GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Die Übernutzung nicht valider Ratschläge:
Warum schlechte Ratschläge über Gebühr berücksichtigt werden

Dissertation

Zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

„Doctor rerum naturalium“

Der Georg-August-Universität Göttingen

vorgelegt von

Thomas Schultze

aus Darmstadt

Göttingen, 2010

Mitglieder des Betreuungsausschusses:

Referent: Prof. Dr. Stefan Schulz-Hardt

Koreferent: Prof. Dr. Michael Waldmann

Tag der mündlichen Prüfung: 04.11.2010

Gewidmet meinem

verstorbenen Großvater

Dr. Rudi Schultze

Danksagung

Meine Dissertation markiert den Abschluss einer Karrierestufe in meiner wissenschaftlichen Laufbahn, genauso wie das Ende eines Lebensabschnitts. Vielmehr aber ist sie das Ergebnis eines langen Prozesses, der von anfänglichen Gedankenspielen und theoretischen Überlegungen über Konzeptionen und Revisionen der erdachten Konzepte, die konkrete Planung und Durchführung der Experimente, die Auswertung und Deutung der Ergebnisse bis hin zur letztendlichen Verschriftlichung führte. Das erfolgreiche Gelingen dieses Prozesses war nur möglich, weil ich stets auf die Unterstützung zahlreicher Personen bauen konnte, die mir auf vielfältige Art und Weise zur Seite standen. Diesen Personen möchte ich hiermit meine tiefe Dankbarkeit ausdrücken.

Allen voran gilt mein Dank meinem Doktorvater, Freund und Mentor Stefan Schulz-Hardt. Im Einzelnen darzulegen, wie viel ich ihm auch weit über meine Dissertation hinaus verdanke, würde wohl ein ganzes Kapitel meiner Dissertation in Anspruch nehmen. Stefan war mir zugleich Lehrer, Ratgeber, Kritiker und Vorbild. Er hat mich in vielerlei Hinsicht durch Höhen und Tiefen begleitet und wird das hoffentlich auch in Zukunft weiterhin tun.

Gleichermaßen möchte Michael Waldmann danken, sowohl dafür, dass er sich bereit erklärt hat, als Zweitgutachter meiner Arbeit zu fungieren, als auch für seine intensive fachliche Unterstützung und die hervorragende Zusammenarbeit in den vergangenen Jahren. Mein Dank gilt weiterhin meinem Freund und Kollegen Andreas Mojzisch, der mich überhaupt erst auf die Idee gebracht hat, mich mit dem Judge-Advisor-Paradigma zu befassen und dessen Anregungen und Rückmeldungen einen ganz wesentlichen Beitrag zur Entwicklung meines Forschungsprogramms geleistet haben. Danken möchte ich auch Margarete Boos, York Hagmayer und Hannes Rakoczy, die sich trotz voller Terminpläne die Zeit genommen haben, das Prüfungskomitee für meine Dissertation zu vervollständigen.

Ich möchte ferner all denjenigen Personen danken, ohne deren Hilfe die Durchführung der Experimente, auf denen meine Dissertation basiert, nicht möglich gewesen wäre. Mein Dank gebührt Simon Palmer für die ursprüngliche Programmierung des ersten Experiments und Thorsten Albrecht für seine Unterstützung, als ich mit der Modifikation des Experimentes einmal nicht weiterkam. Ferner danke ich den fleißigen Hilfskräften und Diplomanden, die als Versuchsleiter unermüdlich Probanden akquirierten und testeten, nämlich Cari-

na Cohrs, Ricarda Otto, Penninah Jones, Anette Opielka, Alex Stern, Johanna Theyson und Christoph Ehrling, sowie den vielen Versuchspersonen, die das nicht immer spannende Experiment über sich haben ergehen lassen.

Spezieller Dank gebührt den Mitgliedern meiner Kochgruppe, Jakob und Katharina Bierwagen, Nadira Faulmüller, Lars Kasper, Frederik Köpper, Johannes Schmidt-Hieber, Felicitas Sedlmair und Nora Wender für die sozial-emotionale Unterstützung vor allem in den Zeiten, in denen meine Motivation für das Anfertigen einer Dissertation sich auf suboptimalem Niveau bewegte. Johannes Schmidt-Hieber möchte ich dabei außerdem noch für die Zusammenarbeit bei der Erstellung des normativen Modells der Ratgebergewichtung danken, das ich am Ende der Dissertation kurz beschreibe.

Zum Abschluss möchte ich noch zwei Personen meinen ganz besonderen Dank aussprechen, weil ich ohne ihren Einfluss nie in die Situation gekommen wäre, im Fach Psychologie zu promovieren. Zum einen gilt daher mein Dank meiner Mutter und Kollegin Marianne Ponto-Schultze, die in mir die Begeisterung für mein Fach weckte. Zum anderen gilt mein Dank gleich in zweifacher Hinsicht Dieter Heyer, nämlich einerseits dafür, dass ich durch ihn meine Leidenschaft für die Wissenschaft entdeckte, zum anderen deshalb, weil erst durch sein Wirken der Kontakt zu Stefan Schulz-Hardt entstand.

Inhaltsverzeichnis

1. Einleitung

Urteilen und Entscheiden gehört zu den integralen Bestandteilen unseres alltäglichen Lebens. Jeden Tag treffen wir eine Vielzahl von Entscheidungen und geben zahlreiche Urteile oder Einschätzungen ab. Einige dieser Urteile oder Entscheidungen wie die Wahl des Mittagsmenus oder die Einschätzung, wie viel Rotwein für den geselligen Abend benötigt wird, mögen trivial erscheinen und haben, wenn überhaupt, nur wenig bedeutsame Konsequenzen. Andere Urteile und Entscheidungen hingegen können weitreichende Folgen haben, sowohl für diejenige Person, die das Urteil oder die Entscheidung fällt, als auch andere Personen oder Personengruppen. In diese Kategorie fallen beispielsweise Finanzprognosen, deren Akkuratheit über Reichtum oder Ruin der Investoren entscheiden kann, oder politische Entscheidungen, die Implikationen für Millionen von Personen haben können.

Wegen ihrer offensichtlichen Relevanz sind Urteile und Entscheidungen zu einem fachübergreifenden Forschungsgegenstand geworden. Neben der Philosophie und den Wirtschaftswissenschaften hat insbesondere die Psychologie in den vergangenen Jahrzehnten intensiv erforscht, wie Individuen Entscheidungen treffen oder Urteile fällen, wann oder warum diese Urteile oder Entscheidungen gemessen an einem normativen Maßstab gut oder schlecht ausfallen, und welche Heuristiken Menschen anwenden, um Urteile oder Entscheidungen zu fällen (siehe z.B. Bazerman & Neale, 1992; Gilovich, Griffin & Kahneman, 2002, Kahneman & Tversky, 2000; Stanovich, 1999). Ungeachtet der zentralen Bedeutung der Erkenntnisse, die bezüglich des Urteilens und Entscheidens gewonnen wurden, bezieht sich der Großteil der Forschung auf rein individuelle Urteils- und Entscheidungsprozesse. Das bedeutet, dass in der Regel jeglicher soziale Kontext nicht nur vernachlässigt, sondern gezielt ausgeblendet wird.

Menschliches Urteilen und Entscheiden findet jedoch in den seltensten Fällen außerhalb sozialer Kontexte statt. Stattdessen erhält eine Person, die ein Urteil oder eine Entscheidung fällen soll – ob gewünscht oder nicht – häufig Ratschläge in Form von Einschätzungen, Meinungen oder Empfehlungen anderer Personen. Häufig werden Ratschläge sogar aktiv eingeholt, und mitunter teuer bezahlt, weil man sich erhofft, durch die Ratschläge eine bessere Entscheidung oder ein akkurateres Urteil zu fällen. In der Tat zeigt die bisherige For-

schung zum Umgang mit Ratschlägen, dass deren Berücksichtigung einen förderlichen Effekt auf die Urteils- und Entscheidungsqualität hat (Bonaccio & Dalal, 2006, Yaniv 2004a).

Gleichermaßen kommt die Forschung aber auch zu dem Ergebnis, dass Ratschläge in der Regel nicht ausreichend berücksichtigt werden. Das Phänomen, dass diejenige Person, die ein Urteil oder eine Entscheidung treffen soll, einen Ratschlag weniger stark berücksichtigt als sie, basierend auf der objektiven Akkuratheit des Ratschlags, sollte, wird als *advice discounting* (Bonaccio & Dalal, 2006; Yaniv, 2004a, 2004b; Yaniv & Kleinberger, 2000) bezeichnet. Allerdings herrscht in der Literatur auch eine gewisse Unschärfe bei der Verwendung dieses Begriffs, da Yaniv (2004b) sowie Yaniv und Kleinberger (2000) sowohl dann von advice discounting sprechen, wenn ein Ratschlag nicht so stark gewichtet wird, wie es seine Qualität nahelegt, als auch dann, wenn ein Ratschlag weniger als vollständig befolgt wird. Im letzten Fall wird advice discounting mit dem Ausmaß gleichgesetzt, zu dem eine Person ihr eigenes Urteil zu Ungunsten eines Ratschlags beibehält. Um diese begriffliche Unschärfe im Folgenden zu vermeiden, werde ich im Folgenden immer dann von advice discounting sprechen, wenn ein Ratschlag nicht vollständig übernommen wird, und zwar unabhängig davon, wie stark der entsprechende Ratschlag gewichtet werden sollte. Wenn dagegen ein Ratschlag weniger stark gewichtet wird, als seine Qualität nahelegt, werde ich dies stattdessen als *Unternutzung* des Ratschlags bezeichnen.

Die Unternutzung von Ratschlägen, die als eine Form der Irrationalität des menschlichen Urteilens und Entscheidens aufgefasst werden kann, stellt das wohl robusteste und am besten untersuchte Phänomen in der Forschung zur Nutzung von Ratschlägen dar (Bonaccio & Dalal, 2006; Yaniv 2004a). In der Summe entsteht dabei der Eindruck, dass Ratschläge generell diskontiert, also zu gering gewichtet werden. Dieser Schluss könnte jedoch vorschnell sein, da durch den einseitigen Fokus auf die Unternutzung von Ratschlägen die mögliche Kehrseite der Medaille bis dato vollkommen vernachlässigt wurde. Mit anderen Worten: Bisher wissen wir gar nicht, ob Menschen nicht vielleicht unter bestimmten Bedingungen Ratschläge auch zu stark gewichten.

Da die Konsequenzen einer solchen übermäßigen Nutzung von Ratschlägen, falls sie existiert, nicht minder nachteilig sein sollten als die Folgen der bereits wohlbekannten Unternutzung von Ratschlägen, soll es das Ziel meiner Dissertation sein, diese vernachlässigte Kehrseite der Medaille zu beleuchten. In einer Serie von fünf Experimenten untersuche ich,

ob auch das gegenteilige Phänomen auftreten kann, nämlich die systematische *Übernutzung* von Ratschlägen. Für den Fall, dass eine solche Übernutzung auftritt, sollen die Experimente ferner mögliche vermittelnde Mechanismen aufdecken, die Rückschlüsse auf die Prozesse zulassen, die der systematischen Übernutzung von Ratschlägen zugrunde liegen. Abschließend diskutiere ich Implikationen der Befunde für die weitere Forschung.

2. Theoretischer und empirischer Hintergrund

Die systematische Untersuchung des Umgangs mit Ratschlägen beim Urteilen und Entscheiden stellt ein relativ junges Forschungsfeld dar. Die erste empirische Arbeit, die explizit individuelles Entscheiden unter Berücksichtigung von Ratschlägen untersuchte, wurde vor weniger als 25 Jahren veröffentlicht (Brehmer & Hagafors, 1986); das heute gängige Forschungsparadigma, das als *Judge-Advisor-Paradigma* bezeichnet wird (vgl. Abschnitt 2.2), wurde sogar erst vor gut 15 Jahren in der Literatur eingeführt (Sniezek & Buckley, 1995). Dieser Tatsache ist es geschuldet, dass die Forschung zum Umgang mit Ratschlägen bisher noch weitgehend ohne spezifische theoretische Grundlagen operiert. Tatsächlich existiert bislang nur ein einziges Modell, das den Umgang mit Ratschlägen erklärt bzw. Vorhersagen darüber trifft, ob eine Person einen Ratschlag befolgt oder ignoriert (Jungermann, 1999, 2005). Besagtes Modell ist außerdem sehr spezifisch und daher für die Mehrzahl der Situationen, in denen die Nutzung von Ratschlägen untersucht wird, nicht anwendbar. Entsprechend ist die Forschung zur Nutzung von Ratschlägen bisher stark ergebnisorientiert und fokussiert vor allem auf den Grad der Nutzung von Ratschlägen, auf die Veränderung der Akkuratheit der Schätzungen durch die Berücksichtigung der Ratschläge sowie auf die Veränderung der subjektiven Sicherheit durch den Erhalt von Ratschlägen (Bonaccio & Dalal, 2006; Yaniv, 2004a).

Von besonderem Interesse für meine Dissertation ist dabei der erstgenannte und wohl am stärksten beforschte Aspekt, also der Grad der Nutzung von Ratschlägen. Ich werde daher im Folgenden, nach einer kurzen Begriffsklärung dessen, was ich unter Urteilen und Entscheiden verstehe (Abschnitt 2.1), sowie nach einer Beschreibung des in der Forschung zum Umgang mit Ratschlägen gängigen Forschungsparadigmas (Abschnitt 2.2) darlegen, welche Erkenntnisse die Forschung bisher bezüglich der Nutzung von Ratschlägen geliefert hat. Ich werde dann zunächst erläutern, welche gängigen Maße verwendet werden, um das

Ausmaß zu bestimmen, in dem Ratschläge berücksichtigt werden, welche Probleme mit den einzelnen Maßen verbunden sind und welches Maß aus welchen Gründen in meiner Dissertation Anwendung findet (Abschnitt 2.3). Im Anschluss werde ich das Phänomen der Unternutzung von Ratschlägen beschreiben und die zugehörigen Erklärungsansätze kurz skizzieren (Abschnitt 2.4). Danach erläutere ich, warum die Auffassung, Ratschläge würden generell zu wenig berücksichtigt, im besten Falle voreilig ist. Zu diesem Zweck stütze ich mich zum einen auf Teilbefunde aus existierenden Studien zum Umgang mit Ratschlägen, die als Hinweis auf eine mögliche Übernutzung von Ratschlägen aufgefasst werden können (Abschnitt 3.1); zum anderen ziehe ich Parallelen zu ähnlichen Phänomenen aus der Forschung, zum Beispiel zum Anker-Effekt (Tversky & Kahnemann, 1974), der als eine Art Übernutzung nicht valider Hinweisreize aufgefasst werden kann (Abschnitt 3.2). Aus der Zusammenführung dieser Argumentationsstränge konkretisiere ich dann meine Forschungsfragen, die ich dann im Rahmen von 5 Experimenten (Abschnitte 4 bis 8) beantworte.

2.1 Begriffsklärung: Urteilen und Entscheiden

Urteilen und Entscheiden sind Begriffe, die in der psychologischen Forschung häufig in einem Atemzug genannt werden; obwohl sie zwei unterscheidbare Prozesse bezeichnen, werden diese beiden Begriffe aber mindestens ebenso häufig synonym oder austauschbar verwendet. Daher scheint es mir angebracht, die beiden Begriffe abgrenzend voneinander zu definieren, wobei ich mich auf die Definitionen von Hastie (1986) sowie Stasser und Dietz-Uhler (2001) beziehe. Deren zentrales Merkmal, das Urteile von Entscheidungen abgrenzt, ist die formale Beschreibung des jeweiligen Antwortformats. Während das Antwortformat einer Entscheidung durch nominal oder ordinal skalierte disjunkte Kategorien abgebildet wird, ist das Antwortformat eines Urteils durch in der Regel intervallskalierte (aber zumindest ordinale) kontinuierliche Variablen repräsentiert. Eine Entscheidung ist daher die Auswahl genau einer Alternative aus einer Menge von mindestens zwei disjunkten Alternativen. In Abgrenzung dazu bezeichnet Urteilen die quantitative und im Idealfall stufenlose Einschätzung einer Sache oder eines Sachverhalts hinsichtlich einer Bewertungsdimension.

