .316; and for percent estimates: no anchor = .584; anchor = .560. CVs did not differ within scale format, that is, between the conditions with and without an anchor, both ts < 1, but all four differences between scale formats (within and across anchor conditions) were significant at a Bonferroni corrected criterion of .008, all ts $\geq |3.74|$, all ps < .001, all rs $\geq .29$.

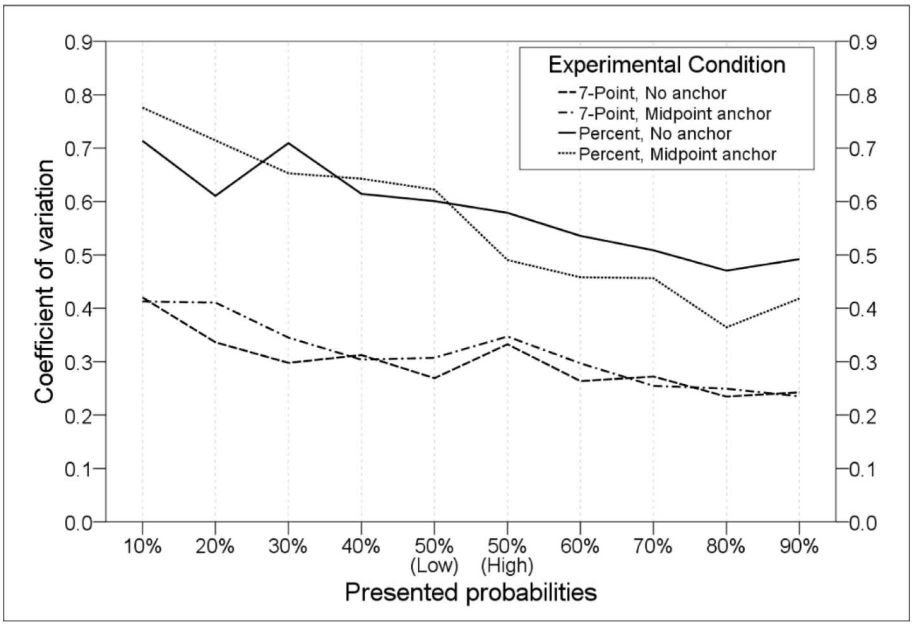

Figure 9. Coefficients of variation as variability profiles over all probability judgments for all between-subjects conditions.

Common stimulus. Probability judgments of the common stimulus are shown in the upper half of Figure 10. In line with our expectations, an anchor led to percent estimates that were significantly smaller in the low probability range than in the high range, $t(24) = -3.30$, $p = .003$, $r = .54$ while estimates were consistent across contexts when no anchor had been provided, $t < |1|$. Judgments on the 7-point rating scale, on the other hand did not differ across contexts irrespective of an anchor, ts $< |1|$. Additionally, none of the ratings differed significantly from the scale's midpoint value of 4, ts < 1.99, ps $\geq .058$.

Frequency judgments. We excluded 22 out of the total of 630 trials (3.49%) as outliers and analyzed frequency judgments analogously to the probability judgments. Appendix B presents scatter plots and the detailed analyses regarding the sensitivity of and relative error in frequency judgments. To sum up the findings, subjects in the anchor condition who used the percent format for probability judgments judged frequencies to be slightly higher than subjects using the rating scale. Individual sensitivity (range = .31–.49) was slightly, though not significantly, lower for higher frequencies and did not differ as a function of the scale format or encoding mode. Aggregate score sensitivity (range = .16–.25) was similar to percent estimates and lower than for judgments on the rating scale but also did not differ between conditions or the stimulus ranges. More noteworthy, though, *CV*s (range = 0.545–0.636) were very close to those of the percent estimates and did not differ irrespective of condition.

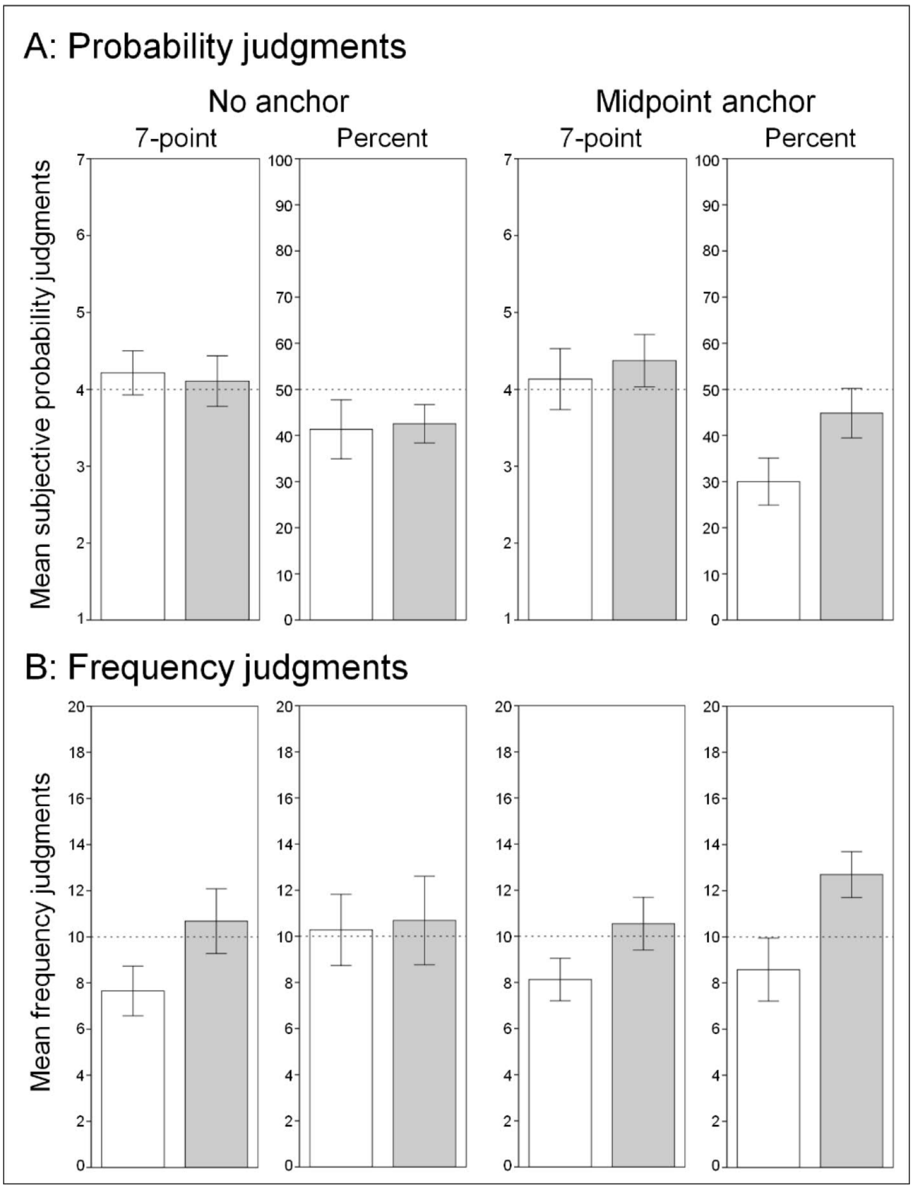

Figure 10. Judgments of the common stimulus. White bars = low range, grey bars = high range. Error bars = 95% within-subjects CIs. A: Presented probability = 50%. The dotted line represents the scale's respective midpoint. B: Presented frequency = 10 (dotted line).

The lower half of Figure 10 presents frequency judgments of the common stimulus. When an anchor had been provided, frequency estimates of the common stimulus were significantly lower in the low frequency range than in the high range, irrespective of the format used for probability estimates, 7-Point: $t(24) = -3.18$, $p = .004$, $r = .54$; percent: $t(24) = -4.41$, $p < .001$, $r = .67$. Without an anchor, frequency estimates were consistent across ranges when subjects provided percent estimates, $t < |1|$, but showed the same within-context regression when subjects had used the rating scale, $t(27) = -2.90$, $p = .007$, $r = .49$.

Discussion

The results regarding probability judgments were in line with our expectations. Scale sensitivity at the individual and the aggregate score level did not differ between scale formats. *CV*s showed that, again, the error in representations had a stronger effect on the percent format than on the rating scale. Crucially, none of these variables differed between encoding conditions, which indicates that the presentation modes were equivalent in terms of error.

Introducing an anchor, by retaining a minimum of aggregate information in the event sample, had the expected effects on both formats. Percent estimates were regressed relative to this bound which led to two different judgment functions across contexts and significantly different estimates of the common stimulus. Judgments on the rating scale, on the other hand, were anchored to the scale's midpoint resulting in consistent ratings of the common stimulus across contexts. As the probability judgments are based on representations of frequency, a comparison of the former with estimates of the latter will help to interpret these findings.

In the anchor condition frequency judgments were regressed within-context irrespective of the format used for probability judgments. The resulting two different judgment functions (Figure C1) as well as the differences between estimates of the common stimulus (Figure 10) resembled closely the pattern of corresponding percent estimates. Judgments on the rating scale, on the other hand, did not follow frequency estimates.

For the condition without an anchor we had expected consistent judgment functions for frequency estimates as well as for probability judgments (possibly with some of the previously observed fluctuations on the rating scale). In the percent condition we found the predicted distributions and a close match between frequency and percent estimates. Subjects who used the rating scale, however, produced frequency estimates that were also strongly regressed within-context. While this pattern, again, was not reflected in the corresponding probability ratings, it is nonetheless surprising and hard to explain.

Despite this discrepancy, we believe that these results indicate that a rating scale may deliver results that appear to be normatively consistent but fail to map the actual imprecision in the representations that underlie subjective probability judgments. The error-related dependent variables lend further support to this interpretation. The aggregate sensitivity scores as well as the *CV*s of all frequency estimates were very similar to those found for the percent estimates but differed markedly from those of the rating scale. Thus, in Experiment 3 we manipulated the error in encoding to investigate this further.

Experiment 3

In the previous two experiments subjects sampled event frequencies under extremely error-prone conditions and were generally not able to differentiate the five respective probabilities presented within one context. As a consequence, we were not able to observe the conflict between stimulus discrimination and probability mapping on the rating scale as described in the introduction. Instead we found that subjects anchored their ratings to the scale's midpoint, which led to judgment distributions that were inconsistent with corresponding frequency estimates. This led us to conclude that the rating scale does not reflect the error in the underlying representations which, in contrast, was clearly captured by the percent format.

Thus, in this experiment we manipulated the error in encoding while retaining the anchor in the stimulus material. We expected that more precise representations would increase the sensitivity of both scale formats and possibly render the percent format more sensitive. Depending on the actual amount of reduction in error we expected percent estimates to still be regressed within-context but to a lesser degree resulting in judgments of the common stimulus to be closer to each other. Regarding the rating scale, we were interested to find out whether less error would result in two distinct judgment functions based on discriminating the stimuli within-context or whether judgments would still be anchored to the midpoint.

Method

The experiment implemented a 2 (scale format: 7-point rating scale vs. percent estimate) × 2 (encoding mode: high error vs. low error) between-subjects design with one within-subjects factor (low range of probabilities vs. high range).

Procedure. The procedure was identical to Experiment 2 with the exception that we employed a different numeracy measure.

Subjects. A total of 105 students at a German university took part in this lab-based study, either for a payment of €3 (approximately US$3.75) or for course credit. Thirteen subjects had to be excluded: Nine did not read out loud the stimuli during encoding, for three subjects the reading could not be confirmed due to technical difficulties with the recording equipment, and one subject counted. Thus, the final sample included $N = 92$ subjects, 71 (77.2%) of whom were female, with *n*s for individual analyses ranging from 21 to 25. Mean age was 21.64 years ($SD = 2.75$) and mean grade in the Abitur exam was 2.18 ($SD = 0.55$). Subjects were randomly assigned to one of the four between-subjects conditions.

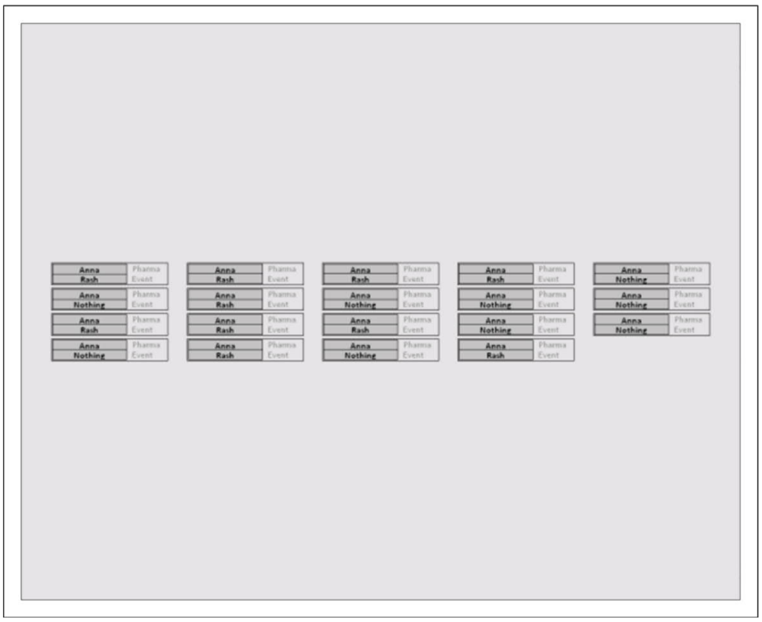

Figure 11. Presentation of one probability in the low-error encoding condition (with an anchor). Original materials were in German.

Encoding of probabilities. The high-error encoding condition was identical to the anchor condition in Experiment 2. For the low-error condition we wanted to facilitate the discrimination between probabilities while retaining the anchor. Thus, we presented the five probabilities within one trial separately and in ascending order while we kept the list-wise presentation format (Figure 11). Subjects still encoded all five probabilities before making their judgments.

Numeracy. Since we observed no systematic effects of numeracy in Experiments 1 and 2, in this experiment we employed for economic reasons the single-item version of the Berlin Numeracy Test recommended for educated samples (Cokely et al., 2012). The single-item numeracy test classified 35.9% of the sample (n = 33) as highly numerate. Numeracy did not differ between conditions, $\chi^2(3) = 6.21$, $p = .102$.

We only found one significant effect of numeracy. Highly numerate subjects using the rating scale were more sensitive to high probabilities in the low-error condition, $t(22) = -2.53$, $p = .019$, $r = .47$. Thus, again we omit numeracy from all further analyses.