Der Unterschied hinsichtlich des Antwortformats ist aber nicht allein ein Merkmal zur definitorischen Abgrenzung von Urteilen und Entscheiden, sondern hat direkte Implikationen für den Umgang mit Ratschlägen. Bei Entscheidungen stellt sich dabei in der Regel die Frage, ob die Person, die die Entscheidung treffen muss, diejenige Alternative wählt, die sie selbst

ursprünglich favorisiert hat, oder stattdessen einem Ratschlag folgt, indem sie diejenige Alternative wählt, die der Ratgeber[1] vorgeschlagen hat. Ganz gleich, für welche der beiden möglichen Alternativen sich die Person letztendlich entscheidet, läuft der Prozess nach dem Prinzip „the winner takes it all" ab, d.h. eine der Alternativen wird zu 100% übernommen, während die konkurrierende Alternative komplett ausgeschlagen wird. Im Gegensatz dazu lässt ein Urteil mehr Spielraum für die Berücksichtigung von Ratschlägen, weil sie auch graduell erfolgen kann. Konkret bedeutet das, dass die Person, die ein Urteil fällt, genau wie bei einer Entscheidung, sowohl an ihrem eigenen ersten Urteil festhalten oder den Ratschlag vollständig übernehmen kann; zusätzlich aber kann sie ihr Urteil aber auch beliebig weit in Richtung des Ratschlags anpassen, während bei einer Entscheidung nur die Wahl zwischen Annahme und Ablehnung des Ratschlags bleibt (Soll & Larrick, 2009). Im Falle des Urteils kann daher auch genauer festgestellt werden, wie stark der Einfluss des Ratgebers letztendlich war, nämlich anhand der prozentualen Anpassung des finalen Urteils in Richtung des Ratschlags (eine konkretere Ausführung bezüglich der Messung des Einflusses von Ratschlägen findet sich im Abschnitt 2.3). Aufgrund des höheren Auflösungsgrads bezüglich der Gewichtung von Ratschlägen hat sich die Mehrheit der empirischen Arbeiten zum Umgang mit Ratschlägen Urteilen anstelle von Entscheidungen bedient (siehe Bonaccio & Dalal, 2006). Aus dem gleichen Grund verwende ich in den Experimenten meiner Dissertation ebenfalls Urteilsaufgaben in Form quantitativer Schätzungen. Entsprechend werde ich im Folgenden meine Argumentation auch weitgehend auf Urteile beschränken und nicht immer gesondert auf Entscheidungen eingehen. Ebenfalls in Anlehnung an die existierende Forschung verwende ich das dort gängige experimentelle Paradigma, das ich im Folgenden kurz beschreibe.

2.2 Das Judge-Advisor-Paradigma

Die Forschung zum Umgang mit Ratschlägen bedient sich im Wesentlichen eines festen Schemas, das als Judge-Advisor-Paradigma oder Judge-Advisor-System bezeichnet wird (Sniezek & Buckley, 1995; für einen Überblick siehe Bonaccio & Dalal, 2006). Das Judge-Advisor-Pradigma entstammt der Forschung zu quantitativen Urteilen in Kleingruppen und trägt der Beobachtung Rechnung, dass auch dann, wenn die Gruppe offiziell führerlos ist und

[1] Um die Lesbarkeit des Textes zu erhalten, wurde auf die separate Nennung beider Geschlechter verzichtet. In der Regel wurden geschlechtsneutrale Formulierungen verwendet. An Stellen, wo dies der Lesbarkeit des Textes abträglich gewesen wäre, wurde das generische Maskulinum verwendet. Gemeint sind jedoch explizit beide Geschlechter.

alle Gruppenmitglieder gleichberechtigt sind, in der Regel ein Gruppenmitglied eine Führungsrolle einnimmt (Sniezek & Buckley, 1995). Dieses Gruppenmitglied hat dann qua Status einen stärkeren Einfluss auf das Urteil der Gruppe als die übrigen Mitglieder. Das Judge-Advisor-Paradigma bildet nun den Extremfall ab, in dem der Einfluss des Führungsmitglieds innerhalb ihrer Gruppe maximal ist. Es stellt also eine Situation dar, in der eine Person ein Urteil weitestgehend eigenständig fällt. Diese Person wird in der Regel als Judge bezeichnet und zwar unabhängig davon, ob tatsächlich ein Urteil gefällt oder eine Entscheidung getroffen wird. Der Judge trifft das Urteil zwar individuell, jedoch innerhalb eines sozialen Kontextes, der mindestens eine weitere Person umfasst. Diese Person, als Advisor bezeichnet, hat selbst keine Entscheidungsgewalt. Sie kann aber indirekt auf das Urteil Einfluss nehmen, indem sie den Judge mit zusätzlichen Informationen versorgt oder ihm mitteilt, wie sie an seiner Statt urteilen würde, indem sie also Ratschläge erteilt. Die Interaktion von Judge und Advisor läuft im Judge-Advisor-Paradigma ebenfalls nach einem festen Schema ab. Zunächst fällt der Judge individuell ein Initialurteil, und zwar ohne jedwede Interaktion mit dem Advisor. Anschließend erhält der Judge einen Ratschlag vom Advisor, meist in Form des Urteils, das der Advisor individuell gefällt hat. Anschließend gibt der Judge sein Finalurteil ab und hat damit Gelegenheit, sein Initialurteil zu revidieren. Es sei an dieser Stelle angemerkt, dass dieser Ablauf an vielen Stellen variiert wird, um verschiedenen Fragestellungen gerecht zu werden. So kann beispielsweise variiert werden, von wie vielen Advisors der Judge einen Ratschlag erhält, ob Judge und Advisor tatsächlich real interagieren oder der Judge lediglich die Meinung des Advisors als schriftlichen Ratschlag erhält, ob der Judge ein Initialurteil fällen soll oder nicht und vieles mehr (siehe Bonaccio & Dalal, 2006).

2.3 Maße für die Nutzung von Ratschlägen im Judge-Advisor-Paradigma

Die zentrale Eigenschaft des Judge-Advisor-Paradigmas besteht darin, dass genau gemessen werden kann, wie stark ein Ratschlag berücksichtigt wurde, sofern sich Ratschlag und Initialurteil voneinander unterscheiden. Im Falle von Urteilen kann ein Ratschlag vollständig übernommen, vollständig abgelehnt und auch graduell berücksichtigt werden (Soll & Larick, 2009). Hier ist entscheidend, wie stark der Judge sein Finalurteil in Richtung des erhaltenen Ratschlags anpasst. In der bisherigen Forschung zum Umgang mit Ratschlägen existieren im Prinzip zwei Klassen von Ansätzen, um den Grad der Nutzung von Ratschlägen zu bestimmen, nämlich einerseits regressionsbasierte Ansätze und andererseits formelbasierte

Ansätze, die den Grad der Nutzung als anteilige Gewichtung von Initialurteil und Ratschlag verstehen (siehe Bonaccio & Dalal, 2006). Ich werde im Folgenden die gängigsten Maße der beiden Klassen kurz skizzieren und vor allem darauf eingehen, welche methodischen Probleme diese Maße jeweils mit sich bringen.

Die regressionsbasierten Ansätze versuchen grundsätzlich, das Finalurteil eines Judges mittels multipler Regression aus dem zugehörigen Initialurteil und den erhaltenen Ratschlägen zu berechnen (z.B. Harvey, Harries & Fischer, 2000). Es wird also für jeden Judge eine eigene Regressionsanalyse durchgeführt. Diese Methode setzt dementsprechend voraus, dass Judge und Advisor(s) über mehrere Durchgänge hinweg zusammenarbeiten. Der Einfluss, den die Initialurteile des Judges sowie die einzelnen Ratgeber auf die Finalschätzungen haben, ist dann durch die entsprechenden Regressionsgewichte abgebildet. Das zentrale Problem der regressionsbasierten Ansätze besteht in der Annahme, dass der wahre Wert der Gewichtung der Ratschläge über alle Durchgänge hinweg konstant ist und beobachtete Schwankungen allein durch einen unsystematischen Fehler verursacht werden. Ein zweites Problem der regressionsbasierten Maße besteht in der Multikollinearität, die dann auftritt, wenn Judge und Advisor(s) keine vollständig unabhängigen Schätzungen abgeben, und die Interpretierbarkeit der Regressionsgewichte erschwert. Wenn nämlich beispielsweise die Schätzungen von Judge und Advisor nicht unabhängig sind und demnach Varianz teilen, dann wird einer der beiden Prädiktoren rein statistisch die Varianzaufklärung im Sinne des Regressionsgewichts auf sich ziehen. Allerdings kann daraus noch nicht gefolgert werden, dass der Judge den entsprechenden Schätzungen auch tatsächlich mehr Beachtung geschenkt hat.

Das Problem der Multikollinearität kann mit einer zweiten Variante der Regression basierten Ansätze teilweise ausgeräumt werden, nämlich durch die Verwendung so genannter *Utilization Indices* (z.B. Azen & Budescu, 2003; Budescu & Azen, 2004). Diese Indizes werden berechnet, indem zunächst das volle Regressionsmodell unter Einschluss aller Ratgeber berechnet wird. Im zweiten Schritt wird dann derjenige Ratgeber, dessen Index bestimmt werden soll, als Prädiktor ausgeschlossen. Der Utilization Index ergibt sich dann als Differenz der Erklärungskraft der beiden Regressionsmodelle (z.B. Gomez-Beldarrain, Harries, Garcia-Monco, Ballus & Grafman, 2004; Harvey, Harries & Fischer, 2000). Die Grundidee ist also,

dass ein Ratgeber dann einen umso stärkeren Einfluss auf die Finalschätzung hatte, je stärker sein Ausschluss aus dem Regressionsmodell dessen Erklärungskraft verringert.

Regressionsbasierte Ansätze zur Bestimmung der Gewichtung von Ratschlägen werden häufig als Methode der Wahl bezeichnet, im Falle von Utilization Indices sogar als Königsweg (LeBreton, Polyheart & Ladd, 2004). Allerdings weisen diese Ansätze ein weiteres zentrales Problem auf. Da sie auf Korrelationen basieren, ist für den Grad der Nutzung nur relevant, wie stark Ratschlag und Finalschätzung systematisch kovariieren. Nicht berücksichtigt wird hingegen der Bias des Advisors, und zwar weder in der Form einer systematischen Über- oder Unterschätzung um einen festen Betrag, noch in Form einer fehlerhaften Skalierung mit einem Faktor ungleich 1, da beides lediglich Lineartransformationen sind, die keinen Einfluss auf das Ausmaß der linearen Abhängigkeit haben. Ein Judge, der sich entscheidet, als Finalschätzung stets exakt drei siebtel dessen anzugeben, was sein Advisor schätzt, würde sich nach der Logik der regressionsbasierten vollständig an den Ratschlägen orientieren, obwohl die Ratschläge stets diskontiert würden. Es wird deutlich, dass die regressionsbasierten Ansätze deshalb davon ausgehen müssen, dass der Judge einen eventuellen Bias des Advisors kennt oder zumindest erkennen kann und bei der Finalschätzung für diesen Bias korrigiert.

Während für Situationen mit mehreren Ratgebern nur regressionsbasierte Ansätze in Frage kommen, existieren zumindest für Situationen mit nur einem Ratgeber einfachere Methoden, um die Nutzung von Ratschlägen zu quantifizieren, die nicht auf Korrelationen zwischen Ratschlag und Finalschätzung basieren und für die daher ein Bias des Ratgebers kein Problem darstellt. Es handelt sich dabei um formelbasierte Ansätze, die die Finalschätzung des Judges als gewichtetes Mittel aus Initialschätzung und Ratschlag verstehen. Der wohl prominenteste dieser Ansätze ist der *weight of advice* (WOA), der von Yaniv (2004a) wie folgt definiert wurde:

Der WOA setzt also die absolute Veränderung von Initialschätzung zu Finalschätzung des Judges ins Verhältnis zum absoluten Abstand zwischen Initialschätzung des Judges und Ratschlag des Advisors. Diese Kennzahl ist nur definiert für Fälle, in denen Initialschätzung

und Ratschlag nicht identisch sind. Sie nimmt immer dann den Wert 0 an, wenn Finalschätzung und Initialschätzung identisch sind, der Judge also den Ratschlag vollständig ignoriert. Für den Fall, dass der Judge den Ratschlag vollständig übernimmt, nimmt AT den Wert 1 an, während Werte größer 1 bedeuten, dass die Finalschätzung so stark in Richtung des Ratschlags angepasst wird, dass sie über den Ratschlag hinausschießt. Werte zwischen 0 und 1 geben an, wie stark der Ratschlag prozentual in die Finalschätzung eingeflossen ist. In ca. 95% der Fälle liegt der WOA tatsächlich zwischen 0 und 1 (Gino, 2008; Gino, Shang & Croson, 2009; Yaniv, 2004a), weshalb davon ausgegangen wird, dass der WOA relativ gut widerspiegelt, wie Ratschläge genutzt werden. Etwas seltener findet sich das gegenteilige Maß zum WOA, der so genannte *weight of estimate* (WOE), der entsprechend angibt, welches prozentuale Gewicht die Initialschätzung erhalten hat (z.B. Yaniv & Kleinberger, 2000). Der WOE ist wie folgt definiert:

Daraus ergibt sich, dass der WOE im Prinzip dieselben Eigenschaften hat wie der WOA, also in der Regel zwischen 0 und 1 liegt, und dass sich WOA und WOE für jede definierte Situation zu 1 aufsummieren. Das zentrale Problem von WOA und WOE besteht darin, dass jeweils der Betrag der relevanten Differenzen herangezogen wird. Deshalb bewerten diese formelbasierten Ansätze es gleichermaßen als Nutzung eines Ratschlags, wenn der Judge sich bei der Finalschätzung 50% auf die Schätzung des Ratgebers zubewegt und wenn er sich 50% von der Schätzung wegbewegt. Diese Annahme ist aber nur dann sinnvoll, wenn man davon ausgeht, dass der Judge in der Lage ist jeweils zu erkennen, ob der Ratschlag möglicherweise eine negative Validität aufweist und daraufhin konsequent das Gegenteil dessen tut, was der Ratschlag nahelegt. Ein wohl noch drastischeres Problem, das ebenfalls durch die Beträge der Differenzen entsteht, liegt darin, dass jegliche Veränderung zwischen Initial- und Finalschätzung automatisch immer als Nutzung des Ratschlags interpretiert wird. WOA und WOE bilden also nur unter der Annahme die echte Nutzung eines Ratschlags ab, dass weder unsystematische Schwankungen noch systematische Veränderungen, die nicht auf den Ratschlag zurückzuführen sind, existieren. Geht man hingegen davon aus, dass unsystematische Schwankungen und weitere systematische Veränderungen zwischen Initialschätzung und Finalschätzung, beispielsweise im Sinne einer Fehlerkorrektur, denkbar sind,

dann laufen WOA und WOE Gefahr, das Ausmaß der Nutzung der Ratschläge systematisch zu überschätzen.

Der dritte und letzte formelbasierte Ansatz ist in der Lage, dieses Problem zu umgehen. Es handelt sich dabei um die Kennzahl a*dvice taking* (AT), die von Harvey und Fischer (1997) wie folgt definiert wurde:

Der AT entspricht also im Wesentlichen dem WOA, jedoch ohne die Absolutbeträge im Zähler und Nenner. Damit lässt der AT Werte kleiner 0 (also negative Werte) zu. Negative Werte von AT deuten darauf hin, dass der Judge seine Finalschätzung entgegen des Ratschlags anpasst. Obwohl Werte kleiner als 0 theoretisch möglich sind, kommen sie in der Praxis relativ selten vor, nämlich in ca. 1% der Fälle (Gino et al., 2009). Der entscheidende Vorteil des AT gegenüber WOA und WOE besteht dabei aber vor allem darin, dass unsystematische Schwankungen sich ausmitteln können und das Ausmaß der Nutzung von Ratschlägen nicht artifiziell überschätzt wird. Dies ist von höchster Wichtigkeit für meine Dissertation, da hier je gerade ein Nachweis für die Übernutzung von Ratschlägen erbracht werden soll. Da der AT im Vergleich zum WOA die Nutzung von Ratschlägen eher konservativ bewertet, eignet es sich hervorragend für einen kritischen und methodisch sauberen Test der Übernutzungshypothese. Nutzte ich hingegen den WOA, so bestünde das Risiko, dass artifiziell der Eindruck entstünde, Ratschläge würden über Gebühr genutzt.

Unter Abwägung der Vor- und Nachteile der hier dargestellten Methoden, wird in der vorliegenden Arbeit der AT als Maß für die Nutzung von Ratschlägen ausgewählt. Die formelbasierten Ansätze sind inhaltlich leichter interpretierbar als die regressionsbasierten Ansätze, da erstere direkt die Prozentuale Gewichtung des Ratschlags angeben, letztere jedoch zunächst nur ein Regressionsgewicht, das erst im Vergleich mit dem Regressionsgewicht der Initialschätzung interpretierbar wird. Durch die Wahl eines formalisierten Ansatzes ist zudem die Vergleichbarkeit mit der überwiegenden Mehrheit der bisher veröffentlichten Studien mit nur einem Ratgeber gewährleistet, da diese Studien fast ausschließlich entweder den WOA, den WOE oder den AT nutzen (für Ausnahmen siehe Gardner & Berry, 1995; Lim & O'Connor, 1995). Innerhalb der formelbasierten Ansätze ist wiederum der AT dasjenige Maß,

das im Vergleich zu den übrigen formelbasierten Maßen meine Hypothese, dass Ratschläge unter bestimmten Bedingungen zu stark gewichtet werden, nicht artifiziell begünstigt, da es mögliche unsystematische Veränderungen zwischen Initial- und Finalschätzung nicht als Nutzung eines Ratschlags interpretiert. Entsprechend wird die Nutzung von Ratschlägen in den Experimenten meiner Dissertation also im Sinne des AT quantifiziert.

2.4 Die systematische Unternutzung von Ratschlägen

Wie eingangs geschildert, zeigt die Literatur zum Umgang mit Ratschlägen vor allem, dass Ratschläge nicht stark genug berücksichtigt werden, was in meiner Dissertation als Unternutzung bezeichnet wird (siehe Abschnitt 1). Es sei dabei gleich angemerkt, dass keine einheitlichen Kriterien dafür existieren, wie stark ein Ratschlag überhaupt gewichtet werden sollte. In so fern beruht die Aussage, Ratschläge würden über Gebühr diskontiert, häufig auf Plausibilitätsüberlegungen, allen voran die Überlegung, dass Ratschläge einer Person, die genau so kompetent ist wie der Judge, normativ exakt zu 50% gewichtet werden sollte (Soll & Larrick, 2006). Aus dieser zentralen Überlegung kann man ferner ableiten, dass die korrekte Gewichtung einer Person, die weniger kompetent ist als der Judge, unter 50% liegen muss, während Ratgeber, die kompetenter sind als der Judge, zu mehr als 50% gewichtet werden müssten. Die exakte Gewichtung kann hier aber für beide Fälle nicht bestimmt werden. Anhand dieser Überlegungen kann man zumindest zwei Fälle ableiten, in denen definitiv eine Unternutzung von Ratschlägen vorliegt, nämlich dann, wenn ein gleich kompetenter Ratgeber zu weniger als 50% gewichtet wird und wenn ein kompetenterer Ratgeber zu 50% oder weniger gewichtet wird.

In verschiedenen Untersuchungen zeigt sich nun, dass Personen, gemessen an den gerade beschriebenen Plausibilitätsüberlegungen, Ratschläge in der Tat zu wenig berücksichtigen. So wurde mehrfach gezeigt, dass Menschen Ratschläge von Personen, die entweder genauso kompetent sind wie sie selbst oder bei denen aufgrund fehlender Informationen die Annahme gleicher Kompetenz nahe liegt, nur zu ca. 30% anstelle der zu erwartenden 50% gewichten (Harvey & Fischer, 1997; Soll & Larrik, 2009; Yaniv, 2004b; Yaniv & Kleinberger, 2000, Yaniv & Milyawsky, 2007). Gleichermaßen werden Personen, die objektiv deutlich kompetenter sind als der Judge, nur zu ca. 50% gewichtet, obwohl die objektive Akkuratheit der Schätzungen ein höheres Gewicht nahelegt. Dieser Effekt bleibt sogar dann stabil, wenn die Probanden veridikales Feedback erhalten, anhand dessen sie feststellen können, dass die

Ratschläge akkurater sind als ihre eigenen Schätzungen (Yaniv & Kleinberger, 2000). Ein ähnliches Muster findet sich auch bei Harvey und Fischer (1997), nämlich dass Laien bei Prognosen den Ratschlag von Personen, die bereits in der Aufgabe trainiert waren, nur zu ca. 40% gewichteten. Selbst unter der Annahme, dass durch das Training die Fähigkeit, akkurate Prognosen zu liefern, unbeeinflusst geblieben wäre, wäre jedoch eine Gewichtung zu 50% sinnvoll gewesen. Dass also ein Expertenratschlag durch Laien weniger als 50% gewichtet wird, deutet auf eine Unternutzung der Ratschläge hin.

Weitere Evidenz dafür, dass Ratschläge über die Gebühr diskontiert werden, findet sich bei Lim und O'Connor (1995) sowie bei Gardner und Berry (1996). In der Studie von Lim und O'Connor (1995) wurden die Probanden gebeten, über 30 Durchgänge hinweg anhand der eigenen Prognose und einer statistischen Vorhersage die Verkaufszahlen eines Produktes zu schätzen. Mittels einer Regressionsgleichung wurde gezeigt, dass der Einfluss der eignen Initialschätzung im Mittel doppelt so hoch war wie der der statistischen Vorhersage, und zwar auch dann, wenn die Probanden jeweils explizit darauf hingewiesen um wie viele Prozentpunkte diese Vorhersage im Mittel akkurater war als ihre Initialschätzungen. Ferner konnte anhand von Regressionsgleichungen berechnet werden, wie stark die jeweiligen Ratschläge in Form der statistischen Vorhersage hätten berücksichtigt werden müssen, um den maximalen Grad an Akkuratheit zu erreichen. Hier zeigte sich konsistent in drei Experimenten, dass die tatsächliche Gewichtung deutlich geringer war als die objektiv optimale Gewichtung und die Ratschläge gemessen an der objektiven Qualität zu wenig genutzt wurden.

Den wohl eindeutigsten Nachweis für die Unternutzung von Ratschlägen liefern jedoch Gardner und Berry (1995), die ihre Probanden baten, in einer medizinischen Simulation über mehrere Durchgänge hinweg einem fiktiven Patienten bestimmte Dosen verschiedener Medikamente zu verabreichen, um verschiedene Zielgrößen wie Puls und Blutdruck auf einem optimalen Niveau zu stabilisieren. Den Probanden stand dabei in jedem Durchgang die Empfehlung eines medizinischen Expertensystems zur Verfügung, das auch jeweils die optimale Dosierung für jedes Medikament vorschlug. Dies wurde den Probanden auch explizit mitgeteilt, das heißt, die Probanden waren sich bewusst, dass die Ratschläge des Expertensystems stets die bestmögliche Dosierung darstellen würden. Die normativ optimale Strategie hätte also darin bestanden, die Ratschläge stets vollständig zu übernehmen. Stattdessen

zeigte sich, dass die Probanden die Ratschläge des Expertensystems bis zu 50% diskontierten.

Die Unternutzung von Ratschlägen scheint dabei nicht auf Situationen mit nur einem Ratgeber beschränkt zu sein. Die bisher einzige Studie, in der mehrere Ratgeber eingesetzt werden und bei der der Judge gleichzeitig auch eine Initialschätzung abgibt (Yaniv & Milyawsky, 2007)[2], zeigt ebenfalls, dass Ratschläge diskontiert werden. Allerdings werden die Ratschläge nicht alle in gleichem Maße diskontiert, sondern umso stärker, je weiter sie von der Initialschätzung des Judges abweichen. Judges erhielten in dieser Studie entweder Ratschläge von zwei, vier oder acht Ratgebern. Die Finalschätzung des Judges konnte dabei am akkuratesten durch ein Gewichtungsschema erklärt werden, das die beiden Ratschläge, die am weitesten von der Initialschätzung entfernt sind, vollständig ignoriert, und über die übrigen Ratschläge sowie die Initialschätzung mittelt.

In der Summe lässt sich also festhalten, dass Ratschläge oftmals weniger stark gewichtet werden, als ihre objektive oder auch die subjektiv wahrgenommene Qualität nahelegt, was sowohl im Widerspruch zu normativen Modellannahmen steht (Soll & Larrick, 2009) als auch nachteilig für die Akkuratheit der Finalschätzungen ist (Gardner & Berry, 1995, Lim & O'Connor, 1995; Yaniv, 2004b). Da die Unternutzung von Ratschlägen also ein robustes und situationsübergreifendes Phänomen ist, das die Qualität von Urteilen gefährdet, ist es leicht verständlich, dass der Großteil der Forschung zum Umgang mit Ratschlägen darauf abzielt, ein besseres Verständnis für das Zustandekommen des Phänomens zu erlangen (Clement & Krueger, 2000; Harvey & Fischer, 1997; Harvey & Harries, 2004; Lim & O'Connor, 1995; Yaniv, 2004a, 20004b; Yaniv & Kleinberger, 2000), Moderatoren des Effekts zu ermitteln (Feng & MacGeorge, 2006; Gino, 2008; Gino & Schweitzer, 2008; Gino et al., 2009; Harvey & Fischer, 1997; Harvey et al., 2000; Sniezek & van Swol, 2001) oder effektive Interventionen zu entwickeln, die zu einer stärkeren Gewichtung von Ratschlägen führen (Sniezek, Schrah, & Dalal, 2004).

Bezüglich möglicher Erklärungen für die systematische Unternutzung von Ratschlägen kann man drei Ansätze unterscheiden. Den ersten Versuch, die Unternutzung von Ratschlä-

[2] In der Mehrzahl der Studien, in der mehr als ein Advisor Ratschläge erteilen, gibt der Judge keine Initialschätzung ab (siehe Budescu & Rantilla, 2000; Harvey et al., 2000, Harvey & Harries, 2004; Gomez-Beldarrain et al, 2004).

gen zu erklären, unternahmen Lim und O'Connor (1995), die Unternutzung von Ratschlägen als eine spezielle Form des Ankereffekts verstehen, wobei die Initialschätzung als Anker wirkt und daher nur unzureichend in Richtung des Ratschlags adjustiert wird. Dieser Erklärungsansatz gilt jedoch weitestgehend als widerlegt, da Ratschläge auch dann unzureichend berücksichtigt werden, wenn formal kein Anker vorliegt. Das ist beispielsweise der Fall, wenn gar keine Initialschätzung abgegeben wird (Clement & Krueger, 2000). Gleichermaßen zeigt sich auch dann eine Unternutzung von Ratschlägen, wenn man Probanden anstelle der eigenen Initialschätzung die Schätzung einer anderen Person vorlegt, sie aber im Glauben lässt, es handele sich dabei um ihre jeweilige Initialschätzung (Harvey & Harries, 2004). In diesem Falle waren die vermeintlichen Initialschätzungen den Probanden vorher nicht bekannt, und sie wurden parallel zu den Ratschlägen dargeboten, so dass sie nicht als Anker hätten wirken können. Dennoch zeigte sich die Unternutzung der Ratschläge im Vergleich zu den vermeintlichen Initialschätzungen.

Als Alternative zu Anker-Effekten bzw. der unzureichenden Adjustierung als Erklärung der Unternutzung von wurde ein egozentrischer Bias postuliert (Clement & krueger, 2000; Harvey & Harries, 2004; Krueger, 2003), der besagt, dass Judges ihre eigenen Initialschätzungen präferieren, weil sie deren Qualität überschätzen. Ein dritter und zu dem egozentrischen Bias komplementärer Erklärungsansatz stammt von Yaniv und Kleinberger (2000; siehe auch Yaniv, 2004a, 2004b). Sie gehen davon aus, dass Ratschläge deshalb zu wenig genutzt werden, weil eine Informationsasymmetrie zwischen Judge und Advisor besteht. Während der Judge durch Introspektion vollständige Information darüber hat, aus welchen Gründen er eine bestimmte Schätzung abgibt und diese dadurch argumentativ sowohl nachvollziehen als auch rechtfertigen kann, sind genau diese Überlegungen des Advisors in der Regel nicht oder nicht vollständig verfügbar. Als Konsequenz erscheint dem Judge die Glaubwürdigkeit seiner eigenen Initialschätzung höher als die des Ratschlags. Obwohl also noch keine einheitliche Theorie der Unternutzung von Ratschlägen existiert (Schrah, Dalal & Sniezek, 2006), liegen mit dem egozentrischen Bias und der Informationsasymmetrie zwei plausible Erklärungsansätze vor, die einheitlich davon ausgehen, die Unternutzung von Ratschlägen ließe sich vor allem auf eine Fehlwahrnehmung der relativen Qualität von Initialschätzungen und Ratschlägen zurückführen.

Im Gegensatz zu der sehr umfangreichen Forschung zur Unternutzung von Ratschlägen bleibt die Frage nach der Kehrseite der Medaille fast vollständig unberücksichtigt, also die Möglichkeit, dass Ratschläge unter bestimmten Bedingungen systematisch zu stark gewichtet werden. Damit leite ich zu der Fragestellung meiner Dissertation über.

3. Ableitung der Fragestellung

Wie ich eingangs kurz dargelegt habe, verfolgt meine Dissertation das Ziel zu zeigen, dass Ratschläge nicht generell zu wenig berücksichtigt werden, sondern dass auch der gegenteilige Fall eintreten kann. Obwohl die Unternutzung von Ratschlägen ein robustes und situationsübergreifendes Phänomen ist, vertrete ich die Auffassung, dass unter bestimmten Bedingungen Ratschläge nicht diskontiert sondern sogar stärker gewichtet werden, als sie sollten. Diese Auffassung begründe ich durch zwei Kernargumente: zum einen sind in der Forschung zum Urteilen und Entscheiden Phänomene bekannt, in denen Hinweisreize eindeutig zu stark ins Urteil einbezogen werden, allen voran der Anker-Effekt (Tversky & Kahnemann, 1974); zum anderen lassen sich auch in der Literatur zum Umgang mit Ratschlägen erste Hinweise darauf finden, dass eine systematische Übernutzung von Ratschlägen stattgefunden haben könnte. Ich werde im Folgenden sowohl bereits anerkannte Nachweise einer systematischen Übernutzung von Hinweisreizen im Allgemeinen als auch die entsprechenden Hinweise auf die Übernutzung von Ratschlägen skizzieren. Abschließend werde ich kurz darstellen, auf welche Weise man einfach und eindeutig den Nachweis einer systematischen Übernutzung erbringen kann.

3.1 Allgemeine Befunde zur Übernutzung von Hinweisreizen

Ein naheliegender Grund für die Annahme, dass auch eine Übernutzung von Ratschlägen denkbar ist, ist die Feststellung, dass es bereits in anderen Forschungsbereichen Nachweise für die Übernutzung von Informationen oder Hinweisreizen gibt. In der Literatur zum Urteilen und Entscheiden lassen sich mindestens zwei solcher Phänomene finden. Eines dieser Phänomene stammt aus der Forschung zu so genannten Multi-Cue-Judgments, also Urteilen auf Basis mehrerer Hinweisreize, die mehr oder weniger stark mit der zu schätzenden Zielgröße in Verbindung stehen. Ein prominentes Beispiel ist die Vorhersage der akademischen Leistung von Studenten anhand bestimmter Kennzahlen wie z.B. Schulnoten, College-Noten und dem Renommee der jeweiligen Universität (z.B. Dawes, 1979). Die optimale

Gewichtung der einzelnen Hinweisreize ergibt sich dabei aus dem so genannten „best linear model“, also der multiplen Regressionsgleichung, die die beste Vorhersagekraft bezüglich der zu schätzenden Variable aufweist. Hier zeigt sich nicht nur, dass statistische Modelle selbst die akkurateste menschliche Prognose übertreffen, sondern auch, dass die Überlegenheit der statistischen Modelle zum Teil daher rührt, dass Menschen nicht oder nur unzureichend in der Lage sind, valide Hinweisreize von nicht validen zu unterscheiden (Dawes, Faust & Meehl, 1989). Das heißt, dass bei diesen Multi-Cue-Judgments wenig oder nicht valide Hinweise zu stark und dadurch valide Hinweise zu wenig einbezogen werden, und zwar vorranging deshalb, weil die Validität der Hinweisreize falsch eingeschätzt wird (Dawes, 1979).

Das zweite prominente Beispiel für die systematische Übernutzung von Hinweisreizen stellt der Anker-Effekt dar (Tversky & Kahneman, 1974). Dieser Effekt besagt, dass – insbesondere quantitative – Urteile unter bestimmten Bedingungen durch vollkommen irrelevante numerische Werte verzerrt werden. In der klassischen Studie von Tversky und Kahnemann wurden die Probanden gebeten zu schätzen, welcher Anteil der Staaten Afrikas zum Zeitpunkt der Studie Mitglied der UNO waren. Zuvor sollten die Probanden jedoch an einem Glücksrad mit den Zahlen 1 bis 100 drehen und vor ihrer Schätzung angeben, ob der Anteil höher oder niedriger wäre als die vom Glücksrad angezeigte Zahl. Tatsächlich war das Glücksrad so eingestellt, dass es die Hälfte der Probanden mit einer hohen Zahl, nämlich 65, und die andere Hälfte mit einer niedrigen Zahl, nämlich 10, konfrontierte. Entgegen normativer Annahmen fielen die Schätzungen der Probanden, die zuvor die Zahl 65 gedreht hatten, mit 45% deutlich höher aus als die der Probanden, die eine 10 gedreht hatten und darauf folgend im Mittel 25% schätzten. Das Urteil der Probanden war also aufgrund eines willkürlichen und für den aktuellen Sachverhalt erkennbar irrelevanten Zahlenwertes verzerrt. Dieser Effekt ist mittlerweile in vielen Bereichen repliziert worden, so zum Beispiel im Bereich des Allgemeinwissens (z.B. Jacowitz & Kahmemann, 1995; Strack & Mussweiler, 1997) oder bei der Bewertung von Immobilien- oder Autopreisen (Northcraft & Neale, 1987; Mussweiler, Pfeiffer & Strack, 2000). Außerdem gilt der Anker-Effekt als extrem robust. So tritt er auch dann auf, wenn finanzielle Anreize für die Akkuratheit der Schätzungen geboten werden oder wenn man die Probanden zuvor explizit über den Anker-Effekt aufklärt und sie instruiert, sich nicht von dem Anker beeinflussen zu lassen (Wilson, Houston, Etling & Brekke, 1996). Allerdings gibt es einen zentralen Unterschied zwischen Anker-Paradigma und Judge-

Advisor-Paradigma, der in der Validität der jeweiligen präsentierten Stimuli besteht. Während Anker in der Regel willkürlich gewählte Werte ohne Informationswert darstellen, sind Ratschläge in bestimmtem Maße valide, das bedeutet, dass ihre Berücksichtigung anders als beim Anker-Effekt durchaus rational und für die Urteilsqualität förderlich ist. Weiterhin müssen Personen im Judge-Advisor-Paradigma kein komparatives Urteil dahingehend fällen, ob der wahre Wert über oder unter dem Ratschlag liegen wird, während dieses komparative Urteil als entscheidend für die Entstehung von Anker-Effekten angesehen wird.