Results

We excluded 9 out of the total of 552 trials (1.63%) as outliers. Figure 12 presents mean subjective probability judgments. Visual inspection reveals much steeper judgment functions in the low-error condition on both scale formats. Irrespective of error, percent estimates showed the expected within-context regression. Judgments on the rating scale,

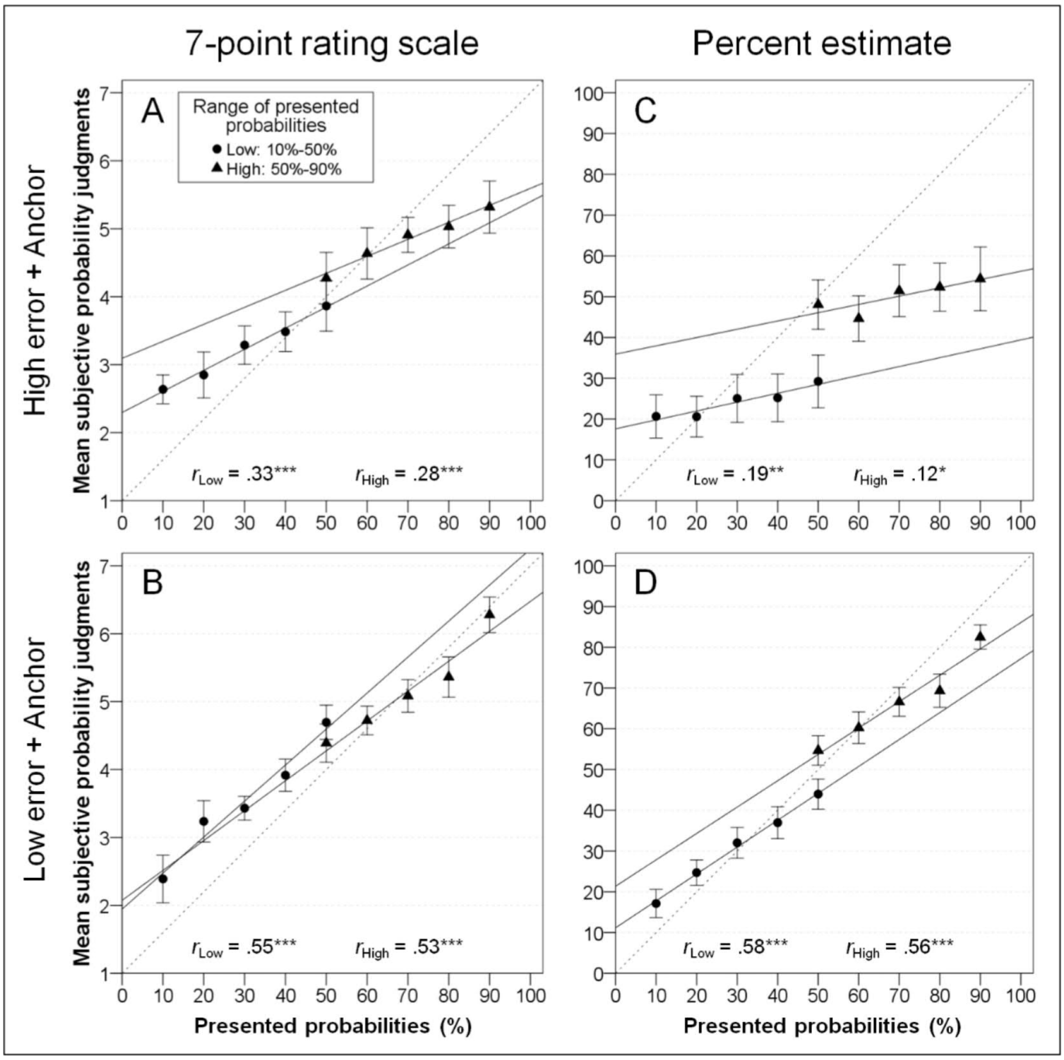

Figure 12. Mean subjective probability judgments as a function of presented probabilities (A: $n = 22$, B: $n = 24$, C: $n = 21$, D: $n = 25$). Solid lines represent the best linear fit. Correlation coefficients indicate scale sensitivity at the aggregate score level. Error bars = 95% within-subjects CIs. *$p < .05$. **$p < .005$. ***$p < .001$.

again, appeared to be anchored around the scale's midpoint, though the relation of the two respective common stimulus judgments was reversed as a function of encoding error.

Sensitivity. We recoded perfect sensitivity in five out of 526 (0.95%) correlations. Figure 13 presents mean sensitivity. The scores in the high-error condition (range = .31–.54) were comparable to Experiment 2, though percent estimates were a little less sensitive. The scores in the low-error condition (range = .71–.88) were, as expected, much higher. Accordingly, a mixed-design ANOVA found a large effect of the encoding condition on individual sensitivity, $F(1, 86) = 41.63$, $p < .001$, $\eta_p^2 = .33$. Additionally, subjects were a little less sensitive to high probabilities, $F(1, 86) = 4.35$, $p = .040$, $\eta_p^2 = .05$. This effect was more pronounced in the low-error condition, interaction: $F(1, 86) = 2.90$, $p = .092$, $\eta^2_p = .03$. There was no main effect of the scale format on individual sensitivity, $F < 1$. However, a significant interaction indicated that in the high-error condition percent estimates were less sensitive than judgments on the rating scale while the opposite was the case in the low-error condition, $F(1, 86) = 3.97$, $p = .049$, $\eta^2_p = .04$.

We observed the same pattern of results at the aggregate score level. The omnibus *Q*-test found significant differences between aggregate sensitivities for both ranges, low: $Q = 50.39$, $df = 3$, $p < .001$; high: $Q = 61.47$, $df = 3$, $p < .001$. Single comparisons using Fisher's *z*-test revealed that aggregate sensitivity was significantly higher in the low-error condition within as well as across scale formats, low range: zs $\geq |3.57|$, ps $< .001$; high: zs $\geq |3.90|$, ps $< .001$. Additionally, like on the individual level, when the error was high, percent estimates were less sensitive than judgments on the rating scale, low range: $z = 1.92$, $p = .055$; high: $z = 2.20$, $p = .028$. However, assuming a Bonferroni corrected significance criterion of .008, neither effect would be qualified as significant. It is further noteworthy that in the low-error condition, aggregate sensitivity was virtually identical for both scale formats, zs < 1. Aggregate sensitivity did not differ between ranges of probabilities, ts $\leq |1.03|$, ps $\geq .307$.

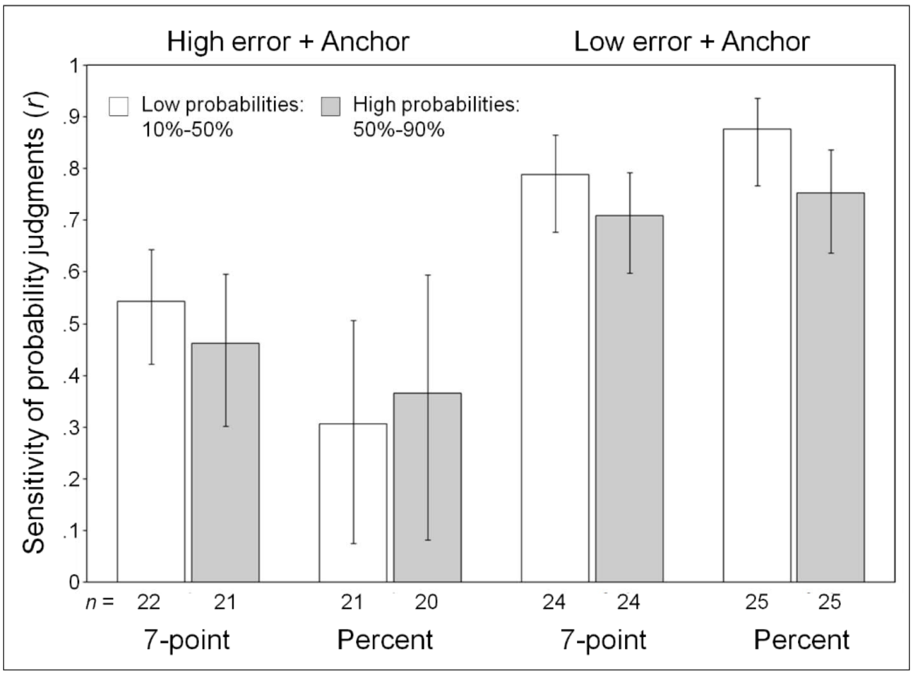

Figure 13. Sensitivity at the individual score level of both scale formats for both ranges of probabilities and both encoding conditions. Error bars = 95% CIs.

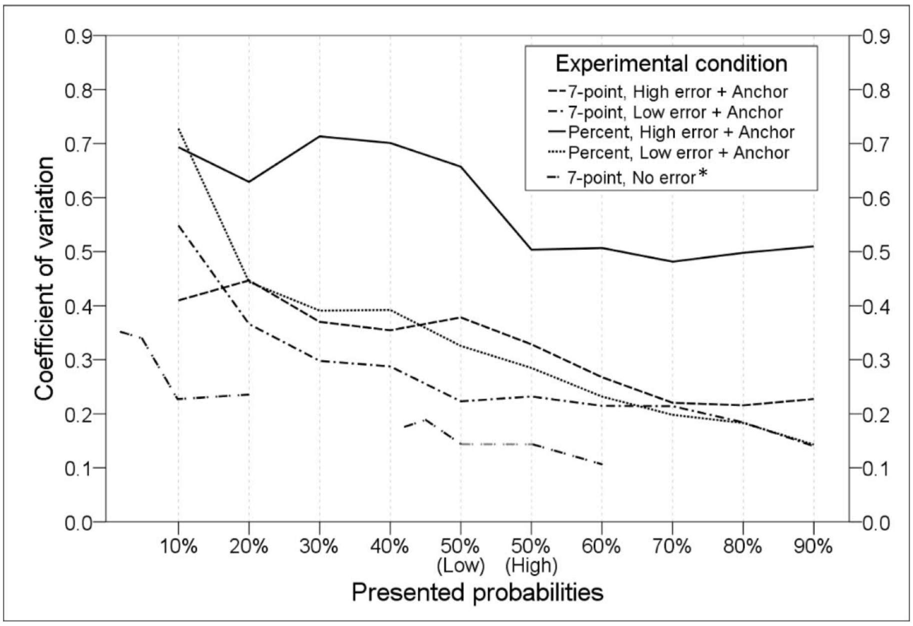

Figure 14. Coefficients of variation as variability profiles over all probability judgments for all between-subjects conditions. *The disconnected profile sections are based on data from Haase et al. (2013). The semi-transparent connection between the two 50% judgments indicates that the estimates of higher probabilities formed one context. Subjects in that study judged 50% only once.

Error in judgments. Mean *CV*s for judgments on the rating scale were: high-error = 0.322, low = 0.271; and for percent estimates: high-error = 0.589, low = 0.332 (Figure 14).

Remarkably, the manipulation of encoding error only affected percent estimates significantly, $t(136) = 3.42$, $p = .001$. $r = .28$ while the variation of judgments on the rating scale did not differ between encoding conditions, $t(136) = 1.29$, $p = .199$, $r = .11$. High-error percent estimates also showed more variation than the rating scale, $t\text{s} \geq |3.53$, $p\text{s} \leq .001$, rs $\geq .30$ while low error estimates did not differ from it, $t\text{s} \leq |1.57|$, $p\text{s} \geq .119$

Common stimulus. The upper half of Figure 15 presents probability judgments of the common stimulus. In both encoding conditions percent estimates of the common stimulus were significantly lower in the low probability range than in the high range, high-error: $t(20) = -3.61$, $p = .002$, $r = .63$; low: $t(24) = -4.56$, $p < .001$, $r = .68$. Judgments on the rating scale, again, did not differ significantly between probability ranges. However, it is noteworthy that the trend in the score difference was reversed between encoding conditions, high-error: $t(21) = -1.34$, $p = .195$. $r = .28$; low: $t(23) = 1.44$, $p = .16$. $r = .29$.

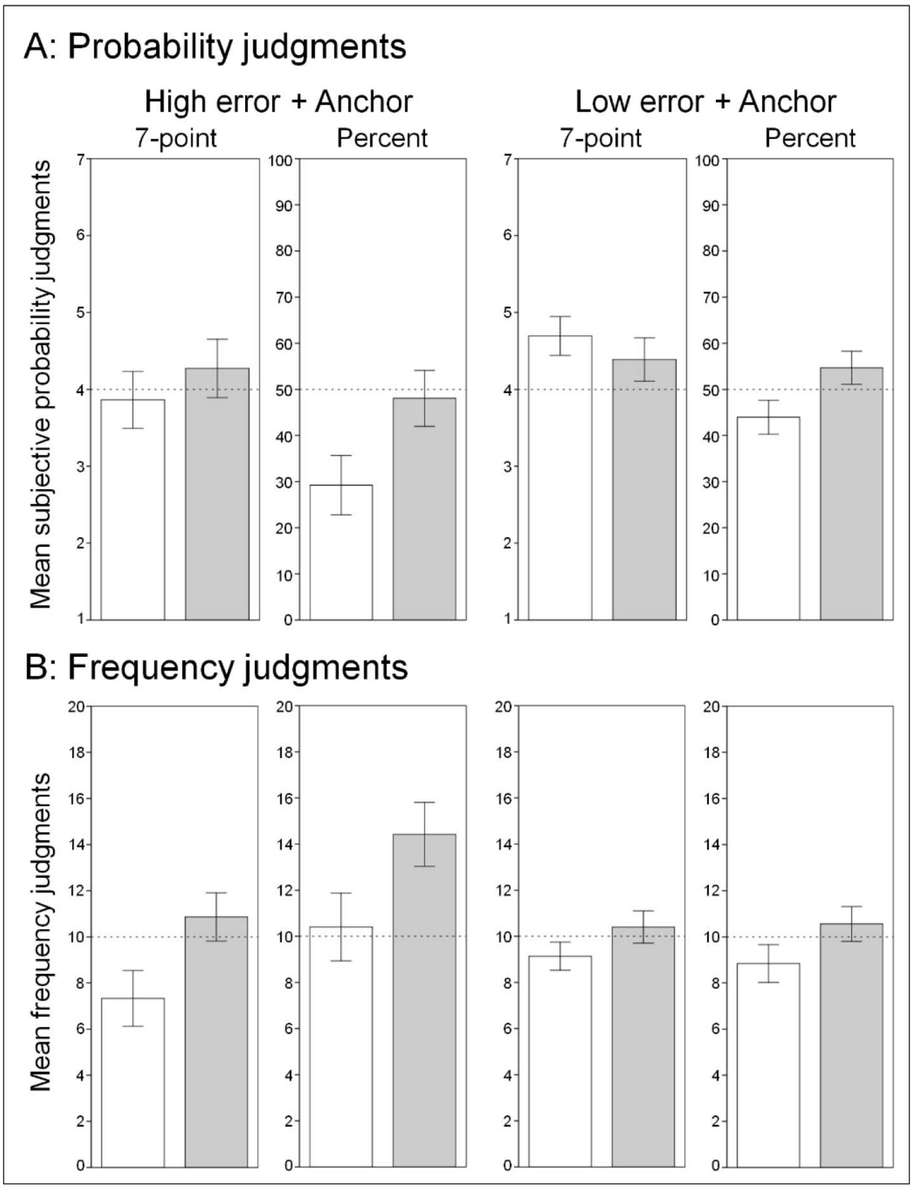

Figure 15. Judgments of the common stimulus. White bars = low range, grey bars = high range. Error bars = 95% within-subjects CIs. A: Presented probability = 50%. The dotted line represents the scale's respective midpoint. B: Presented frequency = 10 (dotted line).

Frequency judgments. We excluded 16 out of the total of 552 trials (2.90%) as outliers. Appendix C presents scatter plots and the detailed analyses regarding the sensitivity of and relative error in frequency estimates. In the high-error condition (like in the identical anchor condition in Experiment 2), subjects who used the percent format estimated frequencies to be higher than those using the rating scale while in the low-error conditions, frequency estimates did not differ as a function of the probability judgment format.

Individual sensitivity was significantly higher when encoding error was low (range = .83–.91) as compared to when it was high (range = .33–.56). It was generally lower for higher frequencies but did not differ as a function of the scale format. At the aggregate level we found mostly the same pattern of results, though sensitivity was generally lower (high-error range = .12–.38, low-error range = .34–.67). *CV*s were higher when the error was high and did not differ between scale formats (7-point, high-error = 0.483, low = 0.383; percent, high-error = 0.548, low = 0.332).

The lower half of Figure 15 presents frequency estimates of the common stimulus. In all four conditions, estimates were significantly lower in the low range as compared to the high range. This effect was only slightly larger in the high-error encoding condition, 7-point: $t(21) = -3.91$, $p = .001$, $r = .65$; percent: $t(20) = -4.15$, $p < .001$, $r = .68$, as compared to low-error encoding, 7-point: $t(23) = -2.37$, $p = .027$, $r = .44$; percent: $t(24) = -2.89$, $p = .008$, $r = .51$.

Discussion

The reduction of random error in the event sampling process led to the expected increase of scale sensitivity at the individual as well as the aggregate score level in both formats. There were additional noteworthy trends in the data. Like in the previous experiments, when the encoding error was extremely high, percent estimates tended to be less sensitive than judgments on the rating scale (especially at the aggregate level). When the error was significantly lower, on the other hand, this trend tended to be reversed. Furthermore, with less error, there were some indications of a range effect which is in line with the findings in Haase et al. (2013). High range probability judgments as well as frequency estimates tended to be less sensitive as it is harder to track many instances of the focal event than just few. For this effect to be visible, however, we believe that a certain level of precision in representations is necessary. In all other conditions it was drowned out by the generally high level of error.

With regard to capturing the error in the frequency representations which underlie probability judgments, the *CV*s proved to be very telling. While a reduction of error significantly reduced the variability in percent estimates, it had no effect on the *CV*s of judgments on the rating scale. For comparison we included *CV*s from the study by Haase et al. (2013) in Figure 14. These are based on a control condition, in which subjects simply read probabilities of an icon array and transferred these into judgments on different scale formats. For the percent format, variation and *CV*s dropped essentially to zero (not shown in the figure). Judgments on the rating scale, however, varied almost to the same degree as in the present study. Thus, it appears that the variation in rating scale judgments does not, or only to a small degree, stem from error in representations but rather from the scale-inherent imprecision.

This disconnect was again evident in the judgments of the common stimulus. Percent estimates were regressed within-context and differed accordingly for the common stimulus in both encoding conditions. The difference in actual scores was smaller in the low-error condition, though the effect sizes were, due to the higher measurement precision, very similar. Crucially, this pattern matched the corresponding frequency estimates very closely. In contrast, judgments on the rating scale were again anchored around the scale's midpoint. When the error was high and the five probabilities could not be clearly discriminated, the placement of all judgments was presumably only guided by the anchor and thus the difference between the common stimulus judgments, though not significant, followed the same trend as percent and frequency estimates. However, when discrimination was easier and judgment functions accordingly steeper, this trend was reversed and did not match the corresponding within-context regressed frequency estimates (which, due to their higher resolution, allow for discrimination without violating the anchor; Figure C1).

We must note one caveat in our interpretation. The frequency estimates served as a standard of comparison for the probability judgments and, thus, should have been identical irrespective of the scale format. However, in the high-error condition, subjects who used the percent format provided higher frequency estimates than subjects who used the rating scale. These differences were consistent across stimuli, that is, the judgment functions remained parallel (Figure C1) and thus do not impede our conclusions which are based on the relation between estimates rather than their absolute value. Nonetheless, they might indicate unexpected carry-over effects between probability and frequency estimates that should be avoided in future research.

Additional Analysis: Recoding the 7-Point Rating Scale

The accuracy of scale formats is typically assessed as deviations in judgments from an objective norm, which, of course, requires the compared values to occupy identical value arrays. However, transforming scale values to match a predefined range is problematic because any transformation of this kind is somewhat arbitrary as any number of different transformation functions is conceivable. For instance, in the present context one could simply divide the value array of 0–100 into seven equal segments and assign the segments' midpoints to the categories on the rating scale. One could just as well fix the endpoints first at 0 and 100 and then divide the remaining array in equal segments (e.g., Parducci & Wedell, 1986), which would result in drastically different values. Additionally, both these approaches assume that subjects interpret category labels to represent equally large sections of the probability continuum and to be spread equidistantly along it. Both assumptions have been refuted

(Bocklisch et al., 2010) and, furthermore, research indicates that there are vast interindividual differences in the translation of verbal qualifiers into numeric probability expressions (e.g., Budescu & Wallsten, 1985).

In light of these findings, some researchers have suggested to collect individual transformation functions in order to translate vague quantifier judgments into numeric estimates (Al Baghal, 2014; Bradburn & Miles, 1979). Thus, in Experiments 2 and 3 we asked subjects to assign numeric percent values to the verbal labels of the 7-point rating scale to find out whether it is at least theoretically viable. Subjects could either provide point estimates or an interval of values. In the latter case we used the interval's midpoint to assign a numeric value to the respective category. Figure 16 presents recoded as well as the original scale judgments. Note that the right *y*-axis is scaled to represent the 7-point rating scale.

For the analysis we pooled the data from both experiments, though separate analyses revealed identical results. For the most part, subjects provided intervals (88%). These differed significantly in width, $F(6, 570) = 22.96$, $p < .001$, $\eta_p^2 = .20$. On average, the percent intervals assigned to the central categories 3, 4, 5, and 6 were 1.5 times as wide as those assigned to categories 2 and 7, and almost 3 times as wide as the interval assigned to the smallest category 1. This explains why the recoded judgments in Figure 16 (solid black lines) diverge more from the original rating scale judgments (grey lines) in the lower half of all graphs as well as at the upper extreme in the low-error panel on the right.

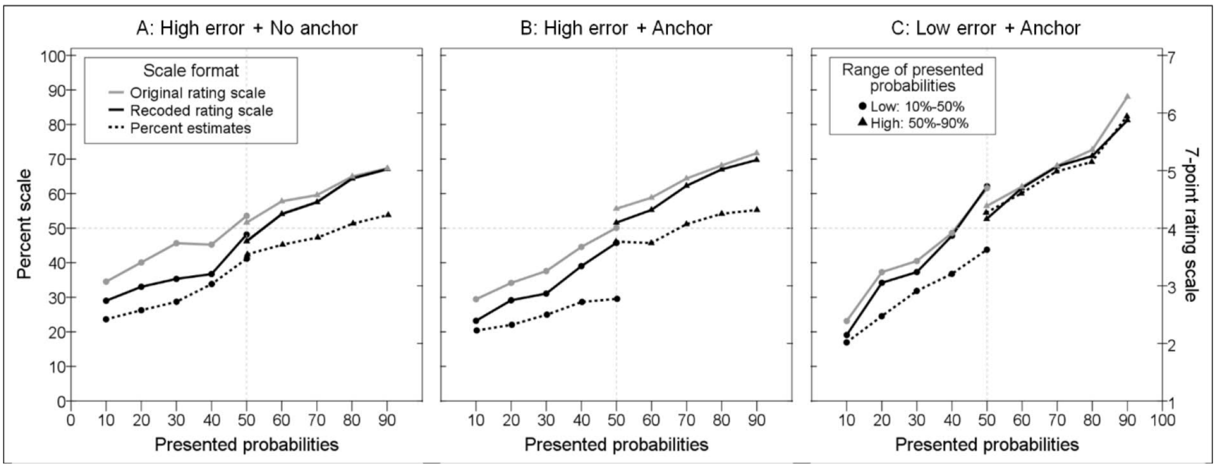

Figure 16. Recoded and original judgments in Experiments 2 and 3 (Rating scale, A: $n = 28$, B: $n = 47$, C: $n = 24$; Percent estimates, A: $n = 27$, B: $n = 46$, C: $n = 25$).

A mixed-design ANOVA found that the transformation functions did not differ between encoding modes, $F < 1$. Keep in mind that subjects assigned percent values to rating scale categories in a theoretical or idealized sense. This can be understood as the counterpart to using the rating scale for actual judgments with perfect knowledge (as explained in the introduction). However, all experiments indicated that when representations are perturbed by random error, subjects do not use the categories to coherently map the probability continuum but follow other principles such as expressing the relation of stimuli to an anchor or for ordinal stimulus discrimination. Therefore, when these category ratings are transformed with a function that relies on an error-free matching, the resultant percent values will be similarly distorted. Accordingly, we found in three separate ANOVAs that the recoded ratings were significantly higher than actual percent estimates in all encoding conditions, Fs ≥ 5.96, ps $\leq .018$, η^2_ps $\geq .10$, though in the low-error condition (Figure 16 C) this difference was not significant for high range probabilities, $F(6.62, 311.16) = 3.17$, $p = .004$, $\eta^2_p = .06$ (Mauchly's test: $\chi^2(44) = 69.33$, $p = .009$; Greenhouse-Geisser $\varepsilon = .74$). Even though percent estimates may have been artificially depressed to some degree (discussed below), if one follows the logic behind the transformation approach, Figure 16 C, for example, would indicate that subjects really judged the common stimulus in the low range to be higher than in the high range while actual percent estimates and, more importantly frequency estimates, suggest the opposite.

To sum up, setting aside that it seems at least impractical if not unreasonable to employ a verbal rating scale in any applied setting and to additionally ask subjects for a translation of that scale, our findings indicate that whenever the judged representations contain some error, this approach will likely lead to false conclusions. Only under conditions of perfect knowledge can such a transformation lead to reasonable estimates. But under such conditions, a low-resolution rating scale will, of course, also entail a loss of sensitivity. Thus, we conclude that rating scales cannot be used to measure absolute perceptions of subjective probability.

General Discussion

In three experiments subjects encoded the same two ranges of objective probabilities as sequences of option-outcome pairs and judged subjective probability either on a verbally labeled 7-point rating scale or in the form of percent estimates. The two ranges shared one common stimulus and were presented within-subjects. We varied the error in encoding as well as the presentation format and found that both rather subtle manipulations changed the judgments on both scales markedly and can lead to false conclusions in either case. The main reason for this lies in the different ways these scale formats react to random error in the representations on which judgments are based.

In Experiment 1 we established that in this paradigm both scale formats were used to express the same construct and observed that under highly error-prone conditions the high resolution of the percent format offered no advantage in terms of sensitivity and might even lead to a decrease at the aggregate level. Additionally, imprecise representations, rather than the stimulus distribution, can result in apparent context effects on the rating scale but are unlikely in percent estimates. In Experiment 2 we found that an anchor in the stimuli can tether judgments on the rating scale to the scale's midpoint but in turn result in inconsistent judgment functions on the percent format. These could be interpreted as context effects, though they are, again, not caused by the stimulus distribution but an expression of regression toward the mean. We concluded further that the rating scale may deliver results that appear to be normatively consistent but fail to capture the actual imprecision in the representations which underlie subjective probability judgments. In Experiment 3 we followed up on this and discovered that most of the variation in rating scale judgments stems from the scale-inherent imprecision rather than unreliable representations. The results further indicated, though, that imprecise information changes the way the rating scale is used while percent estimates remain consistent.

Before we discuss the two scale formats in more detail we should note that the results, especially of Experiments 1 and 2, present strong evidence for the hypothesis that event frequencies are encoded automatically (Hasher & Zacks, 1984; Zacks & Hasher, 2002). The sampling process in those experiments was extremely difficult and resulted in highly imprecise estimates. Nonetheless, these estimates captured the relative relations of event frequencies and probabilities. Still, this raises the question about the applicability of our results, that is, in which research settings can we expect to collect probability or frequency judgments that are perturbed by this much noise? Unfortunately, the literature does not provide a clear answer as sensitivity is rarely reported or it is based on scores that reduce some of the error in the estimates. For instance, Attneave (1953) reports a correlation of .79 between the frequency of letters in the English language and median direct estimates. Lichtenstein, Slovic, Fischhoff, Layman, and Combs (1978) report in their study of judged frequency of lethal events correlations between .82 and .91 that are based on geometric means. In the same study, analyses of individual sensitivity (.28–.92) were based on log transformed responses. Notwithstanding this lack of clarity, the main purpose of the high error in the present experiments was to elucidate how it affects scale usage.

The Verbally Labeled 7-Point Rating Scale

Theoretically, the probability continuum can be mapped on a rating scale with a limited number of categories when both arrays share the same endpoints, that is, when the extreme categories indicate endpoints (e.g., *almost certain*) rather than just a relativistic increase (e.g., *very very likely*). Such a mapping necessarily implies that similar probabilities share a category. However, Haase et al. (2013) showed that when representations of the probabilities are imprecise, subjects using a vague rating scale tend to abandon a consistent mapping in favor of ordinal stimulus discrimination. The authors concluded that a rating scale therefore does not allow a meaningful quantification of probability judgments and thus no between-subjects comparisons.