Sowohl der Nachweis übermäßiger Gewichtung von Hinweisreizen im Multi-Cue-Judgment als auch das Auftreten von Anker-Effekten zeigen, dass eine systematische Übernutzung von Hinweisreizen bei individuellen Urteilen auftreten kann. Daher lässt sich schlussfolgern, dass zumindest theoretisch die Möglichkeit besteht, dass ähnliche Phänomene auch im Judge-Advisor-System auftreten könnten.

3.2 Empirische Evidenz für die Übernutzung von Ratschlägen

Bis dato existiert noch keine veröffentlichte Arbeit, die systematisch die Übernutzung von Ratschlägen untersucht hat. Nichtsdestotrotz finden sich in der Forschung zum Umgang mit Ratschlägen zumindest zwei empirische Arbeiten, deren Befunde im Sinne einer Übernutzung von Ratschlägen interpretiert werden können. Zum einen handelt es sich um eine Studie von Gino (2008), die zeigte, dass Ratschläge dann besonders stark berücksichtigt werden, wenn die Probanden dafür zahlen müssen. So werden dieselben Ratschläge im Mittel zu ca. 40% gewichtet, wenn sie kostenlos sind, und zu ca. 60%, wenn die Probanden dafür zahlen mussten. Die Höhe der Kosten hat dabei ebenfalls einen Einfluss auf den Grad der Nutzung, wobei mit steigenden Kosten die Nutzung höher ausfällt. Diesen so genannten Paid-Advice-Effekt erklärt Gino über versunkene Kosten (Arkes & Blumer, 1985). Nun bestünde die Möglichkeit, dass Probanden Ratschläge, für die sie Geld zahlen mussten, als qualitativ höherwertig einstuften, weshalb eine höhere Gewichtung dieser Ratschläge dann einem rationalen Kalkül folgte. Diese Alternativerklärung konnte Gino jedoch ausschließen, indem sie die wahrgenommene Qualität der Ratschläge erfasste und statistisch kontrollierte; der Paid-Advice-Effekt blieb dabei aber unverändert bestehen (wobei kritisch angemerkt werden muss, dass die wahrgenommene Qualität der Ratschläge keinen Einfluss auf den Grad der Nutzung hatte – die entsprechenden Angaben der Probanden sind daher nicht eindeutig interpretierbar). Gino schlussfolgerte daraus, dass dieselben Ratschläge ungeachtet der

wahrgenommenen Qualität aufgrund versunkener Kosten unterschiedlich stark gewichtet werden. Es besteht insofern die Möglichkeit, dass bezahlte Ratschläge systematisch zu stark gewichtet wurden. Allerdings kann dies nicht überprüft werden, da Gino (2008) weder Aussagen über die normativ korrekte Gewichtung der Ratschläge macht, noch darüber, ob sich durch die erhöhte Gewichtung der Ratgeber auch die Akkuratheit der Finalschätzungen verschlechtert. Wenn nämlich – wie häufig argumentiert – die Qualität von Ratschlägen relativ zur eigenen Initialschätzung systematisch unterschätzt wird (Clement & Krueger, 2000; Harvey & Harries, 2004; Krueger, 2003), dann wäre es denkbar, dass erst durch den Paid-Advice-Effekt dasjenige Niveau der Gewichtung erreicht wird, das der objektiven Qualität der Ratschläge gerecht wird.

Eine weitere Studie, deren Ergebnisse auf eine mögliche Übernutzung von Ratschlägen hindeuten, stammt von Harvey und Fischer (1997), die in einer Serie von drei Experimenten untersuchten, wie sich die wahrgenommene Kompetenz eines Ratgebers auf dessen Gewichtung auswirkt. Dabei waren sämtliche Ratschläge ohne Wissen der Probanden computergeneriert. Die wahrgenommene Kompetenz des vermeintlichen Ratgebers wurde über die Vorinformation darüber manipuliert, wie viel Erfahrung der jeweilige Ratgeber zuvor mit der Aufgabe sammeln konnte. Ein vermeintlich unerfahrener Ratgeber hatte demnach noch keine Vorerfahrung mit der Aufgabe, während ein vermeintlich erfahrener Ratgeber bereits 100 Übungsdurchgänge und ein vermeintlicher Expertenratgeber sogar 240 Übungsdurchgänge absolviert hatte. Ferner wurde zwischen den drei Experimenten manipuliert, wie kompetent der Judge bezüglich der Aufgabe war, und zwar analog zur vermeintlichen Kompetenz der Ratgeber. In Experiment 1 waren alle Probanden Laien, das heißt sie hatten keine Vorerfahrung mit der Aufgabe. In Experiment 2 hingegen absolvierten die Probanden vor Beginn des eigentlichen Experiments 100 Übungsdurchgänge und erhielten nach jedem Durchgang eine Rückmeldung über die Akkuratheit ihrer Schätzung, während die Probanden in Experiment 3 sogar 240 Durchgänge absolvierten. Insgesamt sind damit die Probanden in Experiment 1 hinsichtlich Kompetenz vergleichbar mit den Laienratgebern, die Probanden in Experiment 2 entsprechen den erfahrenen Ratgebern und die Probanden aus Experiment 3 sind ähnlich kompetent wie die Expertenratgeber. Eine Analyse über die drei Experimente hinweg, bei der die Datensätze der einzelnen Experimente quasi zu einem zweifaktoriellen Design mit jeweils 3 Faktorstufen zusammengefasst wurden, zeigte zunächst, dass die Probanden mit steigender eigener Kompetenz Ratschläge weniger stark berücksichtigten und

dass Ratgeber umso stärker gewichtet wurden, je kompetenter sie waren. Es zeigte sich aber auch eine unerwartete Interaktion der beiden Faktoren, die dadurch zustande kam, dass der vermeintlich inkompetente Laienratgeber in allen drei Experimenten zu ca. 20% gewichtet wurde. Diese Gewichtung von 20% kann in Experiment 1 im Sinne eine Unternutzung interpretiert werden, da sowohl Judge als auch Advisor hier denselben Kenntnisstand hatten, was basierend auf Plausibilitätsannahmen eine Gewichtung von 50% nahelegt (Soll & Larrick, 2006). Im Sinne solcher Plausibilitätsüberlegungen kann man nun eine weitere Schlussfolgerungen ableiten, nämlich, dass zumindest die Probanden aus Experiment 3, die sehr viel Erfahrung mit der Aufgabe hatten, die Ratschläge von untrainierten Laien, wenn überhaupt, dann nur in sehr geringem Umfang nutzen sollten, was aber nicht der Fall ist. Vor allem aber sollte die Gewichtung der Laienratgeber zwischen den drei Experimenten nicht invariant sein. Sollte die Gewichtung inkompetenter Ratgeber tatsächlich wie in den Ergebnissen von Harvey und Fischer (1997) angedeutet von der Kompetenz des Judges unabhängig und bei ca. 20% konstant sein, so wird zwangsläufig immer dann eine Übernutzung des Ratschlags vorliegen, wenn eine kritische Differenz zwischen der Kompetenz des Judges und der des Advisors überschritten wird und die normativ korrekte Gewichtung dadurch weniger als 20% beträgt. Ob diese Bedingung in den Experimenten von Harvey und Fischer bereits vorlag, lässt sich nicht objektiv bestimmen. Dennoch ist es zumindest theoretisch möglich, dass tatsächlich eine Übernutzung der Ratschläge des untrainierten Ratgebers durch Judges mit hohem Maß an Vorerfahrung vorlag.

In der Summe liefern die Ergebnisse von Harvey und Fischer (1997) sowie Gino (2008) Hinweise darauf, dass eine systematische Übernutzung von Ratschlägen stattgefunden haben könnte, auch wenn der eindeutige Nachweis in beiden Studien nicht geführt werden kann. Deshalb erscheint es mir einerseits fruchtbar und anderseits dringend geboten, einen solchen systematischen Nachweis zu führen.

3.3 Eine Methode zum Nachweis der Übernutzung von Ratschlägen

Wie bereits dargestellt, ist es nicht trivial, einen eindeutigen Nachweis dafür zu liefern, dass ein Ratschlag zu stark gewichtet wurde, weil die Bestimmung der optimalen Gewichtung bisher nicht ohne Weiteres möglich ist. Eine Möglichkeit, dieses Problem zu umgehen, besteht darin, die Ratschläge so zu gestalten, dass die optimale Gewichtung aus der Validität des Ratschlags direkt ableitbar ist, beispielsweise dann, wenn die Kompetenz des

Advisors mit der des Judges identisch ist, was entsprechend der bereits erwähnten Plausibilitätsüberlegungen eine optimale Gewichtung von 50% nahelegt (Soll & Larrick, 2006)[3]. Unabhängig davon, dass sich eine solche Deckungsgleichheit der individuellen Kompetenz zwischen Judge und Advisor schwer herstellen lässt, wurde für diesen Fall bereits gezeigt, dass hier eine deutliche systematische Unternutzung stattfindet (z.B. Harvey & Fischer, 1997). Insbesondere liefern aber Harvey und Fischer einen Hinwies darauf, unter welchen Bedingungen Ratschläge höchstwahrscheinlich über Gebühr berücksichtigt werden, nämlich dann, wenn die Ratschläge im Vergleich zu den Initialschätzung nur wenig akkurat sind, also mit anderen Worten eine geringe Validität aufweisen.

Daher scheint eine zweite Plausibilitätsüberlegung hier vielversprechender, nämlich Ratschläge mit einer Validität von Null darzubieten. Basierend auf derselben Logik, die bei gleicher Validität von Initialschätzungen und Ratschlägen eine Gewichtung zu 50% als normativ korrekt definiert, sollte ein Ratschlag mit einer Validität von Null gar nicht gewichtet werden, weil keine überzufällige Verbesserung der Schätzung zu erwarten ist. Dies gilt vor allem unabhängig von der Kompetenz des Judges. Sofern Ratschläge mit einer Validität von Null aber überzufällig gewichtet werden, liegt vergleichen mit der normativ korrekten Gewichtung eine systematische Übernutzung vor. Die folgenden Experimente bauen auf dieser Grundidee auf und verwenden deshalb unter anderem auch solche Ratschläge, deren Validität augenscheinlich Null ist, nämlich Zufallszahlen.

Zufallszahlen als Ratschläge erlauben also auf sehr einfachem Wege, die systematische Übernutzung von Ratschlägen nachzuweisen, und genau dieser Nachweis soll in den folgenden Experimenten erbracht werden. Ich postuliere dementsprechend, dass Probanden im Judge-Advisor-Paradigma ihre Finalschätzungen überzufällig an Zufallszahlen anpassen, auch wenn klar sein sollte, dass diese Zahlen keinerlei Informationswert bezüglich der zu schätzenden Größe haben. In Experiment 1 wird aufbauend auf der Idee, Zufallszahlen als Ratschläge darzubieten, erstmalig überprüft, ob eine systematische Übernutzung nicht valider Ratschläge auftritt.

[3] Es sei angemerkt, dass hier – wie in der bisherigen Forschung – davon ausgegangen wird, dass der Ratgeber tatsächlich die Absicht hat, dem Judge zu helfen, und daher eine möglichst genau Schätzung abgibt.

4. Experiment 1

4.1 Zielsetzung und Hypothesen

Experiment 1 verfolgt das Ziel, erstmalig einen eindeutigen Nachweis für die Übernutzung von Ratschlägen zu erbringen. Zu diesem Zweck wird zusätzlich zu den Ratschlägen eines vermeintlich kompetenten und eines vermeintlich wenig kompetenten Ratgebers eine dritte Kategorie von Ratschlägen verwendet, deren normativ korrekte Gewichtung eindeutig bestimmbar ist, nämlich vermeintliche Zufallszahlen, die in keinem Bezug zu der aktuellen Schätzaufgabe stehen. Werden solche Zahlen systematisch berücksichtigt, dann liegt eine eindeutige Übernutzung vor. Zunächst wird aber, basierend auf früheren Befunden zur Gewichtung von Ratschlägen (Harvey & Fischer, 1997; Yaniv & Kleinberger, 2000), erwartet, dass die Probanden die Ratgeber entsprechend ihrer vermeintlichen Kompetenz unterschiedlich stark gewichten, und damit zumindest in Grenzen rational handeln. Gleichermaßen sollte der vermeintlich kompetente Ratgeber stärker gewichtet werden als die vermeintliche Zufallszahl. Daraus ergeben sich folgende zwei Hypothesen:

Hypothese 1: Ratschläge des vermeintlich kompetenten Ratgebers werden stärker berücksichtigt als Ratschläge des vermeintlich wenig kompetenten Ratgebers und als die vermeintlichen Zufallszahlen.

Hypothese 2: Personen übernutzen vermeintliche Zufallszahlen, d.h. sie passen ihre Finalschätzung überzufällig in Richtung der Zufallszahl an.

Die Bestätigung von Hypothese 1 dient zwei Zwecken: erstens gibt sie Aufschluss darüber, ob die Manipulation der Ratgeberkompetenz bzw. der Art des Ratschlags erfolgreich war; zweitens kann, wenn Hypothese 2 angenommen wird, davon ausgegangen werden, dass das Verhalten der Probanden in Grundzügen rational ist und aus dem Bemühen resultiert, nach Möglichkeit diejenigen Ratschläge stark zu nutzen, die zu einer starken Verbesserung der Finalschätzung führen sollten. Andernfalls wäre eine Übernutzung von Zufallszahlen inhaltlich nicht interpretierbar.

Um die systematische Gewichtung von Ratschlägen sauber nachzuweisen, muss sie schließlich auch von zufälligen Veränderungen zwischen Initial- und Finalschätzung abgegrenzt werden. Deshalb wurde zusätzlich zu den beiden Ratgebern und der Zufallszahl eine

Kontrollbedingung verwendet, in der die Probanden keine Ratschläge erhielten. Sofern in dieser Kontrollbedingung unsystematische Veränderungen von Initialschätzung zu Finalschätzung auftreten, sollten diese in der Summe Null ergeben.[4] Dies führt zu

Hypothese 3: In der Kontrollbedingung ohne Ratschlag finden nur unsystematische Veränderungen zwischen Initial- und Finalschätzung statt, d.h. die Veränderungen der Schätzungen mitteln sich über die Kontrolldurchgänge des Experiments aus.

4.2 Stichprobe und Design des Experiments

An Experiment 1 nahmen 26 Studierende unterschiedlicher Fachrichtungen der Georg-August-Universität Göttingen teil, davon 12 weibliche Studierende (46%). Das Durchschnittsalter der Teilnehmer lag bei 24,46 Jahren (*SD* = 3,66 Jahre). Experiment 1 folgt einem einfaktoriellen Innersubjekt-Design mit der Art des Ratschlags als 4-stufigem Faktor, der folgende Ausprägungen aufweist: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl vs. Kontrollbedingung ohne Ratschlag.

4.3 Methode

Alle Probanden in Experiment 1 bearbeiteten eine computergestützte Urteilsausgabe, die mittels der Software Presentation® (Version 13.0, www.neurobs.com) programmiert und dargeboten wurde (der entsprechende Programmcode für Experiment 1 ist im digitalen Anhang enthalten).

Die Probanden wurden zunächst durch den Versuchsleiter begrüßt und über den Ablauf des Experiments sowie die Höhe des Versuchspersonenhonorars von 5 Euro informiert. Danach wurden sie in separaten Räumen an einen Computerarbeitsplatz geführt. Die genaue Instruktion der Probanden erfolgte nach Eingabe demographischer Daten (Geschlecht, Alter und Studienfach) schriftlich über das Computerprogramm. Die schriftliche Instruktion wurde dabei auf mehrere Seiten verteilt, und die Probanden konnten die jeweils nächste Seite durch Drücken der Leertaste aufrufen.