In the present study we expanded this notion to within-subjects designs and even more error-prone conditions. We observed that when representations are too imprecise to allow for ordinal discrimination judgments on the rating scale will be anchored around the scale's midpoint. This finding is consistent with the theoretical framework of fuzzy-trace theory (Reyna, 2012) in that it represents a shift from an ordinal to a nominal scale. The probability continuum has a clearly defined and meaningful midpoint at 50%. Probabilities above indicate that an event will rather happen, those below that it will rather not happen. When representations only allow a dichotomous discrimination, anchoring the judgments at the midpoint is reasonable and, in a sense, even represents a consistent scale use across context. However, as we have shown, it can also lead to the false conclusion that a common stimulus from two different judgment contexts was perceived to be equal in magnitude while direct estimates of the underlying frequency representations clearly indicated a difference.

In the third experiment we have shown that different levels of gist representation can even have a competing influence on judgments on a rating scale. The nominal discrimination pushed the common stimulus toward its respective context while the ordinal discrimination pushed in the opposite direction. It would be interesting to see how a removal of the anchor from the stimuli would affect judgments in a low-error condition. More importantly, though, the manipulation of encoding error revealed that the variation in rating scale judgments does not reflect the imprecision in representations. We would therefore argue that, even when used for estimation, probability judgments on a rating scale are not actually magnitude estimates but merely a form of categorization.[12] More importantly, the rules which guide this categorization change as a function of the precision in the underlying representations.

A likely reason for this is that the vague qualifiers of a verbal scale do not pressure subjects into concerns about normative accuracy (Windschitl & Wells, 1996). There is evidence that subjects' preferences for communicating probabilities in a verbal or numerical format differs with the precision of the representations that are to be conveyed, which supports this reasoning. With increasingly precise knowledge preferences shift from verbal to numerical (Wallsten, Budescu, Zwick, & Kemp, 1993; Wallsten & Budescu, 1995; Witteman, Renooij, & Koele, 2007). As a way to counteract the arbitrariness of the verbal scale, it has been suggested to combine verbal and numerical labels which would fasten segments of the probability continuum to the categories (Witteman & Renooij, 2003). While we would be curious to see how such a scale fares in our research paradigm we do not really expect any benefit. On the contrary, we would expect a drop in sensitivity because, on the one hand, the numerical labels suppress the scale-free ordinal discrimination while, on the other hand, the scale, of course, still only has a very low resolution.

What remains then is the question of whether a verbal rating scale can ever be used reasonably in research on subjective probability. The only application that we can think of is the assessment of ordinal relations of a limited number of different probabilities with the number of categories as an upper limit. Why this should be favored over simply asking for a direct ranking, however, is beyond us.

Percent Estimates

In most judgment and decision research, the percent format has a bad reputation as being too hard for subjects to handle. This is mainly due to a number of studies that have shown that biases in Bayesian reasoning disappear when likelihood information is presented in the form of natural frequencies (not relative frequencies) instead of probabilities (e.g., Gigerenzer & Hoffrage, 1995). However, inference is not the same as estimation. For instance, numeracy, which essentially measures the ability to solve simple inference problems, had no effect on percent estimates in any of our experiments.[13] We found percent

[12] Note that this interpretation is different from the notion that verbal ratings express intuitive perceptions of certainty as discussed in the introduction.

[13] For this purpose, we additionally assessed the accuracy of percent estimates as absolute error and as root mean square deviations and also found no relation with numeracy. These analyses are not included in the paper because, as delineated above, they cannot be calculated for the rating scale and thus do not allow a comparison between scale formats.

estimates to provide the most reliable measurements of subjective probability and would recommend them as the format of choice whenever judgments need to be compared to a normative standard or across conditions. They are highly sensitive and relate consistently to the underlying representations. This, however, does not mean that they cannot lead to false conclusions.

The defining features of the percent scale are its high resolution and the exactness of its categories (i.e., integers). These properties are also the reason why this format is often liked the least by subjects (Diefenbach et al., 1993; Eibner et al., 2006). It asks for precise estimates even when representations are not precise. One undesired consequence of this is an overreliance on readily available anchors such as multiples of ten or five when providing percent estimates (Ariely et al., 2000; Manski & Molinari, 2010; Wallsten, Budescu, & Zwick, 1993). Nonetheless the analyses of the *CV*s in this study and the study by Haase et al. (2013) clearly indicates that percent estimates are very sensitive to error and we would argue that this is a strength of the format because the imprecision is a part of the representation. It also means that researchers need to be aware of it.

In the introduction we wrote that the percent scale might introduce more noise due to its high resolution. In light of our findings and in line with previous research (Pleskac, Dougherty, Rivadeneira, & Wallsten, 2009) we want to clarify this statement. The percent format does not introduce but capture noise. For instance, the sensitivity of the percent scale was generally lower than that of the rating scale at the aggregate but not at the individual score level because the percent estimates accurately reflected interindividual differences. Thus, additionally analyzing individual scores may help to avoid false conclusions in many research contexts.

Being cognizant of the error in percent estimates is, of course, much more important in the analyses of mean judgments lest one fall prey to a classic regression fallacy. In Experiments 2 and 3 we observed lower estimates for the 50% probability in the low range as compared to the high range when the stimuli provided an anchor. These differences are an expression of the within-context regression toward the mean. This does not imply, however, that subjects did not truly differ in their estimates—of course they did! It only implies that these differences mean nothing beyond the fact that the representations underlying the estimates were perturbed by random error. It is prudent to keep this in mind when comparing experimental conditions that are supposed to represent another manipulated factor. While the midpoint of the probability continuum is certainly a very salient anchor in many contexts, the same logic applies to the whole range, as well as to any other single or number of arbitrary anchors in the stimuli or on the measuring instrument itself (Hollands & Dyre, 2000). The important point is that the percent format remains consistent in how it translates error and anchors into estimates.

There is one caveat to this statement. We observed that in all high-error conditions the judgment functions of the percent estimates did not cross the identity line at the midpoint (50% without an anchor, 25% and 75% respectively with an anchor) which would be expected as judgments should be regressed symmetrically. Instead, the "turning points" between over- and underestimation lay at around 30% (no anchor) and 20% and 40% respectively (anchor). One explanation for this could be that the representations of the high range stimuli were noisier because it is harder to track many instances of a focal event than just few. In that case, however, the judgments should differ in sensitivity between the ranges which was generally not the case. More importantly, the corresponding frequency estimates should show a similar pattern, but they crossed the identity lines roughly around the respective midpoints. A more likely explanation for this discrepancy could lie in the mental representation of numbers themselves. Keep in mind that subjects estimated both probabilities and frequencies as integers. However, the former covered a range from 0–100 (presumably not starting below 10) while the latter only went up to estimates of around 20. Some research indicates that numbers are represented by magnitudes that have scalar variability, that is, a constant coefficient of variation (e.g., Whalen, Gallistel, & Gelman, 1999). This property implies an increase in noise for estimates of larger magnitudes. Thus, even though subjects based their judgments on identical representations, expressing these judgments with larger numbers might have additionally increased the error (see Cohen, Ferrell, & Johnson, 2002 for comparable findings).

Note, however, that both judgment functions in the low-error condition did not show this pattern but were regressed almost symmetrically. Additionally, their noise did not increase with magnitude. We found high positive correlations between mean percent estimates and their corresponding standard deviations in all high-error conditions, $r\text{s} \geq .79$, $p\text{s} \leq .007$, but not in the low-error condition, $r = .15$, $p = .685$, indicating that those judgments did not have scalar variability. It seems that with increasing noise in the underlying representations the percent scale becomes more alike to analog magnitude estimates and less alike to a high-resolution category scale (see Parducci & Wedell, 1986 for comparable findings and a similar interpretation).

Nonetheless, the property of scalar variability implies flat variability profiles. However, the *CV* profiles for all percent estimates, including high-error conditions, show a marked downward trend (Figures 4, 9, & 14). The *CV*s for the frequency estimates, on the other hand, do not show any trend and remain reasonably level (Figures B3 & C3). Previous research indicates differences in scalar variability between unbounded (e.g., frequency) and bounded (e.g., probability) estimation tasks (Ebersbach, Luwel, Frick, Onghena, & Verschaffel, 2008; Ebersbach, Luwel, & Verschaffel, 2013). Here, this comparison is of course not entirely appropriate because of the highly different numerosities. It does, however, serve to illustrate that percent estimates were not simply mathematical transformations of previously provided frequency estimates but distinct judgments.

Context Effects in Probability Estimates

Each variation in any of the three experiments could be interpreted as a manipulation of the judgment context. However, as we explained in the introduction, we were interested in the kind of context effects that are created by the stimulus distributions and that have been studied predominantly within the framework of range-frequency theory (Parducci, 1965). From this point of view, we should have expected the same results in each experiment as the stimulus distributions were identical. We must note, though, that the applicability of range-frequency theory to our paradigm is certainly questionable. First, it has typically been applied to judgments made on explicitly relativistic scales with no clearly defined lower and upper bounds. The study by Varey et al. (1999), of course, is the most relevant exception. Furthermore, to our knowledge, it has not been applied to this form of stimuli. Conceptually, in each trial we presented five stimuli (i.e., probabilities) simultaneously (i.e., before any judgment). However, each stimulus was itself composed of 20 sequentially encoded separate stimuli (i.e., option-outcome pairs). A later and more elaborated version of the theory assumes that due to working memory constraints the effective judgment context is limited to the 12 previously presented stimuli (Parducci & Wedell, 1986). It is unclear, whether this limitation would be relevant here. Most importantly, range-frequency theory describes context effects in between-subjects designs while we varied the distributions within-subjects. Nonetheless we discuss our results in relation to this theory because they support an alternative account for context effects.

Haase et al. (2013) observed different judgment functions for different stimulus distributions in a between-subjects design. Even though there were marked differences between different scale formats, their findings could be in line with range-frequency theory which takes the number of categories on a scale into account (Parducci, 1982). However, these context effects also varied with the difficulty of encoding the stimuli, that is, with the precision of the underlying representations. Therefore, the authors suggested that fuzzy-trace theory (Reyna, 2012) might provide a better explanation. Our present findings support this conclusion. Furthermore, the within-subjects design provides a stronger test as it precludes the range-frequency account. We have observed that a change in the way a scale is used can give the appearance of context effects (Experiment 1) or, through the same mechanism, the appearance of normatively consistent probability judgments (Experiments 2 and 3). Crucially, as we observed for the latter two instances, these judgments contradicted direct assessments of the underlying frequency. Thus, these apparent context effects and their apparent absence were not located at the level of representation but were created by the specific scale format. In comparison, percent estimates have been consistent across different judgment contexts in between- (Haase et al., 2013) as well as in within-subjects designs (Experiment 1). The provision of an anchor in Experiments 2 and 3 led to context effects. However, the differences between judgments of a common stimulus were reversed to what range-frequency theory would predict and could be fully explained by taking the random error in the representations into account. It would be interesting to test whether the context effect that Varey et al. (1999) found were not, in fact, an expression of noisy judgments.

Limitations

The main limitation to our results and interpretations stems from the rather small sample sizes in all three experiments. Because it was crucial that all subjects based their judgments on identical information we were very conservative in excluding those for whom a proper encoding could not be confirmed. This resulted in a certain lack of power to detect significant differences, especially with regard to analyses based on the z-test.

Additionally, we observed unexpected differences in the frequency estimates that were most likely carry-over effect from the different probability scale formats. However, as explained above, we do not believe that these impede our conclusions.

Conclusion

We investigated the performance of two scale formats for subjective probability when objective probabilities were provided. We found that a verbally labeled 7-point rating scale did not assess actual estimates of subjective probability but was rather used for categorization of stimuli. Moreover, the rules that guided this categorization depended upon the precision of the available information. This can lead to the appearance of context effects or of normatively correct judgments that are, however, not representative of the underlying representations. Thus, as this scale does not allow meaningful comparisons between experimental conditions even in within-subjects designs, we conclude that it should not be used in research on subjective probability.

Instead we recommend assessing probability judgments as percent estimates. The percent format is highly sensitive and captures the underlying representations reliably and for the most part consistently. It is however also sensitive to random error in said representation. Not being aware of this can lead to classic regression fallacies especially when the stimuli present salient anchors.

In light of the many different ways that we have found in which different scale formats for subjective probability can lead to false conclusions we would like to close with a quotation from Stanley Smith Stevens' and Eugene H. Galanter's seminal paper on ratio and category scales (1957, p. 405). In their discussion of judgments of proportion, the authors laconically remark:

There are numerous pitfalls in this business.

References

Al Baghal, T. (2014). Numeric estimation and response options: An examination of the accuracy of numeric and vague quantifier responses. *Journal of Methods and Measurement in the Social Sciences, 5*(2), 58–75. https://doi.org/10.2458/azu_jmmss_v5i2_Al_baghal

Ariely, D., Au, W. T., Bender, R. H., Budescu, D. V, Dietz, C. B., Gu, H., … Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied, 6*(2), 130–147. https://doi.org/10.1037/1076-898X.6.2.130

Attneave, F. (1953). Psychological probability as a function of experienced frequency. *Journal of Experimental Psychology, 46*(2), 81–86. https://doi.org/10.1037/h0057955

Baghal, T. (2011). The measurement of risk perceptions: The case of smoking. *Journal of Risk Research, 14*(3), 351–364. https://doi.org/10.1080/13669877.2010.541559

Baguley, T. (2012). Calculating and graphing within-subject confidence intervals for ANOVA. *Behavior Research Methods, 44*(1), 158–175. https://doi.org/10.3758/s13428-011-0123-7

Baillon, A. (2008). Eliciting subjective probabilities through exchangeable events: An advantage and a limitation. *Decision Analysis, 5*(2), 76–87. https://doi.org/10.1287/deca.1080.0113

Beach, L. R., & Phillips, L. D. (1967). Subjective probabilities inferred from estimates and bets. *Journal of Experimental Psychology, 75*(3), 354–359. https://doi.org/10.1037/h0025061

Beach, L. R., & Wise, J. A. (1969). Subjective probability and decision strategy. *Journal of Experimental Psychology, 79*(1), 133–138. https://doi.org/10.1037/h0026959

Betsch, C., Haase, N., Renkewitz, F., & Schmid, P. (2015). The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions? *Judgment and Decision Making, 10*(3), 241–264. Retrieved from http://journal.sjdm.org/14/141206a/jdm141206a.pdf

Betsch, T., Glauer, M., Renkewitz, F., Winkler, I., & Sedlmeier, P. (2010). Encoding, storage and judgment of experienced frequency and duration. *Judgment and Decision Making, 5*(5), 347–364. Retrieved from http://journal.sjdm.org/10/91221b/jdm91221b.pdf

Bilgin, B. (2012). Losses loom more likely than gains: Propensity to imagine losses increases their subjective probability. *Organizational Behavior and Human Decision Processes, 118*(2), 203–215. https://doi.org/10.1016/j.obhdp.2012.03.008

Birnbaum, M. H. (1974). Using contextual effects to derive psychophysical scales. *Perception & Psychophysics, 15*(1), 89–96. https://doi.org/10.3758/BF03205834

Birnbaum, M. H. (1999). How to show that 9 > 221: Collect judgments in a between-subjects design. *Psychological Methods, 4*(3), 243–249. https://doi.org/10.1037/1082-989X.4.3.243

Bocklisch, F., Bocklisch, S. F., & Krems, J. F. (2010). How to translate words into numbers? A fuzzy approach for the numerical translation of verbal probabilities. In E. Hüllermeier, R. Kruse, & F. Hoffmann (Eds.), *Computational intelligence for knowledge-based systems design. IPMU 2010. Lecture notes in computer science, vol 6178* (pp. 614–623). Berlin-Heidelberg: Springer. https://doi.org/10.1007/978-3-642-14049-5_63

Bradburn, N. M., & Miles, C. (1979). Vague quantifiers. *Public Opinion Quarterly, 43*(1), 92–101. https://doi.org/10.1086/268494

Budescu, D. V., & Wallsten, T. S. (1985). Consistency in interpretation of probabilistic phrases. *Organizational Behavior and Human Decision Processes, 36*(3), 391–405. https://doi.org/10.1016/0749-5978(85)90007-X

Budescu, D. V., Erev, I., & Wallsten, T. S. (1997). On the importance of random error in the study of probability judgment. Part I: New theoretical developments. *Journal of Behavioral Decision Making, 10*(3), 157–171. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<157::AID-BDM260>3.0.CO;2-#

Cohen, D. J., Ferrell, J. M., & Johnson, N. (2002). What very small numbers mean. *Journal of Experimental Psychology: General, 131*(3), 424–442. https://doi.org/10.1037/0096-3445.131.3.424

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin numeracy test. *Judgment and Decision Making, 7*(1), 25–47. Retrieved from http://journal.sjdm.org/11/11808/jdm11808.pdf

Corbin, J. C., Reyna, V. F., Weldon, R. B., & Brainerd, C. J. (2015). How reasoning, judgment, and decision making are colored by gist-based intuition: A fuzzy-trace theory approach. *Journal of Applied Research in Memory and Cognition, 4*(4), 344–355. https://doi.org/10.1016/j.jarmac.2015.09.001

Costello, F., & Watts, P. (2014). Surprisingly rational: Probability theory plus noise explains biases in judgment. *Psychological Review, 121*(3), 463–480. https://doi.org/10.1037/a0037010

Cousineau, D. (2005). Confidence intervals in within-subject designs: A simpler solution to Loftus and Masson's method. *Tutorials in Quantitative Methods for Psychology, 1*(1), 42–45. https://doi.org/10.20982/tqmp.01.1.p042

Cox, E. P. I. (1980). The optimal number of response alternatives for a scale: A review. *Journal of Marketing Research, 17*(4), 407–422. https://doi.org/10.2307/3150495

de Finetti, B. (1970). Logical foundations and measurement of subjective probability. *Acta Psychologica, 34*, 129–145. https://doi.org/10.1016/0001-6918(70)90012-0

Delavande, A., & Rohwedder, S. (2008). Eliciting subjective probabilities in internet surveys. *Public Opinion Quarterly, 72*(5), 866–891. https://doi.org/10.1093/poq/nfn062

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research, 8*(2), 181–192. https://doi.org/10.1093/her/8.2.181

Ebersbach, M., Luwel, K., Frick, A., Onghena, P., & Verschaffel, L. (2008). The relationship between the shape of the mental number line and familiarity with numbers in 5- to 9-year old children: Evidence for a segmented linear model. *Journal of Experimental Child Psychology, 99*(1), 1–17. https://doi.org/10.1016/j.jecp.2007.08.006

Ebersbach, M., Luwel, K., & Verschaffel, L. (2013). Comparing apples and pears in studies on magnitude estimations. *Frontiers in Psychology, 4*(332). https://doi.org/10.3389/fpsyg.2013.00332

Edwards, W. (1962). Subjective probabilities inferred from decisions. *Psychological Review, 69*(2), 109–135. https://doi.org/10.1037/h0038674

Eibner, F., Barth, J., Helmes, A., & Bengel, J. (2006). Variations in subjective breast cancer risk estimations when using different measurements for assessing breast cancer risk perception. *Health, Risk & Society, 8*(2), 197–210. https://doi.org/10.1080/13698570600677407

Erev, I., Wallsten, T. S., & Budescu, D. V. (1994). Simultaneous over- and underconfidence: The role of error in judgment processes. *Psychological Review, 101*(3), 519–527. https://doi.org/10.1037//0033-295X.101.3.519

Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: Development of the subjective numeracy scale. *Medical Decision Making, 27*(5), 672–80. https://doi.org/10.1177/0272989X07304449

Fleiss, J. L. (1993). The statistical basis of meta-analysis. *Statistical Methods in Medical Research, 2*(2), 121–145. https://doi.org/10.1177/096228029300200202

Galanter, E. H. (1962). The direct measurement of utility and subjective probability. *The American Journal of Psychology, 75*(2), 208–220. https://doi.org/10.2307/1419604

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review, 102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

Glöckner, A., Hilbig, B. E., Henninger, F., & Fiedler, S. (2016). The reversed description-experience gap: Disentangling sources of presentation format effects in risky choice. *Journal of Experimental Psychology: General, 145*(4), 486–508. https://doi.org/10.1037/a0040103

Goldstein, D. G., & Rothschild, D. (2014). Lay understanding of probability distributions. *Judgment and Decision Making, 9*(1), 1–14. Retrieved from http://journal.sjdm.org/13/131029/jdm131029.pdf

Gorsuch, R. L., & Lehmann, C. S. (2010). Correlation coefficients: Mean bias and confidence interval distortions. *Journal of Methods and Measurement in the Social Sciences, 1*(2), 52–65. https://doi.org/10.2458/azu_jmmss_v1i2_gorsuch

Haase, N., Renkewitz, F., & Betsch, C. (2013). The measurement of subjective probability: Evaluating the sensitivity and accuracy of various scales. *Risk Analysis, 33*(10), 1812–1828. https://doi.org/10.1111/risa.12025

Hasher, L., & Zacks, R. T. (1984). Automatic processing of fundamental information: the case of frequency of occurrence. *The American Psychologist, 39*(12), 1372–1388. https://doi.org/10.1037/0003-066X.39.12.1372

Haubensak, G. (1992). The consistency model: A process model for absolute judgments. *Journal of Experimental Psychology: Human Perception and Performance, 18*(1), 303–309. https://doi.org/10.1037/0096-1523.18.1.303

Hertwig, R., Pachur, T., & Kurzenhäuser, S. (2005). Judgments of risk frequencies: Tests of possible cognitive mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(4), 621–642. https://doi.org/10.1037/0278-7393.31.4.621

Hilbert, M. (2012). Toward a synthesis of cognitive biases: How noisy information processing can bias human decision making. *Psychological Bulletin, 138*(2), 211–237. https://doi.org/10.1037/a0025940

Hollands, J. G., & Dyre, B. P. (2000). Bias in proportion judgments: The cyclical power model. *Psychological Review, 107*(3), 500–524. https://doi.org/10.1037//0033-295X.107.3.500

Juslin, P., Olsson, H., & Björkman, M. (1997). Brunswikian and Thurstonian origins of bias in probability assessment: On the interpretation of stochastic components of judgment. *Journal of Behavioral Decision Making, 10*(3), 189–209. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<189::AID-BDM258>3.0.CO;2-4

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review, 80*(4), 237–251. https://doi.org/10.1037/h0034747

Keller, C., Siegrist, M., & Visschers, V. (2009). Effect of risk ladder format on risk perception in high- and low-numerate individuals. *Risk Analysis, 29*(9), 1255–1264. https://doi.org/10.1111/j.1539-6924.2009.01261.x

Kenny, D. A. (1987). *Statistics for the social and behavioral sciences*. Boston: Little, Brown & Co.

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766. https://doi.org/10.1016/j.jesp.2013.03.013

Lichtenstein, S., Slovic, P., Fischhoff, B., Layman, M., & Combs, B. (1978). Judged frequency of lethal events. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 551–578. https://doi.org/10.1037/0278-7393.4.6.551

Manski, C. F., & Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business & Economic Statistics, 28*(2), 219–231. https://doi.org/10.1198/jbes.2009.08098

Maraun, M. D., Gabriel, S. M., & Martin, J. (2011). The mythologization of regression towards the mean. *Theory & Psychology, 21*(6), 762–784. https://doi.org/10.1177/0959354310384910

Marsh, H. W. (1983). Neutral-point anchoring in ratings of personality-trait words. *The American Journal of Psychology, 96*(4), 513–526. https://doi.org/10.2307/1422572

Marsh, H. W., & Parducci, A. (1978). Natural anchoring at the neutral point of category rating scales. *Journal of Experimental Social Psychology, 14*, 193–204. https://doi.org/10.1016/0022-1031(78)90025-2

McKelvie, S. J. (1978). Graphic rating scales – How many categories? *British Journal of Psychology, 69*(2), 185–202. https://doi.org/10.1111/j.2044-8295.1978.tb01647.x

McKelvie, S. J. (2001). Factors affecting subjective estimates of magnitude: When is 9 > 221? *Perceptual and Motor Skills, 93*(2), 432–434. https://doi.org/10.2466/pms.2001.93.2.432

Mellers, B. A., & Birnbaum, M. H. (1982). Loci of contextual effects in judgment. *Journal of Experimental Psychology: Human Perception and Performance, 8*(4), 582–601. https://doi.org/10.1037/0096-1523.8.4.582

Mellers, B. A., & Birnbaum, M. H. (1983). Contextual effects in social judgment. *Journal of Experimental Social Psychology, 19*(2), 157–171. https://doi.org/10.1016/0022-1031(83)90035-5

Morey, R. D. (2008). Confidence intervals from normalized data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology, 4*(2), 61–64. https://doi.org/10.3758/s13414-012-0291-2

Nickerson, C. A. E. (1995). Does willingness to pay reflect the purchase of moral satisfaction? A reconsideration of Kahneman and Knetsch. *Journal of Environmental Economics and Management, 28*(1), 126–133. https://doi.org/10.1006/jeem.1995.1009

Parducci, A. (1965). Category judgment: A range-frequency model. *Psychological Review, 72*(6), 407–418. https://doi.org/10.1037/h0022602

Parducci, A. (1982). Category ratings: Still more contextual effects! In B. Wegener (Ed.), *Social Attitudes and Psychophysical Measurement* (pp. 89–105). Hillsdale, NJ: Erlbaum.

Parducci, A., & Perrett, L. F. (1971). Category rating scales: Effects of relative spacing and frequency of stimulus values. *Journal of Experimental Psychology Monograph, 89*(2), 427–452. https://doi.org/10.1037/h0031258

Parducci, A., & Wedell, D. H. (1986). The category effect with rating scales: Number of categories, number of stimuli, and method of presentation. *Journal of Experimental Psychology: Human Perception and Performance, 12*(4), 496–516. https://doi.org/10.1037/0096-1523.12.4.496

Peters, E., & Bjalkebring, P. (2014). Multiple numeric competencies: When a number is not just a number. *Journal of Personality and Social Psychology, 108*(5), 802–822. https://doi.org/10.1037/pspp0000019

Pfeifer, P. E. (1994). Are we overconfident in the belief that probability forecasters are overconfident? *Organizational Behavior and Human Decision Processes, 58*(2), 203–213. https://doi.org/10.1006/obhd.1994.1034

Pleskac, T. J., Dougherty, M. R., Rivadeneira, A. W., & Wallsten, T. S. (2009). Random error in judgment: The contribution of encoding and retrieval processes. *Journal of Memory and Language, 60*(1), 165–179. https://doi.org/10.1016/j.jml.2008.08.003

Preston, C. C., & Colman, A. M. (2000). Optimal number of response categories in rating scales: Reliability, validity, discriminating power, and respondent preferences. *Acta Psychologica, 104*(1), 1–15. https://doi.org/10.1016/S0001-6918(99)00050-5

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment and Decision Making, 7*(3), 332–359. Retrieved from http://journal.sjdm.org/11/111031/jdm111031.pdf

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: Some foundational issues. *Learning and Individual Differences, 7*(2), 145–162. https://doi.org/10.1016/1041-6080(95)90028-4

Sarris, V., & Parducci, A. (1978). Multiple anchoring of category rating scales. *Perception & Psychophysics, 24*(1), 35–39. https://doi.org/10.3758/BF03202971

Schapira, M., Davids, S., McAuliffe, T. L., & Nattinger, A. B. (2004). Agreement between scales in the measurement of breast cancer risk perceptions. *Risk Analysis, 24*(3), 665–673. https://doi.org/10.1111/j.0272-4332.2004.00466.x

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine, 127*(11), 966–972. https://doi.org/10.7326/0003-4819-127-11-199712010-00003

Sedlmeier, P., Hertwig, R., & Gigerenzer, G. (1998). Are judgments of the positional frequencies of letters systematically biased due to availability? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*(3), 754–770. https://doi.org/10.1037/0278-7393.24.3.754

Sokal, R. R., & Braumann, C. A. (1980). Significance tests for coefficients of variation and variability profiles. *Systematic Zoology, 29*(1), 50–66. https://doi.org/10.1093/sysbio/29.1.50

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

Stevens, S. S., & Galanter, E. H. (1957). Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology, 54*(6), 377–411. https://doi.org/10.1037/h0043680

Sun, Y., Wang, H., Zhang, J., & Smith, J. W. (2008). Probabilistic judgment on a coarser scale. *Cognitive Systems Research, 9*(3), 161–172. https://doi.org/10.1016/j.cogsys.2007.03.001

Varey, C. A., Mellers, B. A., & Birnbaum, M. H. (1990). Judgments of proportions. *Journal of Experimental Psychology: Human Perception and Performance, 16*(3), 613–625. https://doi.org/10.1037/0096-1523.16.3.613

Verkooijen, K. T., Stok, F. M., & Mollen, S. (2015). The power of regression to the mean: A social norm study revisited. *European Journal of Social Psychology, 45*(4), 417–425. https://doi.org/10.1002/ejsp.2111

Viscusi, W. K., & Hakes, J. (2003). Risk ratings that do not measure probabilities. *Journal of Risk Research, 6*(1), 23–43. https://doi.org/10.1080/1366987032000047789

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General, 115*(4), 348–365. doi1.10.1037//0096-3445.115.4.348

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39*(2), 176–190. https://doi.org/10.1287/mnsc.39.2.176

Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31*(2), 135–138. https://doi.org/10.3758/BF03334162

Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review, 10*(1), 43–62. https://doi.org/10.1017/S0269888900007256

Wallsten, T. S., & González-Vallejo, C. (1994). Statement verification: A stochastic model of judgment and response. *Psychological Review, 101*(3), 490–504. https://doi.org/10.1037/0033-295X.101.3.490

Weaver, B., & Wuensch, K. L. (2013). SPSS and SAS programs for comparing Pearson correlations and OLS regression coefficients. *Behavior Research Methods, 45*(3), 880–895. https://doi.org/10.3758/s13428-012-0289-7

Wedell, D. H. (1990). Methods for determining the locus of context effects in judgment. In J.-P. Caverni, J.-M. Fabre, & M. Gonzalez (Eds.), *Cognitive biases*. (pp. 285-302). Amsterdam, Netherlands: North-Holland. https://doi.org/10.1016/S0166-4115(08)61329-1

Wedell, D. H., & Parducci, A. (1988). The category effect in social judgment: Experimental ratings of happiness. *Journal of Personality and Social Psychology, 55*(3), 341–356. https://doi.org/10.1037/0022-3514.55.3.341

Wedell, D. H., & Parducci, A. (2000). Social comparisons: Lessons from basic research on judgment. In J. Suls & L. Wheeler (Eds.), *Handbook of social comparison: Theory and research* (pp. 223–251). Dordrecht, Netherlands: Kluwer Academic. https://doi.org/10.1007/978-1-4615-4237-7_12

Wedell, D. H., Parducci, A., & Lane, M. (1990). Reducing the dependence of clinical judgment on the immediate context: Effects of number of categories and type of anchors. *Journal of Personality and Social Psychology, 58*(2), 319–329. https://doi.org/10.1037/0022-3514.58.2.319

Weinstein, N. D., & Diefenbach, M. A. (1997). Percentage and verbal category measures of risk likelihood. *Health Education Research, 12*(1), 139–141. https://doi.org/10.1093/her/12.1.139

Weinstein, N. D., Kwitel, A., McCaul, K. D., Magnan, R. E., Gerrard, M., & Gibbons, F. X. (2007). Risk perceptions: Assessment and relationship to influenza vaccination. *Health Psychology, 26*(2), 146–151. https://doi.org/10.1037/0278-6133.26.2.146

Whalen, J., Gallistel, C. R., & Gelman, R. (1999). Nonverbal counting in humans: The psychophysics of number representation. *Psychological Science, 10*(2), 130–137. https://doi.org/10.1111/1467-9280.00120

Windschitl, P. D. (2002). Judging the accuracy of a likelihood judgment: The case of smoking risk. *Journal of Behavioral Decision Making, 15*(1), 19–35. https://doi.org/10.1002/bdm.401

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364. https://doi.org/10.1037//1076-898X.2.4.343

Witteman, C., & Renooij, S. (2003). Evaluation of a verbal–numerical probability scale. *International Journal of Approximate Reasoning, 33*, 117–131. https://doi.org/10.1016/S0888-613X(02)00151-2

Witteman, C., Renooij, S., & Koele, P. (2007). Medicine in words and numbers: A cross-sectional survey comparing probability assessment scales. *BMC Medical Informatics and Decision Making, 7*(13). https://doi.org/10.1186/1472-6947-7-13

Woloshin, S., Schwartz, L. M., Black, W. C., & Welch, H. G. (1999). Women's perceptions of breast cancer risk: How you ask matters. *Medical Decision Making, 19*(3), 221–229. https://doi.org/10.1177/0272989X9901900301

Woloshin, S., Schwartz, L. M., Byram, S., Fischhoff, B., & Welch, H. G. (2000). A new scale for assessing perceptions of chance: A validation study. *Medical Decision Making, 20*(3), 298–307. https://doi.org/10.1177/0272989X0002000306

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency processing and cognition* (pp. 21–36). Oxford, England: University Press. https://doi.org/10.1093/acprof:oso/9780198508632.003.0002

Appendix A: Stimulus Material

Option names		Outcomes	
Female	Male	Adverse drug reactions	Crimes (Experiment 1 only)
Anna	Andreas	Ausschlag [rash]	Diebstahl [theft]
Ella	Christoph	Erbrechen [vomiting]	Mord [murder]
Karin	Fabian	Fieber [fever]	Raub [robbery]
Katja	Felix	Kopfschmerz [headache]	Schlägerei [brawl]
Klara	Florian	Schwindel [dizziness]	Totschlag [manslaughter]
Laura	Julian	Unruhe [restlessness]	Überfall [mugging]
Lea	Marcel		
Lena	Martin		
Lisa	Matthias		
Lydia	Moritz		
Maria	Peter		
Nadin	Philipp		
Nina	Robert		
Sonja	Stefan		
Sophie	Thomas		

Appendix B: Frequency Judgments in Experiment 2

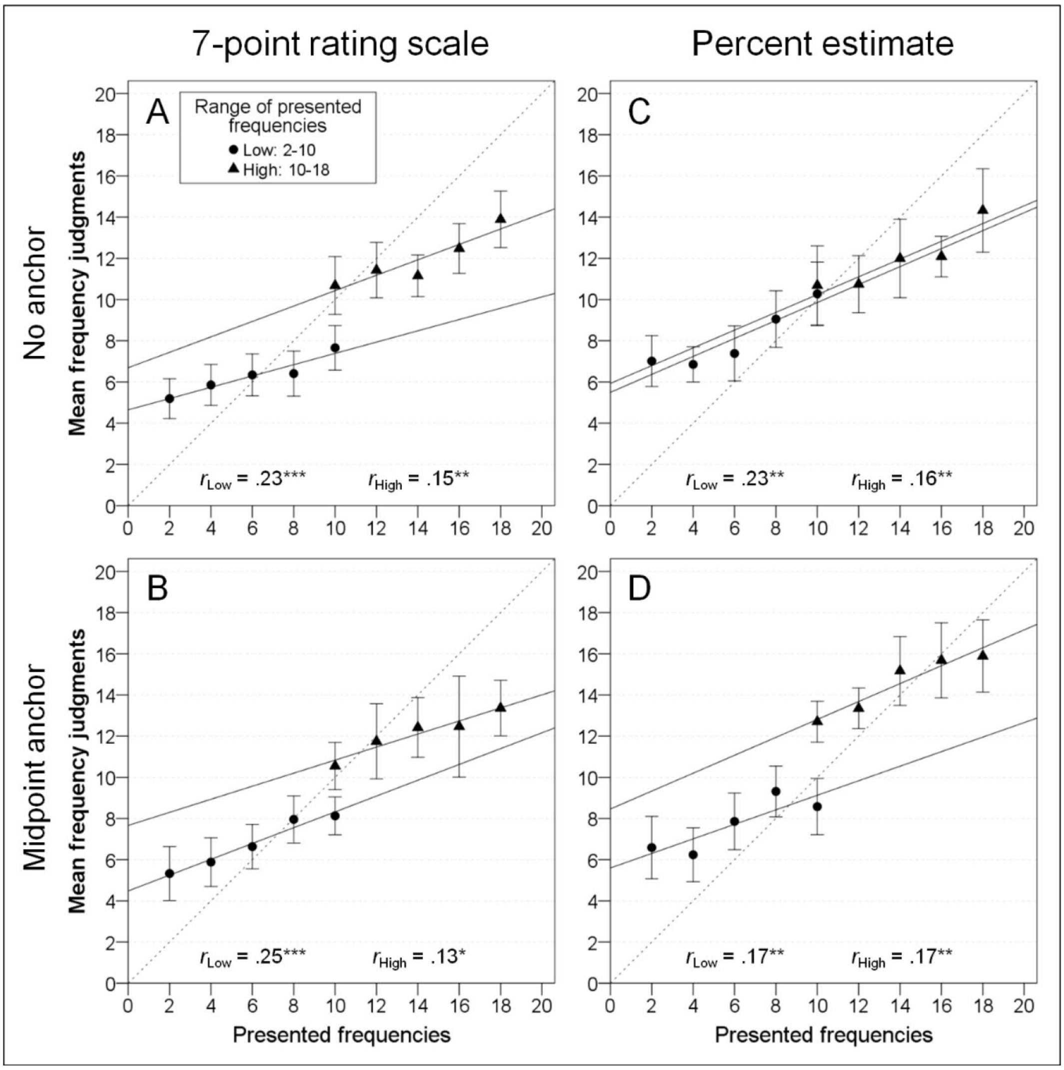

Figure B1. Mean frequency judgments as a function of presented frequencies (A: n = 28, B: n = 25, C: n = 27, D: n = 25). Solid lines represent the best linear fit. Correlation coefficients indicate sensitivity at the aggregate score level. Error bars = 95% within-subjects CIs. *p < .05. **p < .005. ***p < .001.

Differences as a function of probability scale formats

Figure B1 indicates that frequency judgments differed systematically as a function of the formats used probability judgments. Therefore, we calculated two separate mixed design ANOVAs – one for each encoding condition – with the scale format as between-subjects factor and the ten objective frequencies as within-subjects factor. In both cases Mauchly's test indicated a violation of the assumption of sphericity, sequential encoding: $\chi^2(44)$ = 203.65,

$p < .001$; list--wise: $\chi^2(44) = 186.26$, $p < .001$. We report Greenhouse-Geisser corrected tests, sequential: $\varepsilon = .52$, list-wise: $\varepsilon = .49$.

In the condition without an anchor we found a significant effect of the different frequencies indicating, of course, only that higher objective frequencies were judged to be higher, $F(4.66, 247.10) = 36.18$, $p < .001$, $\eta_p^2 = .41$. There was neither a main effect of the scale format on frequency judgments nor an interaction, $F\text{s} \leq 1.58$, $p\text{s} \geq .170$, $\eta_p^2\text{s} \leq .03$.

In the anchor condition we observed the same effect of objective frequencies on frequency judgments, $F(4.40, 211.19) = 47.75$, $p < .001$, $\eta_p^2 = .50$. However, we also found that subjects who used the percent format in their probability judgments, judged frequencies to be higher than those subjects using the rating scale, $F(1, 48) = 4.58$, $p = .037$, $\eta_p^2 = .09$. The interaction was not significant, $F < 1$.

Sensitivity

A mixed-design ANOVA found individual sensitivity to be somewhat lower for higher frequencies (Figure B2), though this effect was not qualified as significant, $F(1, 99) = 3.55$, $p = .062$, $\eta_p^2 = .04$. Individual sensitivity did not differ as a function of encoding mode or scale format, $F\text{s} < 1$.

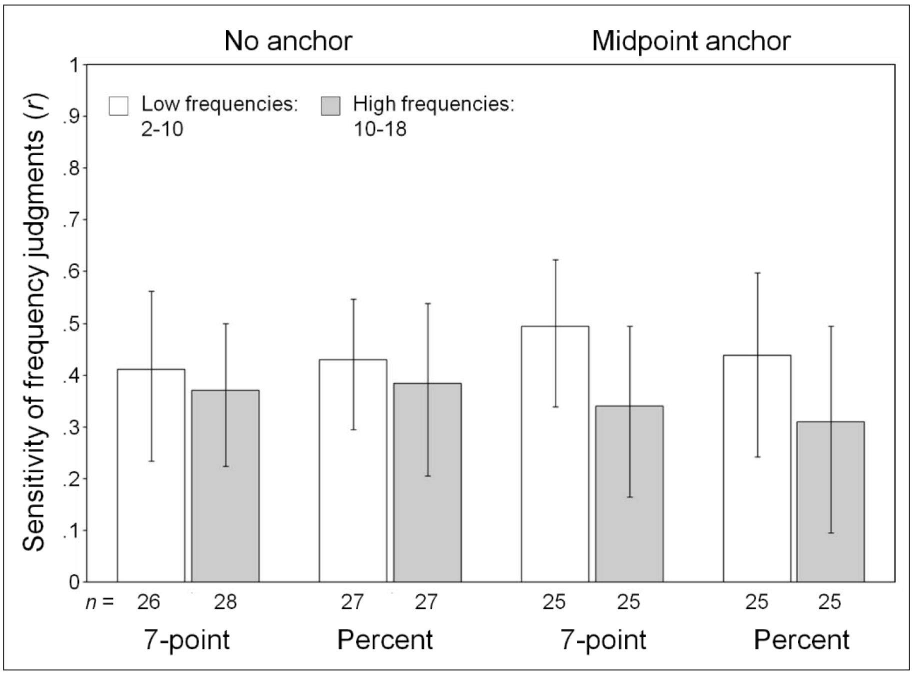

Figure B2. Sensitivity of frequency judgments at the individual score level of both scale formats for both ranges of probabilities and both encoding conditions. Error bars = 95% CIs.

At the aggregate score level, sensitivity did not differ as a function of encoding mode or scale format for either range of frequencies, low: $Q = 1.18$, $df = 3$, $p = .759$, high: $Q = 0.34$, $df = 3$, $p = .953$, or between frequency ranges, $t\text{s} \leq 1.69$, $p\text{s} \geq .093$, $r\text{s} \leq .09$.

Error in judgment

Mean *CV*s were: 7-Point, no anchor = .552, anchor = .590; percent, no anchor = .636, anchor = .545. *CV*s did not differ between conditions, $t\text{s} \leq |1.03|$, $p\text{s} \geq .304$, $r\text{s} \leq .08$.

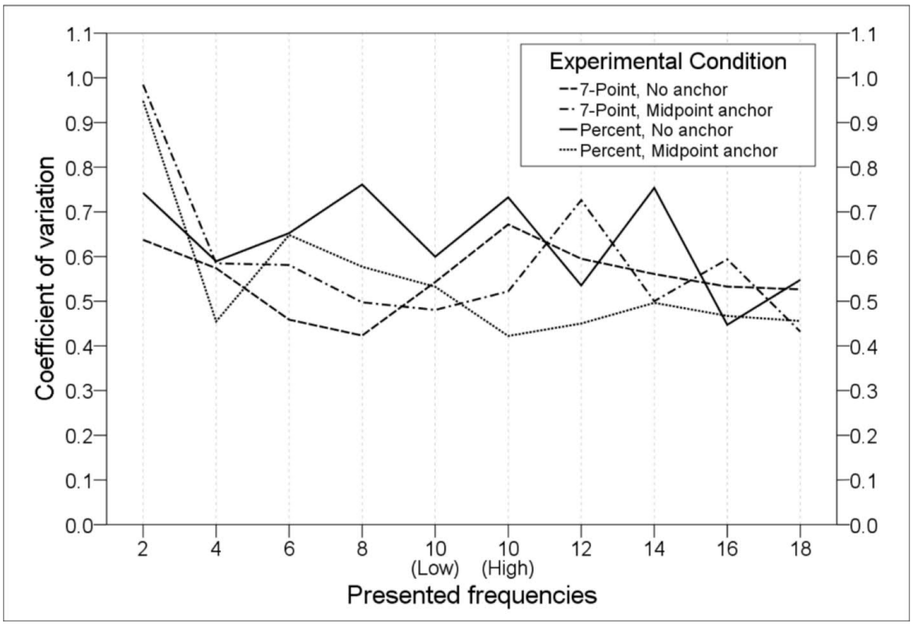

Figure B3. Coefficients of variation as variability profiles over all frequency judgments for all between-subjects conditions.