[4] Es sei hierbei angemerkt, dass ein leichtes methodisches Problem darin besteht, dass in zahlreichen Schätzaufgaben der Wert Null eine natürliche Untergrenze darstellt, dass also eine Schätzung nur um maximal 100% reduziert, jedoch unendlich weit nach oben korrigiert werden kann. Es besteht also wegen dieser Asymmetrie die theoretische Möglichkeit, dass aufgrund extremer, aber unsystematischer Veränderungen von Initialschätzung zu Finalschätzung artifiziell der Eindruck entsteht, Schätzungen würden systematisch nach oben korrigiert; diese Möglichkeit wird jedoch auf Basis früherer Erfahrungen als vernachlässigbar angesehen.

Auf der ersten Seite der Instruktion wurde den Probanden erklärt, dass das Ziel der Studie sei, herauszufinden, wie akkurat Personen unbekannte Größen schätzen können, wobei diese Schätzungen sich im konkreten Experiment auf Luftlinie-Entfernungen zwischen EU-Hauptstädten beziehen würden. Diese Distanzschätzungen wurden gewählt, weil sie bereits vorgetestet und im Rahmen anderer Experimente erfolgreich eingesetzt wurden (Schultze, Mojzisch & Schulz-Hardt, 2010). Vor allem zeigte sich ein mittleres Niveau an Unsicherheit, d.h. Probanden wissen weder die exakten Antworten noch raten sie, womit eine gute Grundlage für die Nützlichkeit von Ratschlägen gelegt ist. Die Probanden wurden instruiert, die jeweiligen Entfernungen so akkurat wie möglich einzuschätzen. Um den Probanden einen Anreiz zu bieten, die Aufgabe gewissenhaft zu bearbeiten und möglichst akkurate Schätzungen abzugeben, wurde zusätzlich zur normalen Vergütung ein Kinogutschein im Wert von 15 Euro in Aussicht gestellt, den die besten fünf Probanden erhalten würden.

Auf der folgenden Seite wurden die Probanden darüber informiert, dass sie bei einigen Durchgängen eine Hilfestellung in Form eines Ratschlags erhalten würden. Zu diesem Zweck würden aus einer Grundgesamtheit von 100 Personen, die bereits zu einem früheren Zeitpunkt dieselbe Aufgabe bearbeitet hatten, zufällig zwei Personen gezogen. Ein Ratschlag bestünde dann immer aus derjenigen Schätzung, die eine der beiden als Ratgeber gezogenen Personen damals bei dem jeweiligen Städtepaar abgegeben hatte. Die Probanden wurden weiterhin darauf hingewiesen, dass die verschiedenen potentiellen Ratgeber unterschiedlich gut abgeschnitten hatten, weshalb bei der Ziehung der Ratgeber auch jeweils angezeigt würde, welchen Platz sie unter den damaligen 100 Probanden belegt hatten. Durch Drücken der Leertaste konnten die Probanden dann die nächste Seite aufrufen und mit der Ziehung der beiden Ratgeber beginnen.

Die Probanden wurden nun darüber informiert, wie die beiden Ratgeber ausgewählt würden. Sie wurden dabei in dem Glauben gelassen, dass die 100 potentiellen Ratgeber, repräsentiert durch Ihren jeweiligen Rang, also die Zahlen 1 bis 100, im Schnelldurchlauf rotierend angezeigt würden. Durch Druck der Leertaste würde dann diejenige Person als erster Ratgeber ausgewählt, deren Rang zum Zeitpunkt des Tatsendrucks angezeigt wurde. Für die Bestimmung des zweiten Ratgebers würde genauso verfahren, das heißt, der Schnelldurchlauf würde erneut einsetzen und diejenige Person, deren Rang beim erneuten Drücken der Leertaste angezeigt würde, würde als zweiter Ratgeber ausgewählt. Die Ziehung

der beiden Ratgeber erfolgte jedoch nur augenscheinlich nach einem Zufallsprinzip. Unabhängig vom Tastendruck der Probanden wurde als erster Ratgeber stets eine Person mit Rang 7 und als zweiter Ratgeber eine Person mit Rang 73 ermittelt. Über die Rangplätze wurde damit die vermeintliche Kompetenz der Ratgeber manipuliert, wobei der Ratgeber mit Rang 7 einen kompetenten Ratgeber und der Ratgeber mit Rang 73 einen wenig kompetenten Ratgeber darstellt. Für beide Ratgeber wurde außer dem Rangplatz auch ein Name angegeben. Mit dem Hinweis auf die Anonymität der Probanden wurden die Probanden darüber informiert, dass die Namen geändert wurden. Um Störeffekte basierend auf dem Geschlecht der Ratgeber auszuschließen, waren die Namen so gewählt, dass immer gleichgeschlechtliche Ratgeber ausgewählt wurden. Für männliche Probanden erhielt der erste Ratgeber mit Rang 7 den Namen Christian, der zweite Ratgeber mit Rang 73 hingegen den Namen Tobias; für weibliche Probanden wurden die Namen Christine für den ersten Ratgeber mit Rang 7 und Katharina für den zweiten Ratgeber mit Rang 73 gewählt.

Nach vollendeter Ziehung der Ratgeber konnten die Probanden die nächste Seite der Instruktion durch erneutes Drücken der Leertaste aufrufen. Dort wurden sie darüber informiert, nach welchem Schema die einzelnen Durchgänge des Experiments ablaufen würden. Der jeweils erste Schritt bestünde daraus, dass ein Städtepaar angezeigt würde, dessen Entfernung die Probanden schätzen sollten. Diese Schätzung sollten die Probanden direkt eingeben, wobei sie darauf hingewiesen würden, dass für die Eingabe aus Zeitgründen ein Limit von 12 Sekunden bestünde. Nach der Eingabe der ersten Schätzung würde eine kurze Pause folgen, in der je eines von vier möglichen Ereignissen einträte, nämlich:

a) kein Ratschlag; die Probanden wären demnach auf sich allein gestellt und könnten die Pause nutzen, um ihre erste Schätzung zu überdenken,
b) das Erscheinen eines Ratschlag des ersten Ratgebers mit Rang 7,
c) das Erscheinen eines Ratschlag des zweiten Ratgebers mit Rang 73, oder
d) das Erscheinen einer computergenerierten Zufallszahl.

Welches dieser vier Ereignisse jeweils einträte, konnten die Probanden dann nach Eingabe der Initialschätzung sehen. Dort wurden die Probanden dann entweder informiert, sie hätten nun kurz Zeit, ihre Initialschätzung überdenken, oder es wurde eine Zahl angezeigt, sowie deklariert, aus welcher Quelle diese Zahl stammt. Zum Beispiel konnten die Probanden im Falle der Zufallszahl den Text „Eine Zufallszahl ist 1400“ lesen. Wenn statt der

Zufallszahl ein Ratgeber als Quelle genannt wurde, wurde außerdem stets der Rang des Ratgebers in Klammern angegeben, um zu vermeiden, dass Verwechslungen der Ratgeber zu einer Fehlwahrnehmung der vermeintlichen Kompetenz führten. Hinsichtlich der Zufallszahl wurden die Probanden außerdem explizit darauf hingewiesen, dass diese Zahl zwar wie eine Entfernungsschätzung aussehen könnte, jedoch völlig zufällig erstellt wurde und keinerlei Information über die zu schätzende Entfernung enthalte. Nach Ablauf der Pause und gegebenenfalls Darbietung eines Ratschlags bzw. einer Zufallszahl hätten die Probanden dann die Gelegenheit, die Entfernung zwischen den beiden Städten erneut zu schätzen, wofür erneut 12 Sekunden zur Verfügung standen. Sie wurden explizit darauf hingewiesen, dass sie dadurch die Möglichkeit hätten, ihre erste Schätzung zu korrigieren oder, für den Fall, dass sie das nicht wünschten, die erste Eingabe zu wiederholen.

Nach erneutem Drücken der Leertaste wurden die Probanden auf der nächsten Seite der Instruktion darüber informiert, dass sie sich nun im Rahmen von vier Übungstrials mit der Aufgabe vertraut machen könnten. Durch Drücken der Leertaste konnten die Probanden den Probedurchgang mit vier Übungsaufgaben starten. Jeder der vier oben genannten Situationen trat in einem der vier Probedurchgänge auf, wobei die Reihenfolge der Ereignisse sowie die Abweichung des jeweiligen Ratschlags bzw. der Zufallszahl von der Eingabe der Probanden einem vorher festgelegten Schema folgte und für alle Probanden identisch war. Nach Ende der vierten Probeaufgabe wurden die Probanden gebeten, auf den Versuchsleiter zu warten, der dann den Hauptteil des Experiments starten würde.

Der Hauptteil des Experiments wurde durch den Versuchsleiter mittels Drücken der Leertaste gestartet und bestand aus 100 Entfernungsschätzungen, die dem oben beschriebenen Schema folgten. Nach Bearbeitung der letzten Entfernungsschätzung wurden die Probanden auf separaten Bildschirmseiten gebeten, vier abschließende Fragen zu beantworten. Die erste Frage bat die Probanden im Sinne einer Verdachtskontrolle anzugeben, worum es ihrer Ansicht nach in dem Experiment gegangen sei. Die beiden folgenden Fragen bestanden daraus, den jeweiligen Rangplatz der beiden Ratgeber anzugeben, und dienten somit als Manipulationskontrolle für die Kompetenz der Ratgeber. In der abschließenden Frage sollten die Probanden dann angeben, ob sie bereits früher an einem Experiment teilgenommen hatten, in dem sie einzeln oder als Teil einer Gruppe Entfernungen schätzen sollten. Diese ab-

schließende Frage diente zur Kontrolle möglicher Vorverfahren der Probanden mit dem verwendeten Aufgabentyp.

Nach Beantworten der Abschlussfragen war das Experiment beendet, und das Programm bat die Probanden, sich an den Versuchsleiter zu wenden, der den Probanden für die Teilnahme am Experiment dankte, ihnen das Versuchspersonenhonorar von 5 Euro auszahlte und sie über das Ziel des Experiments aufklärte.

4.4 Ablauf der einzelnen Durchgänge

Alle 100 Durchgänge des Experiments folgten dem gleichen Schema. Zunächst wurde das zu schätzende Städtepaar angezeigt. Nach drei Sekunden erfolgte die Aufforderung, die Initialschätzung einzugeben. Je nach Bedingung wurde dann ein Ratschlag bzw. eine Zufallszahl oder die Aufforderung, die erste Schätzung zu überdenken, angezeigt. Wiederum nach 3 Sekunden folgte die Aufforderung, die Finalschätzung einzugeben. Für den Fall, dass die Probanden bei der Eingabe einer Schätzung länger als 12 Sekunden benötigten, wurde die Eingabe automatisch beendet, um eine Begrenzung der Länge des Experiments zu erreichen (hätte ein Proband in jedem Durchgang die 12-Sekunden-Grenze voll ausgenutzt, so hätte ein Dauer des Hauptteils von genau 50 Minuten resultiert). Das automatische Beenden wirkte dabei wie das Betätigen der Eingabetaste, d.h. diejenigen Zahlen, die der Proband bis dahin eingegeben hatte, wurden als jeweilige Schätzung eingegeben; unter Umständen konnte dies in Fehleingaben (z.B. 25 km statt intendierten 2500 km) resultieren. Auf dieses Problem wird im Ergebnisteil des Experiments weiter eingegangen.

Welcher Bedingung ein Durchgang des Experiments angehörte, d.h. ob den Probanden nach ihrer ersten Eingabe ein Ratschlag des kompetenten Ratgebers (Rang 7), ein Ratschlag des wenig kompetenten Ratgebers (Rang 73), eine Zufallszahl oder lediglich die Aufforderung, die erste Schätzung zu überdenken (Kontrollbedingung ohne Ratschlag), angezeigt wurde, war durch eine zufällig ermittelte, jedoch für alle Probanden identische Reihenfolge festgelegt. Die Ratschläge sowie die Zufallszahl wurden dabei während des jeweiligen Durchgangs auf Basis der Initialschätzung der Probanden berechnet, und zwar als prozentuale Abweichung. Da Distanzschätzungen, die bis auf die letzte Ziffer ausformuliert sind (z.B. 1444 km), erfahrungsgemäß in studentischen Stichproben selten vorkommen und auch als eher seltsam empfunden werden, wurden alle angezeigten Zahlen auf die Zehnerstelle gerundet (z.B. 1440 statt 1444 km). Die prozentualen Abweichungen der Ratschläge von der

jeweiligen Initialschätzung des Probanden lagen für alle drei Bedingungen, in denen Zahlen angezeigt wurden, im Bereich zwischen ±16% und ±40%. Genauer wurde der Bereich zwischen ±16% und ±40% in 2%-Schritte unterteilt, wodurch insgesamt 26 Abstufungen entstanden. Da das Experiment pro Bedingung jedoch nur 25 Durchgänge hatte, wurde per Zufall eine der 26 Abweichungen, nämlich die Abweichung von -22% ausgeschlossen. Die restlichen 25 Abweichungswerte kamen demnach in allen drei Bedingungen genau einmal vor.

Bei der Berechnung der Abweichungen wurde ein Sicherungsmechanismus eingebaut, um zu verhindern, dass die Probanden nach einer Fehleingabe bemerken, dass die Ratschläge bzw. Zufallszahl lediglich prozentuale Abweichungen ihrer Initialschätzung waren. Kritisch wären zum Beispiel solche Durchgänge gewesen, in denen Probanden versehentlich sehr geringe Werte eingeben. Wäre dann auf die Eingabe 25km ein Ratschlag bzw. eine Zufallszahl von 20 oder 30 gefolgt, so hätten die Probanden schnell das Zustandekommen der entsprechenden Ratschläge durchschauen können. Um derartige Fälle zu vermeiden, wurde immer dann, wenn der angezeigte Ratschlag kleiner als 200km oder größer als 7500km gewesen wäre, anstelle einer prozentualen Abweichung ein zufällig generierter Wert zwischen 500km und 4000km als Ratschlag bzw. Zufallszahl angegeben. Solche Durchgänge wurden dann bei den späteren Analysen nicht einbezogen (siehe Abschnitt 4.5.1).

Wichtig ist ferner, dass zwischen den drei Bedingungen, in denen eine Zahl angezeigt wurde, über das Experiment hinweg die Abweichungen exakt identisch waren und sich nur in der jeweiligen Reihenfolge unterschieden. Dabei wurde für jeden der beiden Ratgeber sowie für die Zufallszahl jeweils eine randomisierte Reihenfolge der prozentualen Abweichungen erstellt, die dann aber für alle Probanden identisch war. In der Summe lässt sich also festhalten, dass das Stimulusmaterial in Form der angezeigten Zahlen im Mittel für alle drei Experimentalbedingungen (beide Ratgeber sowie die Zufallszahl) identisch war; diese Bedingungen unterschieden sich also lediglich in der subjektiven Überzeugung der Probanden, einen guten Ratschlag, einen schlechten Ratschlag oder eine Zufallszahl erhalten zu haben.

4.5 Ergebnisse

4.5.1 Berechnung des Advice Taking und Überprüfung möglicher Störvariablen

Zunächst wurde überprüft, ob Alter und Geschlecht einen Einfluss auf die Gewichtung der Ratgeber oder der Zufallszahl hatten. Zu diesem Zweck wurde zunächst die Kenn-

zahl AT (siehe Kapitel 2.2) für alle 75 Trials berechnet, in denen die Probanden entweder einen Ratschlag oder eine Zufallszahl erhalten hatten. Für jeden Probanden wurden dann alle 75 AT-Werte darauf geprüft, ob sie einen bestimmten Bereich über- oder unterschritten. Extrem große positive und negative Werte von AT deuten darauf hin, dass entweder bei der Initial- oder bei der Finalschätzung eine Fehleingabe vorlag, so dass der zugehörige AT-Wert nicht sinnvoll interpretierbar ist. Als Grenze für den Einschluss von AT-Werten als gültig wurde der Bereich zwischen -1,5 und +1,5 gewählt[5]. Basierend auf der Beobachtung, dass AT-Werte oder Vergleichbare Maße in der Literatur fast ausschließlich im Bereich zwischen 0 und 1 liegen (Bonaccio & Dalal, 2006), sollte bei der Wahl des Intervalls zwischen -1,5 und 1,5 sichergestellt sein, dass keine gültigen Eingaben ausgeschlossen werden. Insgesamt wurden durch diese Prozedur 211 von insgesamt 1950 Durchgängen ausgeschlossen, was einem Anteil von 11% entspricht[6]. Auf Basis der verbleibenden Durchgänge wurden dann für jeden Probanden die mittleren AT-Werte des kompetenten Ratgebers, des wenig kompetenten Ratgebers und der Zufallszahl berechnet und mit dem Alter der Probanden korreliert. Das Alter der Probanden hatte keinen Einfluss auf die Gewichtung der Ratgeber und der Zufallszahl, alle $|r(26)| < .19$, alle $p > .37$. Mögliche Effekte des Geschlechts auf die Gewichtung wurden in einer ANOVA mit dem Geschlecht der Probanden als Zwischensubjektfaktor und der Art des Ratgebers als Innersubjektfaktor überprüft. Hier zeigte sich weder ein Haupteffekt des Geschlechts noch eine Interaktion des Geschlechts mit der Art des Ratgebers, beide $F(1, 25) < 1.65$, beide $p > .21$, beide $\eta^2_p < .07$. Es traten also keine signifikanten Unterschiede in der Gewichtung in Abhängigkeit des Geschlechts der Probanden auf. Schließlich wurde noch überprüft, ob die Richtung der Abweichung der Ratschläge oder der Zufallszahlen von der Initialschätzung einen Einfluss auf deren jeweilige Gewichtung hatte. Dies ist wichtig, da aufgrund der Art der Schätzungen ein natürlicher Nullpunkt existiert, der die Anpassung der Schätzungen nach unten begrenzt, während nach oben keine solche Begrenzung vorliegt. Eine ANOVA mit der Art des Ratgebers und der Richtung der Abweichung als Innersubjekt-

[5] In der Tat gibt es in der bisherigen Literatur keine Richtwerte für die Begrenzung des AT, wohl aber für den WOA, bei dem in der Regel alle Werte größer 1 entweder ausgeschlossen oder auf 1 gesetzt werden (z.B. Gino, 2008; Gino et al., 2009). Damit ist aber per definitionem ohne inhaltliche Begründung ausgeschlossen, dass die Finalschätzung über den Ratschlag hinaus angepasst wird. Um diese Möglichkeit sowie Anpassungen entgegen eines Ratschlags zu erlauben, wurden die Grenzen für den Einschluss eines AT-Werts in den hier dargestellten Experimenten deshalb etwas liberaler gehandhabt als im Falle der WOA-Werte.