Appendix C: Frequency Judgments in Experiment 3

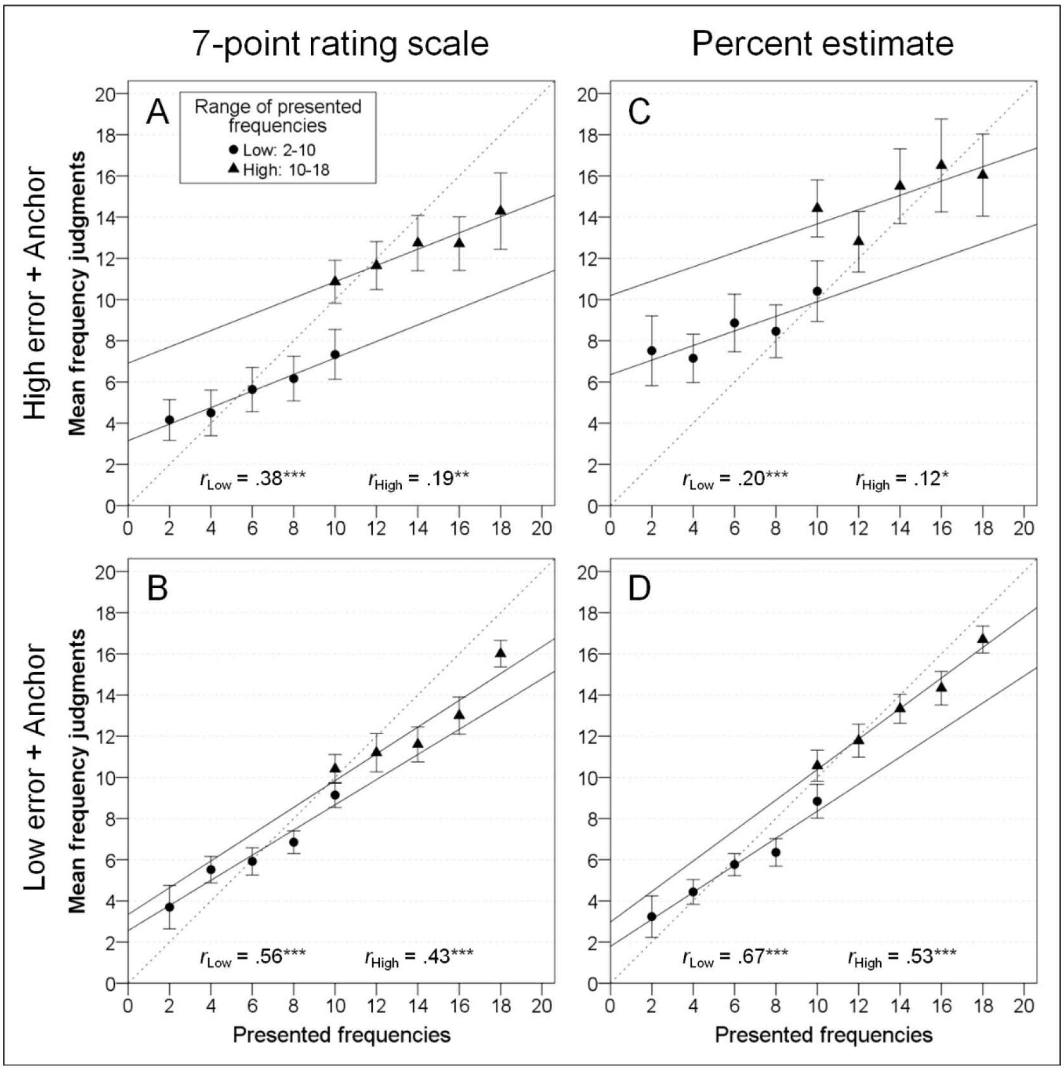

Figure C1. Mean frequency judgments as a function of presented frequencies (A: $n = 22$, B: $n = 24$, C: $n = 21$, D: $n = 25$). Solid lines represent the best linear fit. Correlation coefficients indicate sensitivity at the aggregate score level. Error bars = 95% within-subjects CIs. $^{*}p < .05$. $^{**}p < .005$. $^{***}p < .001$.

Differences as a function of probability scale formats

Figures 15 and C1 indicate that frequency judgments in the high-error condition differed systematically as a function of probability scale formats. We calculated mixed design ANOVAs for each encoding condition with the scale format as between-subjects factor and the ten objective frequencies as within-subjects factor. In both cases Mauchly's test indicated a violation of the assumption of sphericity, high-error: $\chi^2(44) = 183.48$, $p < .001$; low:

$\chi^2(44) = 129.32$, $p < .001$. We report Greenhouse-Geisser corrected tests, high-error: $\varepsilon = .49$; low: $\varepsilon = .53$.

Besides the effect of objective frequencies, $F(4.39, 179.86) = 58.97$, $p < .001$, $\eta_p^2 = .59$, when error was high subjects in the percent group estimated frequencies to be significantly higher than those using the rating scale, $F(1, 41) = 10.60$, $p = .002$, $\eta_p^2 = .21$.

In the low-error condition we observed the same effect of objective frequencies on frequency judgments, $F(4.77, 244.15) = 258.39$, $p < .001$, $\eta_p^2 = .85$ as well as a small but negligible interaction effect, $F(4.77, 244.15) = 2.86$, $p = .017$, $\eta_p^2 = .06$. However, frequency estimates did not vary as a function of the scale format used for probability judgments, $F < 1$.

Sensitivity

We recoded perfect sensitivity in 18 out of 527 (3.42%) correlations. A mixed-design ANOVA found that subjects were more sensitive to frequencies in the low-error condition as compared to the high-error condition, $F(1, 87) = 55.76$, $p < .001$, $\eta_p^2 = .39$. Additionally, subjects were less sensitive to high range frequencies than to low range frequencies, $F(1, 87) = 11.34$, $p = .001$, $\eta_p^2 = .12$. There was no difference between scale formats, nor were there any significant interactions, all $F < 1$.

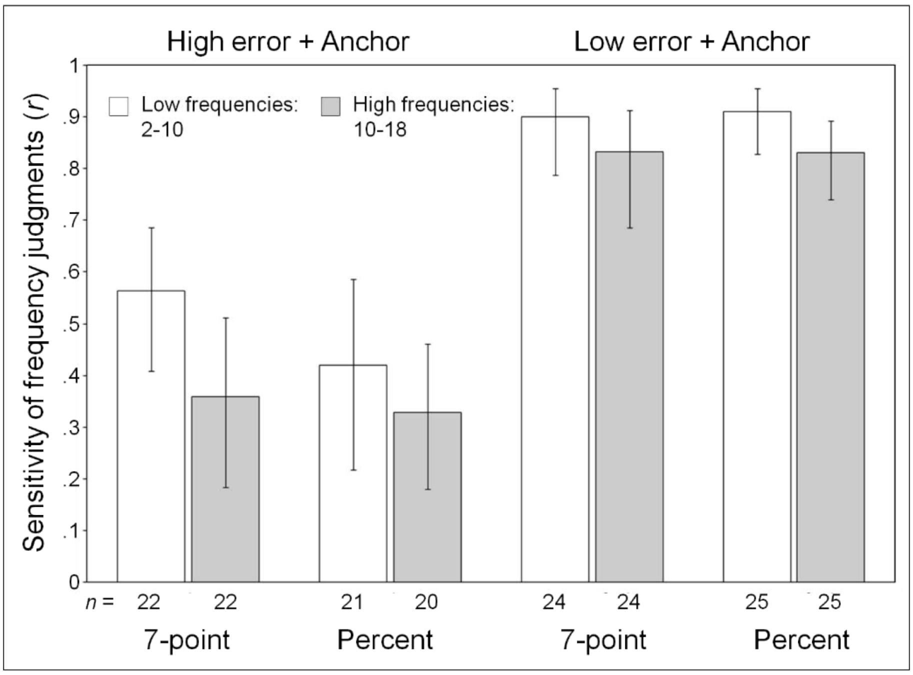

Figure C2. Sensitivity of frequency judgments at the individual score level of both scale formats for both ranges of probabilities and both encoding conditions. Error bars = 95% CIs.

At the aggregate level, the omnibus *Q*-test also confirmed significant differences between sensitivities for both frequency ranges, low: $Q = 69.93$, $df = 3$, $p < .001$; high: $Q = 47.67$, $df = 3$, $p < .001$. Single comparisons using Fisher's *z*-test revealed that aggregate sensitivity was significantly higher in the low-error condition within as well as across probability scale conditions, low range: *z*s ≥ |2.87|, *p*s < .004; high: *z*s ≥ |3.53|, *p*s < .001.

Additionally, for low range frequencies, aggregate sensitivity differed as a function of the scale format. In the high-error encoding, frequency estimates by subjects who used the rating scale were more sensitive than those by subjects using the percent format, $z = 2.50$, $p = .012$. In the high-error condition, this trend was reversed, $z = -2.49$, $p = .013$. However, assuming a Bonferroni corrected significance criterion of .008, neither effect would be qualified as significant. In the high range, aggregate sensitivity did not differ between scale formats within encoding conditions, *z*s ≤ |1.63|, *p*s ≥ .104.

Generally, frequency estimates were more sensitive to lower frequencies then to higher ones, *t*s ≥ 2.48, *p*s ≤ .014, *r*s ≥ .13, except for list-wise-percent condition, $t = 1.01$, $p = .315$, $r = .06$.

Error in judgment

Mean *CV*s were: 7-point rating scale, high-error = 0.483, low = 0.383; percent format, high-error = 0.548, low = 0.332. Within encoding conditions, *CV*s did not differ between scale formats, *t*s ≤ |1.04|, *p*s ≥ .302. Generally, frequency estimates showed less variation in the low-error condition, though differences were not as clear-cut as for probability judgments. Assuming a Bonferroni corrected significance criterion of .008, frequency estimates' variation only differed significantly in the percent condition, $t(136) = 3.12$, $p = .002$, $r = .26$, but not for subjects who had used the rating scale, $t(136) = 1.59$, $p = .115$, $r = .14$. Across scale formats comparisons: *t*s ≥ |2.28|, *p*s ≤ .024, *r*s ≥ .19.

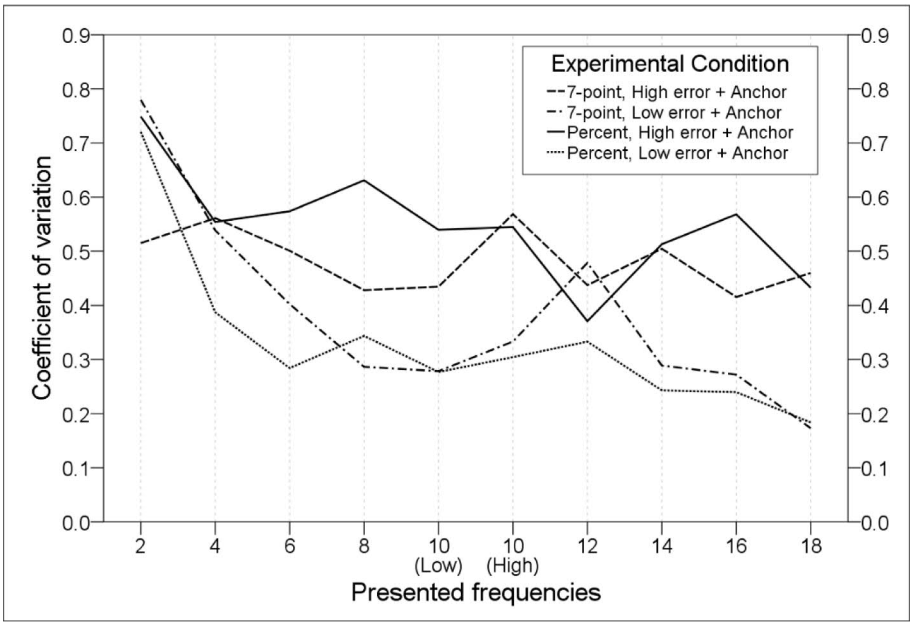

Figure C3. Coefficients of variation as variability profiles over all frequency judgments for all between-subjects conditions.

General Discussion

Summary of the Research Findings

The experiment in the first article (Haase, Betsch, & Renkewitz, 2015) revealed a stable effect of single-case narrative exemplars on the perceived risk and the subjective probability of vaccine adverse events even though a statistic had also been provided. Furthermore, the narratives also biased perceptions of severity. Manipulating the credibility of the statistical information had no effect. However, exemplars from a low-credibility source led to generally reduced estimates on all dependent variables but had no moderating effect on the occurrence of the bias.

The two experiments in Article 2 (Betsch, Haase, Renkewitz, & Schmid, 2015) found that the occurrence of the bias is task- and context dependent: Exemplars had the strongest effect on a broad risk measure while two scales for subjective probability mostly reflected variations in the statistical information. These ratios, however, varied as a function of the order in which the constructs were judged. Furthermore, different cognitive processes seem to underlie these judgments: Exemplars increased rather than decreased risk perceptions while the effect on probability estimates was symmetric. Additionally, numeracy had inconsistent moderating effects. Overall, these findings underscore the important conceptual distinction between risk perceptions and subjective probability. Crucially though, the effect of exemplars was driven by representations of likelihood.

The experiment in the third article (Haase, Renkewitz, & Betsch, 2013) showed that the performance of different measurement formats for subjective probability differs not only due to scale-inherent characteristics, such as their resolution or whether they are verbal or numeric in nature. The differences themselves differed as a function of the noise in the underlying representations. Greater error reduced differences in scale sensitivity but changed the way a low-resolution scale was used which may account in part for commonly observed context effects in between-subjects designs.

Finally, the three experiments in Article 4 (Haase & Betsch, 2016) revealed that high error can create apparent context effects even in a within-subjects design, though the responsible mechanism differs between scale formats. For a low-resolution verbal scale, the error changes the way the scale is used from a focus on mapping probabilities' positions on the probability continuum to differentiating them, effectively reducing the level of measurement to ordinal or even nominal. For a high-resolution numerical scale, high error may lead to classic regression fallacies which could also be interpreted as context effects. However, the percent format remains relatively consistent in how it captures the error in representations.