[6] Häufig liegt der Anteil ausgeschlossener Werte in der Literatur mit 5% bis 6% (z.B: Gino, 2008; Gino et al., 2009) etwas niedriger, was höchstwahrscheinlich auf die Zeitbegrenzung bei der Eingabe der Initial- und Finalschätzung zurückzuführen ist.

faktoren zeigte allerdings weder einen Haupteffekt der Richtung der Abweichung noch eine Interaktion der Richtung mit der Art des Ratgebers, beide $F(1, 25) < 1$, beide $p > .60$. Das Ausmaß der Gewichtung war also unabhängig davon, ob Ratschlag oder Zufallszahl höher oder niedriger als die Initialschätzung lagen.

4.5.2 Gewichtung der Ratschläge und Zufallszahlen

Die mittleren AT-Werte betrugen .50 (*SD* = .24) für den vermeintlich kompetenten Ratgeber, .14 (*SD* = .10) für den vermeintlich wenig kompetenten Ratgeber und .15 (*SD* = .15) für die vermeintliche Zufallszahl (siehe Abbildung 1). Eine ANOVA mit der Art des Ratgebers (kompetenter Ratgeber vs. wenig kompetenter Ratgeber vs. Zufallszahl) als Innersubjektfaktor zeigte einen signifikanten Unterschied der AT-Werte in Abhängigkeit der Art des Ratgebers, $F(2,50) = 56.44$, $p < .001$, $\eta^2_p = .69$. Post-Hoc-Kontraste zeigten, dass der vermeintlich kompetente Ratgeber stärker berücksichtigt wurde als der vermeintliche wenig kompetente Ratgeber einerseits und die Zufallszahl andererseits, beide $t(25) > 7.72$, beide $p < .001$, beide $d > 3.08$, während sich die Gewichtung der Zufallszahl nicht von der des wenig kompetenten Ratgebers unterschied, $t(25) = - 0.33$, $p = .74$. Die Probanden waren also sensitiv für die vermeintliche Qualität der Ratschläge in Abhängigkeit von der Kompetenz bzw. Art des Ratgebers, weshalb die Annahme der Hypothese 1 gerechtfertigt ist.

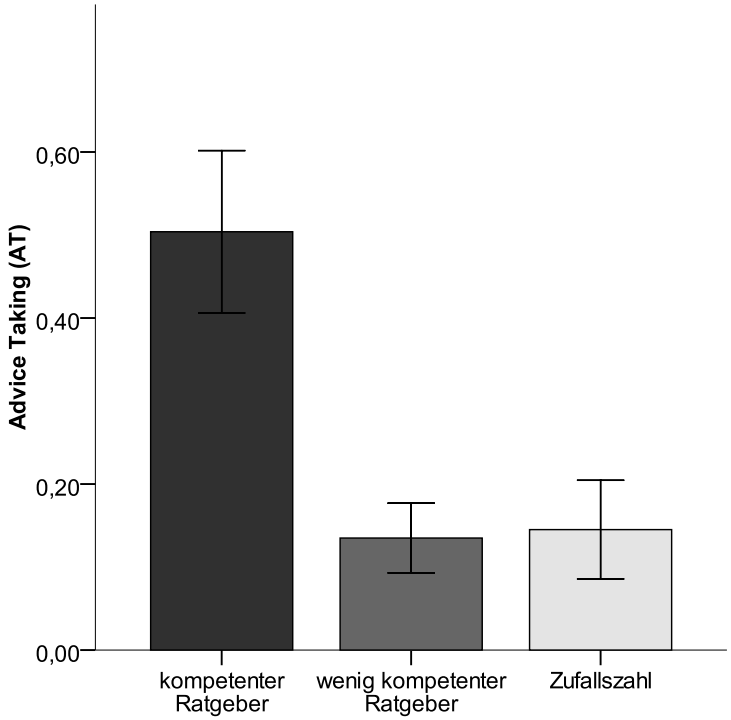


Abbildung 1. Mittlere AT-Werte je nach experimenteller Bedingung in Experiment 1.

Beide vermeintlichen Ratgeber wurden signifikant stärker als Null gewichtet, beide $t(25) > 6.61$, beide $p > .001$, beide $d > 1.29$. Die Gewichtung der vermeintlichen Zufallszahl war ebenfalls signifikant größer als Null, $t(25) = 5.03$, $p < .001$, $d = 0.99$, d.h. die Probanden passten ihre Finalschätzungen systematisch in Richtung der Zufallszahl an. Hypothese 2 wird daher angenommen.

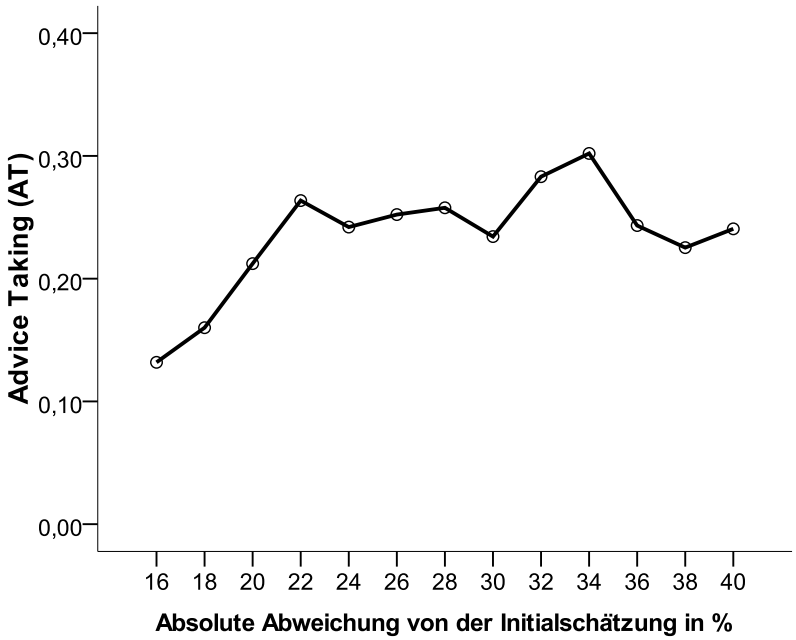
Es besteht nun zumindest theoretisch die Möglichkeit, dass Veränderungen von Initial- zu Finalschätzung, wie sie bei der Zufallszahl beobachtet wurden, auch dann erfolgt wären, wenn den Probanden keinerlei Zahl dargeboten worden wäre. Um dies zu überprüfen, wurde in einer weiteren Analyse neben den drei Experimentalbedingungen die Kontrollbedingung eingeschlossen. Da in der Kontrollbedingung weder Ratschläge noch Zufallszahlen dargeboten wurden, können dort keine AT-Werte berechnet werden. Deshalb wurde eine weitere Kennzahl berechnet, die systematische Abweichungen in allen vier experimentellen Bedingungen abbilden kann, nämlich die mittlere prozentuale Veränderung zwischen Initialschätzung und Finalschätzung. In den Bedingungen mit Ratschlag oder Zufallszahl wurde dabei für jeden Trial die prozentuale Veränderung der Initialschätzung in Richtung der dargebotenen Zahl berechnet. In der Kontrollbedingung wurde in Ermangelung einer vorgegebenen Richtung ein Mittelwert über alle 25 Durchgänge berechnet, bei dem sich unsystema-

tische Schwankungen ausmitteln sollten. Wie auch bei den AT-Werten wurden Grenzwerte für die einzelnen prozentualen Abweichungen definiert, um extrem hohe positive oder negative Werte aufgrund von Fehleingaben herauszufiltern. Als Grenzwert wurde eine prozentuale Veränderung von mehr als 60% nach oben oder unten gewählt, um eine maximale Vergleichbarkeit der Ergebnisse mit den Analysen zum AT zu gewährleisten. Der Wert von 60% ergibt sich daraus, dass in den drei Experimentalbedingungen die Zahlen bis zu 40% von der Initialschätzung abweichen konnten und die Grenze für den AT-Wert bei 1.5 lag. Für jede der vier Bedingungen wurde dann der Mittelwert der gerichteten prozentualen Veränderungen berechnet. Die gerichtete mittlere prozentuale Veränderung der Initialschätzung betrug dabei 14.06 (*SD* = 6.89) in der Bedingung mit kompetentem Ratgeber, 4.54 (*SD* = 3.49) für den wenig kompetenten Ratgeber, 5.30 (*SD* = 5.26) für die Zufallszahl sowie 1.54 (*SD* = 4.41) in der Kontrollbedingung. Zwei *t*-Tests gegen Null zeigten zunächst, dass analog zu den AT-Werten die prozentuale Veränderung der Finalschätzung in Richtung der Zufallszahl signifikant größer als Null war, $t(25) = 5.15$, $p < .001$, $d = 1.01$, während die Veränderung der Schätzungen in der Kontrollbedingung ohne Ratschlag sich nicht statistisch signifikant von Null unterschied, $t(25) = 1.178$, $p = .09$, $d = 0.47$. Entsprechend zeigte auch der Vergleich der beiden Bedingungen, dass die systematische Veränderung der Schätzungen in der Bedingung mit Zufallszahl signifikant stärker ausfiel als in der Kontrollbedingung ohne Ratschlag, $t(25) = 4.43$, $p < .001$, $d = 1.77$. Demnach wird Hypothese 3 angenommen. Die Veränderungen zwischen Initial- und Finalschätzung, die in der Bedingung mit Zufallszahl beobachtet wurden, können damit nicht auf zufällig systematisch wirkende Schwankungen zurückgeführt werden.

4.5.3 Explorative Analyse möglicher Distanz-Effekte auf die Gewichtung

Um die Augenscheinvalidität der Ratschläge zu gewährleisten, wurden unterschiedliche Werte prozentualer Abweichungen von der Initialschätzung als Ratschläge oder Zufallszahlen dargeboten. Dies bietet die Möglichkeit, so genannte Distanz-Effekte zu untersuchen, also den Einfluss der Entfernung zwischen Initialschätzung und Ratschlag auf dessen Gewichtung. Derartige Effekte wurden bisher in nur einer Studie (Yaniv, 2004b) untersucht, wo sich zeigte, dass mit zunehmender Distanz Ratschläge weniger stark gewichtet wurden. Um mögliche Distanzeffekte zu untersuchen, wurden für jeden Ratgebertyp jeweils die AT-Werte derjenigen Durchgänge zu einem Mittelwert zusammengefügt, die gleich weit von der Initialschätzung abwichen, wobei die Richtung der Abweichung vernachlässigt wurde. So wurde

zum Beispiel der Durchgang, in dem die Zufallszahl um 24% über der Initialschätzung lag, mit dem Durchgang zusammengefasst, in dem die Zufallszahl eine prozentuale Abweichung von -24% darstellte. Auf diese Weise ergaben sich für beide Ratgeber sowie die Zufallszahl jeweils 13 Mittelwerte aus den je 25 Durchgängen. Da die Abweichung von 22% nur in einer Richtung vorlag, wurde hier kein Mittelwert gebildet, sondern der AT-Wert des entsprechenden Durchgangs übernommen. Die AT-Werte gingen in eine 3 (*Art des Ratgebers*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 13 (Abweichung von der Initialschätzung: 16%, 18%, 20%, 22%, 24%, 26%, 28%, 30%, 32%, 34%, 36%, 38% sowie 40%) ANOVA mit beiden Faktoren als Innersubjektfaktoren ein. Allerdings fehlte bei 13 der 26 Probanden mindestens einer der Mittelwerte, so dass die Analyse nur auf Basis der verbleibenden 13 Probanden durchgeführt werden konnte. Analog zu den vorigen Analysen zeigte sich erneut der Effekt des Ratgebers, $F(2, 24) = 20.45$, $p < .001$, $\eta^2_p = .63$. Darüber hinaus zeigte sich aber auch ein Effekt der Abweichung, $F(12, 144) = 2.27$, $p = .01$, $\eta^2_p = .16$. Eine genauere Analyse der Art der Unterschiede über Innersubjektkontraste (Trendanalyse) zeigte, dass dieser Haupteffekt am besten durch einen Kontrast zweiter Ordnung erklärt wurde, $F(1, 12) = 15.52$, $p < .01$, $\eta^2_p = .56$ (siehe Abbildung 2)[7].



[7] Da durch einzelne fehlende Werte die Stichprobe für diese Analyse sehr gering war, wurden die fehlenden Werte durch den Mittelwert der Gewichtung des jeweiligen Ratgebers ersetzt und Auf Basis des so vervollständigten Datensatzes die Analyse erneut durchgeführt. Das Befundmuster bleibt dabei identisch.

Abbildung 2. Mittlere AT-Werte in Abhängigkeit der Distanz zwischen Initialschätzung und Ratschlag bzw. Zufallszahl in Experiment 1.

Auffällig war dabei dass insbesondere Ratschläge mit geringer Abweichung weniger stark gewichtet wurden. So betrug der mittlere AT-Wert für Abweichungen von 16% und 18% lediglich .13, und .16. Die Gewichtung für Abweichungen zwischen 20% und 40% waren dagegen sehr ähnlich zueinander und lagen im Mittel zwischen .21 und .30. Der quadratische Kontrast scheint daher abzubilden, dass Ratschläge, die besonders nahe an der Initialschätzung liegen, weniger stark gewichtet wurden, und dass ansonsten kein Distanz-Effekt auftrat. In der Tat verschwindet der Distanz-Effekt vollständig, wenn die beiden geringsten Abweichungen aus der Analyse entfernt werden, $F(10, 120) = 0.66$, $p = .76$, $\eta^2_p = .05$. Die Interaktion zwischen Ratgeber und Abweichung war nicht signifikant, $F(24, 288) = 1.12$, $p = .32$, $\eta^2_p = .09$, was darauf hindeutet, dass der Distanz-Effekt unabhängig von der Art des Ratgebers auftrat.

4.6 Diskussion

Die Ergebnisse des Experiments 1 stellen den erstmaligen eindeutigen Nachweis einer systematischen Übernutzung von Ratschlägen dar. Probanden passten ihre Finalschätzung im Mittel um gut 5% in Richtung der Zufallszahl an, obwohl ihnen explizit mitgeteilt wurde, dass diese Zufallszahlen in keinem Zusammenhang zu der jeweiligen Schätzung stünden. Die normativ rationale Gewichtungsstrategie hätte dagegen darin bestanden, die Zufallszahl vollständig zu ignorieren. Die Anpassung der Schätzungen in Richtung der Zufallszahlen entspricht ferner einer Gewichtung von 15%, ist also nicht unerheblich und vergleichbar mit der Gewichtung des vermeintlich wenig kompetenten Ratgebers (14%).

Jedoch waren die Gewichtungsstrategien der Probanden nicht grundsätzlich irrational. So wurde der kompetente Ratgeber deutlich stärker gewichtet als der vermeintlich wenig kompetente Ratgeber bzw. die Zufallszahl, d.h. die Probanden waren sensitiv für die vermeintliche Qualität der Ratschläge und passten ihre Schätzungen stärker in Richtung eines Ratschlags an, den sie für akkurater hielten. Die Gewichtungen der beiden Ratgeber entsprechen zudem dem üblichen Befundmuster, dass Ratschläge von Experten zu ca. 50% gewichtet werden, während Ratschläge von Laien oder wenig kompetenten Ratgebern nur zu ca. 15 bis 20% berücksichtigt werden (Harvey & Fischer, 1997; Yaniv & Kleiberger, 2000; Yaniv & Milyavsky, 2007). Da das Verhalten der Probanden in Experiment 1 bezüglich der

beiden Ratgeber dem bisher in der Literatur beschrieben Verhalten sehr ähnlich ist, scheint es zumindest unwahrscheinlich, dass die Gewichtung der Zufallszahl auf den gewählten Aufgabentyp oder die Stichprobe zurückzuführen ist. Durch den Vergleich mit der Kontrollbedingung ohne Ratschlag konnte ferner ausgeschlossen werden, dass die Gewichtung der Zufallszahl durch das Zusammenspiel grundsätzlich unsystematischer Veränderungen von Initial- zu Finalschätzung und der Asymmetrie der möglichen Anpassungen bei Schätzaufgaben mit natürlichem Nullpunkt zustande kam. In der Summe konnten also sowohl der Nachweis einer Übernutzung vollkommen invalider Ratschläge erbracht werden, als auch ausgeschlossen werden, dass das Phänomen ein rein artifizielles Resultat der verwendeten Methode ist.

Die Befunde von Experiment 1 führen unmittelbar zu der Frage, welche psychologischen Mechanismen für die Übernutzung von nicht validen Ratschlagen verantwortlich sind. Grundsätzlich kann man hier zwei Arten der Übernutzung nicht valider Ratschläge unterscheiden, nämlich *objektive Übernutzung,* also Übernutzung gemessen am normativ korrekten Grad der Nutzung, und *subjektive Übernutzung*, gemessen an dem, was die jeweilige Person subjektiv als optimalen Grad der Nutzung wahrnimmt. Entsprechend dieser Logik kann eine objektive Übernutzung also erstens dadurch zustande kommen, dass eine Person fälschlicherweise annimmt, ein nicht valider Ratschlag müsse zu einem bestimmten Grad genutzt werden, was im Folgenden als *irrationale Nutzungsüberzeugung* bezeichnet wird, und den Ratschlag dann auch in genau diesem Umfang in ihr Urteil integriert. Dann läge jedoch keine subjektive Übernutzung vor, da die Person ihrer irrationalen Nutzungsüberzeugung nach korrekt gehandelt hat. Zweitens könnte eine objektive Übernutzung vollständig deckungsgleich mit der subjektiven Übernutzung sein, das heißt, die Person erkennt korrekt, dass der nicht valide Ratschlag ignoriert werden sollte, hat also keinerlei irrationale Nutzungsüberzeugung, berücksichtigt den Ratschlag aber dennoch zu einem gewissen Grad. Und schließlich ist eine Kombination aus beidem möglich.

In drei Folgeexperimenten wird nun zunächst der Frage nachgegangen, in wie fern die objektive Übernutzung der Zufallszahlen, die in Experiment 1 beobachtet wurde, auch subjektiv eine Übernutzung darstellt. Mit anderen Worten wird also geprüft, ob die Übernutzung dadurch erklärt werden kann, dass die Probanden – aus welchen objektiv falschen Gründen auch immer – glaubten, dadurch ihre Schätzungen verbessern zu können. Mit anderen Worten könnte also, ähnlich wie im Falle der Multi-Cue-Judgments (Dawes, 1979; Da-

wes et al., 1989), eine Fehlwahrnehmung der Validität von Zufallszahlen ein Grund dafür sein, dass irrationale Nutzungsüberzeugungen auftreten und in der Folge nicht valide Ratschläge systematisch übernutzt werden. Eine solche Fehleinschätzung der Validität ist denkbar, wenn zumindest ein Teil der Probanden einer Fehlkonzeption dessen unterliegt, was genau eine Zufallszahl ist, oder wenn Probanden der Kontrollillusion unterliegen, auch bei zufälligen Zahlen die akkurateren Werte von den weniger akkuraten unterscheiden zu können. Beides sollte bereits zu Beginn des Experiments zu der normativ falschen Überzeugung führen, die Zufallszahlen sollten zu einem gewissen Grad bei der Finalschätzung berücksichtigt werden, um ein möglichst akkurates Ergebnis zu erzielen. Dieser Erklärungsansatz wird in Experiment 2 überprüft, indem die Probanden zu Beginn gebeten werden, anzugeben, wie stark sie die Ratgeber sowie die Zufallszahl gewichten sollten, um möglichst akkurate Finalschätzungen abzugeben.

Es ist nun durchaus denkbar, dass sich die Einschätzung der Validität sich im Verlaufe des Experiments ändert. So kann die Einschätzung der Validität der einzelnen Ratgebertypen dadurch erschwert werden, dass die Probanden abwechselnd mit Ratschlägen von drei vermeintlich unterschiedlich kompetenten Ratgebern konfrontiert sind, wobei die entsprechenden Ratschläge für alle drei Ratgeberarten relativ ähnlich ausfallen. Dies könnte dazu führen, dass die wahrgenommene Validität der drei Ratgeberarten verschwimmt und die Probanden nicht mehr genau differenzieren können, welche Ratschläge sie wie stark integrieren sollten. Ein solcher Effekt ist besonders dann zu erwarten, wenn mehrere Ratgeber eingesetzt werden, weil die Differenzierung der Ratgeber und deren Validität zu einer gewissen Belastung des Arbeitsgedächtnisses führen kann, die sich dann nachteilig auf die korrekte Gewichtung ausübt (Harvey et al., 2000). Experiment 3 untersucht daher, ob die Übernutzung der Ratschläge auch dann auftritt, wenn anstelle eines Innersubjekt-Designs mit allen drei Ratgeberarten ein Zwischensubjekt-Design verwendet wird, bei dem die Art des Ratgebers zwischen den Probanden manipuliert wird und folglich keine Differenzierung der Validitäten verschiedener Ratgeber vorgenommen werden muss.

Drittens besteht noch die Möglichkeit, dass die Probanden die Zufallszahl vorrangig deshalb genutzt haben, weil sie aufgrund der experimentellen Manipulation stets in einem plausiblen Bereich lag, nämlich immer in moderater Entfernung zu den eigenen Initialschätzungen, und damit den vermeintlichen Ratschlägen sehr ähnlich war – so dass die Versuchs-

personen auch bei den Durchgängen mit Zufallszahl nicht aus dem Urteilsmodus „beratenes Urteilen“ aussteigen. Auch in diesem Falle würde sich also im Verlaufe des Experiments die anfängliche Einschätzung der Validität verändern. Experiment 4 untersucht daher, ob das Ausmaß der Übernutzung der Zufallszahlen durch die Schwankungsbreite der Zufallszahlen moderiert wird. Insgesamt decken die Experimente 2 bis 4 damit eine Kategorie von Erklärungsansätzen ab, die die Übernutzung nicht valider Ratschläge auf eine Fehlwahrnehmung der Validität zurückführen. Experiment 2 erfasst dabei das Ausmaß der anfänglichen irrationalen Nutzungsüberzeugung, während die Experimente 3 und 4 das Ziel verfolgen, eine darüber hinausgehende Übernutzung durch eine im Verlauf des Experiments auftretende zusätzliche Fehlwahrnehmung der Validität zu erklären. Gleichzeitig dienen die Experimente natürlich der Überprüfung der Stabilität und auch der Generalisierbarkeit des Effekts der Übernutzung von Zufallszahlen.

5. Experiment 2

5.1 Zielsetzung und Hypothesen

Experiment 2 verfolgt das Ziel, die Ergebnisse des ersten Experiments zu replizieren und gleichzeitig den ersten der drei Erklärungsansätze, die auf eine Fehlwahrnehmung der Validität von Zufallszahlen hindeuten, zu überprüfen. Es soll also in Experiment 2 untersucht werden, in wie weit die Übernutzung von Ratschlägen – wenn auch auf Basis normativ falscher Überzeugungen – intentional erfolgt. Zu diesem Zweck wurde das experimentelle Design von Experiment 1 herangezogen und um die Einschätzung der Probanden hinsichtlich der subjektiv optimalen Gewichtung der beiden Ratgeber sowie der Zufallszahl erweitert. Mittels dieser Abfrage kann vor allem geprüft werden, ob die Probanden korrekt erkennen, dass Zufallszahlen eine Nullvalidität aufweisen und daher nicht zu einer Verbesserung der Schätzungen beitragen können. In diesem Falle müssten die Probanden angeben, dass sie die Zufallszahl gar nicht gewichten würden. Falls jedoch die Probanden der Überzeugung wären, die Zufallszahl sollte verschieden von Null gewichtet werden, so wäre die objektive Übernutzung der Zufallszahl zumindest anteilig auf die Fehlüberzeugung zurückzuführen, die Zufallszahlen sollten zu einem bestimmten Teil berücksichtigt werden. Wenn hingegen zusätzlich noch eine subjektive Übernutzung vorliegt, so sollte die effektive Gewichtung der Zufallszahlen über das Ausmaß der subjektiv als optimal angesehenen Gewichtung hinaus-

gehen. Ob dies der Fall ist, kann man auf zwei Arten prüfen. Zunächst kann über alle Probanden hinweg die subjektiv als optimal angegebene Gewichtung der Zufallszahl mit der tatsächlichen Gewichtung der Zufallszahl verglichen werden. Ein noch härterer Test ist nur dann möglich, wenn eine gewisse Anzahl an Probanden korrekt erkennt, dass die Zufallszahl nicht gewichtet werden sollte. Für diese Subgruppe kann dann direkt getestet werden, ob die tatsächliche Gewichtung der Zufallszahl sich von Null unterscheidet. Sofern die Übernutzung der Zufallszahlen tatsächlich zumindest anteilig durch eine Fehlwahrnehmung der Validität erklärt werden kann, sollten die Probanden im Mittel angeben, die Zufallszahl sollte stärker als Null gewichtet werden, um möglichst akkurate Schätzung zu erhalten. Dies führt zu:

Hypothese 1: Die Probanden geben zu Beginn des Experiments im Mittel an, die Zufallszahl sollte stärker als Null gewichtet werden, um möglichst akkurate Schätzungen zu erreichen. Es liegt also im Mittel eine irrationale Nutzungsüberzeugung vor.

Zunächst muss dann der zentrale Befund aus Experiment 1 repliziert werden, was wie folgt formuliert wird:

Hypothese 2: Personen übernutzen vermeintliche Zufallszahlen, d.h. sie passen ihre Finalschätzung überzufällig in Richtung der Zufallszahl an (Replikation des Befundes aus Experiment 1). Es liegt also objektive Übernutzung vor.

Auch wenn sich die Hypothesen 1 und 2 bestätigen, ist es unwahrscheinlich, dass die Fehlwahrnehmung der Validität das Phänomen der Übernutzung vollständig erklären kann, da verschiede Arbeiten zeigen konnten, dass deutliche Unterschiede zwischen der subjektiv berichteten Gewichtung und der tatsächlichen Gewichtung von Ratgebern auftreten können (Harvey et al., 2000; Gomez-Beldarrain et al., 2004). Es ist also durchaus denkbar, dass auch in gewissem Maße eine subjektive Übernutzung auftreten kann. Insbesondere ist es denkbar, dass auch die Personen, die anfangs korrekt erkannt haben, dass die Zufallszahl ignoriert werden sollte, einer Übernutzung unterliegen. Dies führt zu folgenden Hypothesen:

Hypothese 3: Die effektive Gewichtung der vermeintlichen Zufallszahl übersteigt die von den Probanden anfänglich als optimal eingeschätzte Gewichtung. Es liegt also eine subjektive Übernutzung vor.

Hypothese 4: Probanden, die vor Beginn des Experiments explizit angeben, die Zufallszahlen sollten ignoriert werden, passen ihre Finalschätzung überzufällig in Richtung der Zufallszahlen an.

5.2 Stichprobe und Design

An Experiment 2 nahmen 30 Studierende unterschiedlicher Fachrichtungen der Georg-August-Universität Göttingen teil. Eine Person musste von den Analysen ausgeschlossen werden, da sie in der Mehrzahl der Fälle bei der Finalschätzung den Wert 0 eingegeben hatte. Unter den verbleibenden 29 Probanden waren 14 weibliche Studierende (48%). Das Durchschnittsalter der Teilnehmer lag bei 23,72 Jahren (*SD* = 2,86 Jahre). Experiment 2 folgt ebenso wie Experiment 1 einem einfaktoriellen Innersubjekt-Design mit der Art des Ratschlags als 4-stufigem Faktor, der folgende Ausprägungen aufweist: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl vs. Kontrollbedingung ohne Ratschlag.

5.3 Methode

Ebenso wie Experiment 1 wurde Experiment 2 computergestützt mittels der Software Presentation® (Version 13.0, www.neurobs.com) programmiert und dargeboten (der entsprechende Programmcode für Experiment 2 ist in Dateiform im digitalen Anhang enthalten). Die Vorgehensweise entspricht der von Experiment 1 mit folgender Änderung: Unmittelbar vor den Übungsdurchgängen, also zu einem Zeitpunkt, zu dem die Probanden bereits ausführlich über die Kompetenz der beiden Ratgeber und die Nullvalidität der Zufallszahl informiert worden waren, erfolgte eine Abfrage der Gewichtungsintentionen. Die Probanden wurden dabei gebeten, anzugeben wie stark sie, ihrer Meinung nach, den jeweiligen Ratgeber bzw. die Zufallszahl prozentual gewichten sollten. Um sicherzustellen, dass die Frage nach der prozentualen Gewichtung richtig verstanden wurde, wurde die prozentuale Gewichtung detailliert erläutert. So wurden die Probanden gebeten, Werte zwischen 0 und 100 anzugeben, wobei 0 gleichbedeutend damit wäre, dass der Proband den Ratgeber bzw. die Zufallszahl gar nicht berücksichtigen würde. Ein Wert von 50 würde einer Mittelwertbildung zwischen den eigenen Schätzungen und denen des Ratgebers bzw. der Zufallszahl entsprechen und ein Wert von 100 hieße, dass die Probanden die Ratschläge bzw. die Zufallszahl vollständig übernehmen würden.

5.4 Ablauf der einzelnen Durchgänge

Die einzelnen Durchgänge liefen identisch zu Experiment 1 ab. Die einzige Ausnahme bestand darin, dass die Zeitbeschränkung bei der Eingabe von Initial- und Finalschätzung von 12 Sekunden auf 120 Sekunden angehoben wurde, um den Anteil zeitbedingter Fehleingaben zu reduzieren.

5.5 Ergebnisse

5.5.1 Berechnung des Advice Taking und Überprüfung möglicher Störvariablen

Analog zu Experiment 1 wurde zunächst anhand der mittleren AT-Werte überprüft, wie stark die Probanden im Durchschnitt die präsentierten Ratschläge bzw. Zufallszahlen bei den Finalschätzungen berücksichtigten. Als Grenze für den Einschluss von AT-Werten galt wieder der Bereich zwischen -1,5 und +1,5. Insgesamt wurden durch diese Prozedur 135 von insgesamt 2175 Durchgängen ausgeschlossen, was einem Anteil von 6% entspricht. Diese Ausfallrate ist deutlich geringer als die in Experiment 1 beobachtete und entspricht in etwa den üblichen in der Literatur berichteten Ausfallraten (Gino, 2008; Gino et al., 2009), was die Vermutung zulässt, dass die Erhöhung der Eingabezeit zu weniger Fehleingaben und dadurch zu mehr gültigen Trials führte. Auf Basis der verbleibenden Durchgänge wurden dann für jeden Probanden die mittleren AT-Werte des kompetenten Ratgebers, des wenig kompetenten Ratgebers und der Zufallszahl berechnet.

Diese AT-Werte wurden dann wie bereits in Experiment 1 mit dem Alter der Probanden korreliert, wobei sich erneut kein Effekt des Alters zeigte, alle $|r(29)| < .26$, alle $p > .16$. Mögliche Effekte des Geschlechts auf die Gewichtung wurden wiederum in einer ANOVA mit dem Geschlecht der Probanden als Zwischensubjektfaktor und der Art des Ratgebers als Innersubjektfaktor überprüft. Wie in Experiment zeigte sich kein Haupteffekt des Geschlechts, $F(1,27) = 0.54$, $p = .47$, $\eta^2_p = .02$. Allerdings zeigte sich eine signifikante Interaktion zwischen der Art des Ratgebers und dem Geschlecht der Probanden, $F(2,54) = 4.43$, $p = .02$, $\eta^2_p = .14$. Dieser Effekt kam dadurch zustand, dass weibliche Probanden den vermeintlich kompetenten Ratgeber stärker gewichteten als männliche Probanden, $t(27) = -2.79$, $p = .01$, $d = 1.07$, während sich sowohl bei dem vermeintlich wenig kompetenten Ratgeber als auch bei der Zufallszahl kein Unterschied in der Gewichtung zeigte, beide $|t(27)| < 0.69$, beide $p > .49$, beide $d < 0.27$.