Discussion and Implications

The findings in Article 1 are in line with previous studies using the same research paradigm. Context variables—such as the emotionality of narratives (Betsch, Ulshöfer, Renkewitz, & Betsch, 2011), the credibility of their source (Haase et al., 2015), or the perceived personal susceptibility to experiencing the focal event (Haase, Schmid, & Betsch, 2016)—can have a main effect on risk and sometimes on probability judgments but do not moderate the bias created by single-case exemplars. Only numeracy, that is, the ability to perform operations with numbers and proportions has inconsistently affected its occurrence. Sometimes highly numerate subjects have been less susceptible to the influence of a small sample of exemplars. (Betsch et al., 2015; Haase et al., 2015; Haase et al., 2016 but see Betsch, Renkewitz, & Haase, 2013).

All these findings are in line with the notion that the encoding of frequency information is a fundamental, largely automatic and ultimately inevitable property of perception (Zacks & Hasher, 2002). People are highly sensitive to event frequencies and use them for probabilistic judgments. However, the translation of a frequency representation into a judgment of probability or risk involves other processes as the effects of numeracy in the presented inference tasks indicate. But also in pure estimation tasks, probability judgments are not simply mathematical transformations of encoded absolute frequencies as indicated, for example, by the differences between the error distributions of frequency and probability judgments that were based on the same stimuli in Article 4.

Articles 2–4 address how perceptions of probability that are based on relative frequencies are mapped onto different measurement formats under various conditions. There is one important difference to keep in mind though. In Article 2, the probabilistic information was manipulated in an applied context and subjects were faced with an inference task. The experiments in Articles 3 and 4, on the other hand, employed an almost psychophysical design in that the stimuli were not embedded in an elaborate scenario and the task was focused on estimation only. Consequently, the manipulations of the focal event's severity and domain had no effect. This distinction is crucial as the respective findings relate to different aspects of the scale formats' validity. I will first discuss the results reported in Articles 3 and 4 because the conclusions are rather clear-cut. I will then close with a discussion of Article 2 because its theoretical implications are broader and somewhat ambiguous.

The results from the studies in Articles 3 and 4 indicate that all subjects used their respective scale format to express the same construct, that is, their perception of the objective probability which they had encoded as a relative frequency. Thus, in the sense that the construct of interest is frequentist probability, all formats can be considered valid given the right circumstances (i.e., the research paradigm). However, the formats also provided clearly defined bounds at the bottom and top end of the scale indicating to the subjects that the construct is not just probability but *absolute* probability, in contrast to *relative* probability. Mapping the probability continuum on to a rating scale with a limited number of categories means, of course, that similar probabilities need to be placed in the same category. When subjects have precise knowledge, this is what they do (Article 3, graphical encoding). However, with increasing error in the representations the focus shifts from precise mapping to differentiation (Articles 3 & 4), which is in line with a preference for gist-based reasoning as posited by fuzzy-trace theory (Reyna, 2012). From the subjects' point of view it may seem reasonable to at least successfully convey certain knowledge (i.e., ordinal relations) than to potentially fail at conveying vague impressions (absolute probabilities). For the researcher who uses a low-resolution scale, however, this means that the construct she measures changes as a function of noise. While this may be interpreted as measurement error, it is not an issue of scale reliability—the effects are systematic and repeatable. It rather demonstrates that verbal probability measures have unstable validity.

Thus, whenever a research question asks for a comparison of probability estimates, a low-resolution verbal rating scale is not a valid instrument. This does not only apply to situations where judgments are to be compared to objective probabilities but to any comparison of estimates between different contexts (e.g., experimental conditions), between- as well as within-subjects. Situations like this require a high-resolution numeric instrument. By prompting concerns about accuracy (Windschitl & Wells, 1996) numbers keep the answer scale fixed to the probability continuum, that is, the construct remains absolute probability. More importantly, they allow and force subjects to make use of the high resolution in the first place. Recall, for example, the visual analog scale in Article 3. It provided the same resolution as the two numeric instruments but no numeric feedback. Consequently, estimates were very similar to the rating scales in terms of sensitivity and context dependency. This is in line with previous research reporting that in an applied context (i.e., imperfect knowledge) people typically do not use more than seven categories for judgments and a higher resolution does not improve scale sensitivity (Diefenbach, Weinstein, & O'Reilly, 1993; McKelvie, 1978; Weinstein & Diefenbach, 1997). The strength of the numeric format, however, does not lie in its superiority to capture monotonic change but to capture random error by asking for precise estimates even when representations are vague. The error is part of the representation that one is trying to assess and being sensitive to it makes numerical probability estimates relate consistently to the theoretical probability continuum allowing for comparisons between judgment contexts. Again, a finding which might be construed to indicate a lack of reliability reveals in fact, I would argue, a property of the scale format which speaks to its validity.

From the reasoning thus far one could draw the conclusion that subjective probability should simply always be assessed numerically, preferably in the form of percent estimates. Unfortunately, matters get a lot less clear once we leave the confines of psychophysical research and turn to the applied context. When there is good reason to believe that subjects can sample event frequencies in such a context, the previous argument, of course, still applies. However, we can easily identify a priori circumstances where asking for a percent estimate is counterproductive or does not even make sense. For instance, when people have no information at all (e.g., asking a Franconian farmer about the chances of rain in Tokyo?)[1] or when the event is fundamentally unique (e.g., "How likely is it that Margarete loves me?")[2]

Of course, we still need to communicate about situations like this. After all, humanity had to deal with uncertain prospects long before Blaise Pascal developed probability theory (Zimmer, 1983). But, when we do, we are not referring to objective probabilities but rather to the strength of our beliefs, that is, subjectivist probability. Verbal expressions seem intuitive and natural for this purpose and pose no problem in clear-cut cases like the two examples where comparisons to any objective criterion are dismissed from the outset. However, when there is a normative standard that is defined by an objectivist interpretation of probability—as is the case in the experiments in Articles 1 and 2—the downside of verbal expressions becomes quickly apparent: Very often we cannot be certain about what construct we are referring to, or, more specifically, whether at all or to what extend a belief in an objective probability enters into a verbal probability judgment. For instance, in Article 2 we only observed an effect of the single-case exemplars (i.e., a purely probabilistic manipulation) on the verbal rating scale when we either controlled for perceptions of severity or when subjects had judged the broader risk construct first. Both findings indicate that in an applied context verbal probability judgments might best be understood from a subjectivist perspective.

---

[1] This example assumes that the average Franconian farmer has not spent an extended period in Tokyo to collect data on the frequency of rain.

[2] One could, of course, argue that there are many instances of girls like Margarete and boys like me. Then the question becomes an issue of finding the right reference classes to which we belong. Similarly, one could take Margarete not as a whole but rather analyze specific behavioral patterns and character traits that have previously coincided with loving me (which is the basic working principle of online dating websites).

Incidentally, this is the reason why it was crucial that in Articles 3 and 4 the verbal judgments were not affected by the focal events' severity and domain. Thus, we could be certain that subjects tried to express their perceptions of objective probability and we could interpret that the scale format is also inconsistent regarding the structure of the construct (i.e., absolute vs. relative judgments) and not just its content (objectivist vs. subjectivist interpretation). Note that the subjectivist view theoretically does not include other aspects of an uncertain prospect in its interpretation of probability. However, how to disentangle subjectivist probabilities from values has been the central challenge for this position with some arguing that it is neither possible nor necessary (Nau, 2002).

The argument about the uncertainty over what construct we are actually assessing with verbal probability expressions applies by extension to risk judgments. The key difference is that most theoretical perspectives are explicit about including other constructs than objective probability in their definitions of risk. Nonetheless, the effect of the relative frequency of exemplars on risk judgments changes when subjects first provide probability estimates or when we control for severity. Given their flexibility and sensitivity to context manipulations, it is then not surprising that verbal likelihood formats and risk judgments have consistently been shown to be superior in the prediction of behavioral intentions and actual behavior than numeric probability estimates (Article 2; Baghal, 2011; Weinstein et al., 2007; Windschitl & Wells, 1996). Behavior, of course, has more antecedent than just the belief in the objective probability of behavioral outcomes.

However, when we assess biases we need to apply a normative standard and in most research on biased probabilistic reasoning this standard is rooted in a frequentist interpretation of probability including the bias of single-case exemplars which hinges on the law of large numbers. In Article 2 we observed a large effect on a broad risk measure (both experiments) and only a small (Experiment 1) or even no effect (Experiment 2) on a percent estimate. In Articles 3 and 4 we found the percent format to be highly sensitive to frequentist probability. Taken together, we could then conclude that single-case exemplars increase risk perceptions through other constituents of the construct. For example, repeatedly reading about vaccine adverse events might make them seem more severe and thus make vaccinations seem more risky. If this were the case, one might still argue that the reasoning is biased if one assumed that perceptions of severity should not be changed by a few repeated encounters with an event. However, this bias does not violate probability theory. Given this reasoning, I close with pointing out what I believe to be the most striking finding of this dissertation. Subjects extracted a representation of likelihood (i.e., relative frequency) from a small sample of single-case exemplars (Article 2, Experiment 2). This representation had no effect on an instrument shown to be sensitive to relative frequencies (i.e., percent estimates) but a large effect on the inherently multi-dimensional risk measure. Thus, the relationship between representations of subjective probability and perceptions of risk is not yet fully understood.

References

Baghal, T. (2011). The measurement of risk perceptions: The case of smoking. *Journal of Risk Research, 14*(3), 351–364. https://doi.org/10.1080/13669877.2010.541559

Betsch, C., Haase, N., Renkewitz, F., & Schmid, P. (2015). The narrative bias revisited: What drives the biasing influence of narrative information on risk perceptions? *Judgment and Decision Making, 10*(3), 241–264. Retrieved from http://journal.sjdm.org/14/141206a/jdm141206a.pdf

Betsch, C., Renkewitz, F., & Haase, N. (2013). Effect of narrative reports about vaccine adverse events and bias-awareness disclaimers on vaccine decisions: A simulation of an online patient social network. *Medical Decision Making, 33*(1), 14–25. https://doi.org/10.1177/0272989X12452342

Betsch, C., Ulshöfer, C., Renkewitz, F., & Betsch, T. (2011). The influence of narrative v. statistical information on perceiving vaccination risks. *Medical Decision Making, 31*(5), 742–753. https://doi.org/10.1177/0272989X11400419

Diefenbach, M. A., Weinstein, N. D., & O'Reilly, J. (1993). Scales for assessing perceptions of health hazard susceptibility. *Health Education Research, 8*(2), 181–192. https://doi.org/10.1093/her/8.2.181

Haase, N., Betsch, C., & Renkewitz, F. (2015). Source credibility and the biasing effect of narrative information on the perception of vaccination risks. *Journal of Health Communication, 20*(8), 920–929. https://doi.org/10.1080/10810730.2015.1018605

Haase, N., & Betsch, T. (2016). *Self-report measures of subjective probability: Error and anchor effects.* Manuscript in preparation.

Haase, N., Renkewitz, F., & Betsch, C. (2013). The measurement of subjective probability: Evaluating the sensitivity and accuracy of various scales. *Risk Analysis, 33*(10), 1812–1828. https://doi.org/10.1111/risa.12025

Haase, N., Schmid, P., & Betsch, C. (2016). *Disease risk, personal susceptibility, and the narrative bias in vaccination risk perception.* Manuscript in preparation.

McKelvie, S. J. (1978). Graphic rating scales – How many categories? *British Journal of Psychology, 69*(2), 185–202. https://doi.org/10.1111/j.2044-8295.1978.tb01647.x

Nau, R. (2002). de Finetti was right: Probability does not exist. Theory and Decision, 51(2), 89–124. https://doi.org/10.1023/A:1015525808214

Reyna, V. F. (2012). A new intuitionism: Meaning, memory, and development in Fuzzy-Trace Theory. *Judgment and Decision Making, 7*(3), 332–359. Retrieved from http://journal.sjdm.org/11/111031/jdm111031.pdf

Weinstein, N. D., & Diefenbach, M. A. (1997). Percentage and verbal category measures of risk likelihood. *Health Education Research, 12*(1), 139–141. https://doi.org/10.1093/her/12.1.139

Weinstein, N. D., Kwitel, A., McCaul, K. D., Magnan, R. E., Gerrard, M., & Gibbons, F. X. (2007). Risk perceptions: Assessment and relationship to influenza vaccination. *Health Psychology, 26*(2), 146–151. https://doi.org/10.1037/0278-6133.26.2.146

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied, 2*(4), 343–364. https://doi.org/10.1037//1076-898X.2.4.343

Zacks, R. T., & Hasher, L. (2002). Frequency processing: A twenty-five year perspective. In P. Sedlmeier & T. Betsch (Eds.), *Etc. Frequency processing and cognition* (pp. 21–36). Oxford, England: University Press. https://doi.org/10.1093/acprof:oso/9780198508632.003.0002

Zimmer, A. C. (1983). Verbal vs. numerical processing of subjective probabilities. In R. Scholz (Ed.), *Decision making under uncertainty* (pp. 159–182). Amsterdam, Netherlands: North-Holland. https://doi.org/10.1016/S0166-4115(08)62198-6

Acknowledgments

For professional and personal support, I thank the following in *randomized* order:

Jenny Brenke

Tilmann Betsch

Cornelia Betsch

Johannes Ritter

Frank Renkewitz

More than anyone and more than everyone combined I thank Anna Lang.

In memory of my mother Dr. Inga Haase-Becher.

Dedicated to the smallest detail.

Curriculum Vitae

Education

2005	Freie Universität Berlin, Germany Diploma in Psychology Overall grade: „very good“ Thesis: *The Implicit Associationtest as a measure of psychopathy in a nonforensic population*
2002	John Jay College of Criminal Justice, City University of New York, New York City, USA Master of Arts in Forensic Psychology Grade Point Average: 3.9 (out of 4.0) Dean's List in both academic years

Employment

10/2016-Present	University of Konstanz, Germany Researcher
06/2010-09/2014	University of Erfurt, Germany Researcher
12/2008-03/2010	Otto-von-Guericke University Magdeburg, Germany Officer for Technology Transfer
02/2007-07/2008	Kulturaustausch – Journal for International Perspectives, Berlin, Germany Freelance Editor
12/2006-02/2007	Kulturaustausch – Journal for International Perspectives, Berlin, Germany Editorial Internship
01/2006-10/2006	Institute of Forensic Psychiatry, Charité University Medicine, Berlin, Germany Development of a research project on diagnostic tools for the assessment of risk factors in violent offenders (unpaid)