5.5.2 Subjektiv optimale Gewichtung der Ratgeber und der Zufallszahl

Im Mittel gaben die Probanden zu Beginn des Experiments an, der vermeintlich kompetente Ratgeber sollte zu .63 (*SD* = .22) gewichtet werden, der vermeintlich wenig kompetente Ratgeber zu .17 (*SD* = .13) und die Zufallszahl zu .08 (*SD* = .13). Eine ANOVA mit der Art des Ratgebers als Innersubjektfaktor zeigte, dass die Unterschiede in der subjektiv optimalen Gewichtung der drei Ratgeberarten statistisch signifikant waren, $F(2, 56) = 117.90$, $p < .001$, $\eta^2_p = .81$. Einzelvergleiche der drei Ratgeber zeigten, dass die Probanden angaben, den vermeintlich kompetenten Ratgeber stärker gewichten zu wollen als sowohl den vermeintlich wenig kompetenten Ratgeber als auch die Zufallszahl, beide $F(1, 28) > 147.34$, beide $p < .001$, beide $\eta^2_p > .83$. Gleichermaßen gaben sie an, der vermeintlich wenig kompetente Ratgeber solle stärker gewichtet werden als die vermeintliche zufallszahl, $F(1, 28) = 8.91$, $p < .01$, $\eta^2_p = .24$. Damit zeigen die Probanden in Grundzügen rationales Verhalten, da sich ihre subjektiv optimale Gewichtung stark an der vermeintlichen Qualität der Ratschläge orientiert. Bezüglich der Einschätzung der optimalen Gewichtung ist nun interessant, dass die Probanden im Mittel angaben, die Zufallszahl solle zu ca. 8% gewichtet werden. Dieser Wert ist signifikant von dem unter normativer Rationalität zu erwartenden Wert 0 verschieden, $t(28) = 3.24$, $p < .01$, $d = 0.60$, was darauf hindeutet, dass zumindest einige der Probanden ein falsches Verständnis der Validität zufälliger Zahlen hatten. In der Tat gaben nur 17 der 29 Probanden (59%) an, die Zufallszahl solle zu 0% gewichtet werden, während die verbleibenden Probanden für die optimale Gewichtung der Zufallszahl Werte zwischen 2% und 50% angaben. Hypothese 1 wird somit angenommen.

5.5.3 Gewichtung der Ratschläge und Zufallszahlen

Die mittleren AT-Werte betrugen .50 (*SD* = .20) für den vermeintlich kompetenten Ratgeber, .17 (*SD* = .14) für den vermeintlich wenig kompetenten Ratgeber und .21 (*SD* = .23) für die vermeintliche Zufallszahl (siehe Abbildung 3).

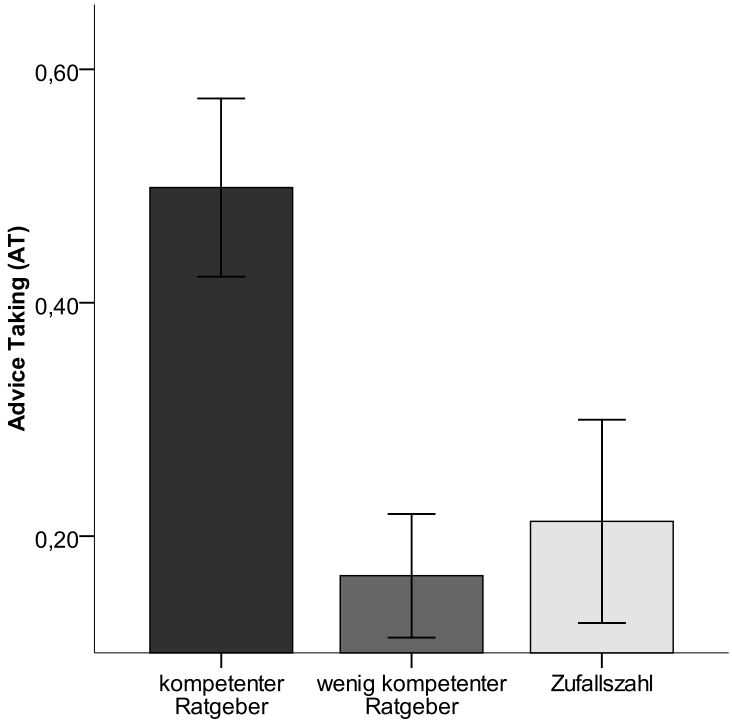


Abbildung 3. Mittlere AT-Werte je nach experimenteller Bedingung in Experiment 2.

Analog zu den Ergebnissen aus Experiment 1 zeigte eine ANOVA mit der Art des Ratgebers als Innersubjektfaktor signifikante Unterschiede bezüglich der Gewichtung der verschiedenen Ratgeber, $F(2, 56) = 29.01$, $p < .001$, $\eta^2_p = .51$, wobei der vermeintlich kompetente Ratgeber stärker gewichtet wurde als der vermeintlich wenig kompetente Ratgeber und sie Zufallszahl, beide $F(1, 28) > 24.19$, beide $p < .001$, beide $\eta^2_p > .46$, während sich die Gewichtung des vermeintlich wenig kompetenten Ratgebers und der Zufallszahl nicht signifikant unterschied, $F(1, 28) = 1.53$, $p = .23$, $\eta^2_p = .05$. Die Gewichtung der vermeintlichen Zufallszahl ist dabei signifikant größer als Null, $t(28) = 5.00$, $p < .001$, $d = 0.93$, d.h. die Probanden passten ihre Finalschätzungen erwartungsgemäß systematisch in Richtung der Zufallszahl an. Dieses Ergebnismuster stellt also eine exakte Replikation der Befunde aus Experiment 1 dar, was zur Annahme von Hypothese 2 führt.

Um zu überprüfen, ob die Übernutzung der Zufallszahlen noch über das Ausmaß hinausging, das durch die anfängliche Fehleinschätzung der Validität erklärt werden kann, wurde zunächst die tatsächliche Gewichtung der Ratgeber gemessen am entsprechenden AT-Wert mit der von den Probanden zu Beginn des Experiments als optimal angegeben Gewichtung verglichen. Eine entsprechende 3 (*Art des Ratgebers*: vermeintlich kompetenter Ratge-

ber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) x 2 (*Art der Gewichtung*: tatsächliche Gewichtung vs. subjektiv optimale Gewichtung) ANOVA mit beiden Faktoren als Innersubjektfaktoren zeigte keinen Haupteffekt der Art der Gewichtung, $F(1, 28) = 0.00$, $p = .99$, aber eine Interaktion zwischen der Art des Ratgebers und der Art der Gewichtung, $F(2, 56) = 14.73$, $p < .001$, $\eta^2_p = .35$. Einzelvergleiche der Art der Gewichtung für die drei Ratgebertypen zeigten, dass der vermeintlich kompetente Ratgeber effektiv weniger stark gewichtet wurde als zu Beginn angegeben ($M = .50$, $SD = .20$ vs. $M = .63$, $SD = .22$), $F(1, 28) = 8.74$, $p < .01$, $\eta^2_p = .24$. Der vermeintlich wenig kompetente Ratgeber wurde so stark gewichtet wie anfänglich angegeben ($M = .16$, $SD = .14$ vs. $M = .17$, $SD = .13$), $F(1, 28) = .03$, $p = .87$, $\eta^2_p = .00$, während die vermeintliche Zufallszahl stärker gewichtet wurde als ursprünglich intendiert ($M = .21$, $SD = .23$ vs. $M = .08$, $SD = .13$), $F(1, 28) = 7.99$, $p < .01$, $\eta^2_p = .22$. Da die Übernutzung der Zufallszahl also über das durch die Fehleinschätzung der Validität erklärbare Ausmaß hinausgeht, wird Hypothese 3 angenommen.

Der noch kritischere Test der Hypothese 4 besteht nun darin, ausschließlich diejenigen Probanden zu betrachten, die explizit angaben, die Zufallszahl solle gar nicht gewichtet werden. Auch für diese Probanden zeigt sich, dass die Finalschätzungen systematisch in Richtung der Zufallszahl angepasst werden. Der entsprechende AT-Wert liegt mit .25 ($SD = .25$) noch geringfügig über dem der Gesamtstichprobe und ist signifikant von Null verschieden, $t(16) = 3.92$, $p < .01$, $d = 0.95$. Es wird also Hypothese 4 angenommen, da selbst die Probanden, die vor Beginn des Experiments im Einklang mit normativen Modellen angeben, eine Zufallszahl sollte gar nicht berücksichtigt werden, ihre Finalschätzungen systematisch an die Zufallszahl anpassten.

Da die Daten von Experiment 2 darauf hinweisen, dass die anfängliche Einschätzung der Probanden, wie stark die jeweiligen Ratschläge gewichtet werden sollten, nicht mit den dann tatsächlich beobachteten Gewichtungen übereinstimmen, wurde zusätzlich überprüft, wie valide die anfänglichen Einschätzungen sind. Für keinen der drei Ratgebertypen zeigte sich hier ein signifikanter Zusammenhang. Während für den vermeintlich kompetenten Ratgeber zumindest deskriptiv der Zusammenhang erwartungskonform ausfiel, $r(29) = .36$, $p = .06$, zeigte sich eine Nullkorrelation zwischen intendierter und tatsächlicher Gewichtung des vermeintlich wenig kompetenten Ratgebers, $r(29) = -.05$, $p = .79$, sowie der vermeintlichen Zufallszahl, $r(29) = .01$, $p = .97$. Dies legt nahe, dass die Probanden entweder nur begrenzt

Einsicht in ihre spätere Gewichtung hatten oder dass sie im Verlauf des Experiments ihre Nutzungsüberzeugung änderten. Da jedoch die Stichprobengröße in Experiment 2 relativ gering war, sollten diese Ergebnisse zunächst nicht überinterpretiert werden.

Analog zu Experiment 1 wurde noch die prozentuale Veränderung zwischen Initial- und Finalschätzung analysiert, um den Vergleich der Bedingungen mit Ratgeber oder zufallszahl mit der Kontrollbedingung ohne Ratschlag zu ermöglichen. Für jede der vier Bedingungen wurde der Mittelwert der gerichteten prozentualen Veränderungen berechnet. Die gerichtete mittlere prozentuale Veränderung der Initialschätzung betrug dabei 13.54 (*SD* = 5.19) in der Bedingung mit kompetentem Ratgeber, 5.48 (*SD* = 4.69) für den wenig kompetenten Ratgeber, 5.57 (*SD* = 6.19) für die Zufallszahl sowie 2.18 (*SD* = 3.08) in der Kontrollbedingung. Die prozentuale Veränderung der Finalschätzung in Richtung der Zufallszahl war dabei signifikant größer als Null, $t(28) = 4.85$, $p < .001$, $d = 0.90$. Die Veränderung der Schätzungen in der Kontrollbedingung ohne Ratschlag las ebenfalls signifikant über Null, $t(28) = 3.806$, $p = .001$, $d = 0.77$, da die Probanden ihre Schätzungen tendenziell stärker nach oben als nach unten anpassten. Dennoch war die Veränderung in der Bedingung mit Zufallszahl wie in Experiment 1 stärker als in der Kontrollbedingung ohne Ratschlag, $t(28) = 2.707$, $p = .01$, $d = 1.02$, so dass auch hier wieder unsystematische Schwankungen als Erklärung für die Übernutzung ausgeschlossen werden können.

5.5.4 Explorative Analyse möglicher Distanz-Effekte auf die Gewichtung

Wie auch in Experiment 1 wurde explorativ untersucht, in wie weit die Entfernung zwischen Initialschätzung und Ratschlag sich auf die Gewichtung des Ratschlags auswirkt. Dazu wurden wieder die AT-Werte der Durchgänge mit derselben Abweichung unter Vernachlässigung der Richtung gemittelt. Diese AT-Werte gingen dann wieder in eine 3 (*Art des Ratgebers*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 13 (Abweichung von der Initialschätzung: 16%, 18%, 20%, 22%, 24%, 26%, 28%, 30%, 32%, 34%, 36%, 38% sowie 40%) ANOVA mit beiden Faktoren als Innersubjektfaktoren ein. Aufgrund fehlender Einzelwerte gingen in diese Analyse nur 25 Probanden ein. Erwartungsgemäß zeigte sich ein Haupteffekt des Ratgebers, $F(2, 48) = 30.24$, $p < .001$, $\eta^2_p = .56$. Anders als in Experiment 1 zeigte sich jedoch kein signifikanter Distanzeffekt, $F(12, 288) = 1.41$, $p = .16$, $\eta^2_p = .06$. Es zeigte sich darüber hinaus zwar noch eine Interaktion der beiden Faktoren, $F(24, 576) = 2.21$, $p < .001$, $\eta^2_p = .08$. Separate Analy-

sen für die drei Arten des Ratgebers zeigten allerdings, dass dieser Effekt schwer interpretierbar ist, da der Haupteffekt der Distanz sich nur für den vermeintlich kompetenten Ratgeber und die vermeintliche Zufallszahl zeigte, $F(12, 300) = 3.28$, $p < .001$, $\eta^2_p = .12$, bzw. $F(12, 312) = 1.95$, $p = .03$, $\eta^2_p = .07$, wobei er in keiner der beiden Bedingungen hinreichend gut durch einen linearen oder quadratischen Kontrast beschrieben werden konnte, alle $F(1, 26) < 3.84$, alle $p > .06$, alle $\eta^2_p < .13$. Für den vermeintlich wenig kompetenten Ratgeber zeigte sich kein Haupteffekt der Distanz, $F(12, 312) = 1.38$, $p = .18$, $\eta^2_p = .05$.[8]

5.6 Diskussion

Die Ergebnisse von Experiment 2 stellen eine Replikation der Ergebnisse von Experiment 1 dar, sowohl bezüglich der Gewichtung der einzelnen Ratgeber als auch bezüglich des kritischen Übernutzungseffekts. Ein nicht valider Ratschlag in Form einer Zufallszahl wurde systematisch in die Finalschätzungen der Probanden einbezogen, wobei die Gewichtung bei ca. 20% und damit in etwa auf dem Niveau eines vermeintlich wenig kompetenten Ratgebers lag. Zum anderen konnte durch Experiment 2 ein Teil der Übernutzung nicht valider Ratschläge erklärt werden, nämlich durch die anfängliche Fehleinschätzung der Validität. So zeigte sich, dass ein Teil der Probanden entgegen der Instruktion und entgegen dem normativen Verständnis von Zufallszahlen der Ansicht war, man solle die Zufallszahl anteilig in das Finalurteil einbeziehen, das heißt, einige der Probanden hatten bereits zu Beginn eine irrationale Nutzungsüberzeugung. Jedoch zeigte sich auch, dass die irrationale Nutzungsüberzeugung das Phänomen der Übernutzung nicht gänzlich erklären kann. So gewichteten die Probanden die Zufallszahlen zum einen im Mittel deutlich stärker als sie zu Beginn des Experiments angegeben hatten; zum anderen trat die systematische Gewichtung der Zufallszahl in unverminderter Stärke auch bei denjenigen Probanden auf, die explizit angegeben hatten, die Zufallszahl solle ignoriert werden. Insgesamt legen die Ergebnisse von Experiment 2 also nahe, dass es sich bei der Übernutzung von Zufallszahlen um einen replizierbaren Befund handelt, der anteilig auf eine anfängliche irrationale Nutzungsüberzeugung von Zufallszahlen zurückgeht, aber auch darüber hinausgeht, wobei die Ursache der zusätzlichen Übernutzung

[8] Ersetzt man wie in Experiment 1 die fehlenden Werte einer Versuchsperson durch deren jeweiligen Mittelwert der Gewichtung des entsprechenden Ratgebers und führt auf Basis der so vervollständigten Datensatzes die Analyse erneut durch, so zeigt sich zwar ein signifikanter Haupteffekt der Distanz, $F(12, 336) = 1.96$, $p = .03$, $\eta^2_p = .07$. Jedoch wird dieser Effekt weder durch einen linearen noch einen quadratischen Kontrast passend erklärt, beide $F(1,28) < 1.88$, beide $p > .18$, beide $\eta^2_p < .07$. Das übrige Befundmuster ist der Analyse ohne Ergänzung fehlender Werte vergleichbar.

noch zu klären wäre. Denkbar wäre dabei, dass es sich dabei um eine subjektive Übernutzung handelt. Allerdings besteht auch die Möglichkeit, dass sich die Nutzungsüberzeugung der Probanden im Verlaufe des Experiments ändert und die Probanden die zufallszahl dann gemäß ihrer der veränderten Nutzungsüberzeugung gewichten.

6. Experiment 3

6.1 Zielsetzung und Hypothesen

Bisher wurde gezeigt, dass eine objektive Übernutzung der Zufallszahlen eintrat, die zu einem gewissen Teil auf anfänglichen irrationalen Nutzungsüberzeugungen basiert, jedoch über das Maß der anfänglichen Nutzungsüberzeugung hinausgeht. Die Ursache für diese zusätzliche Übernutzung könnte sowohl in einer subjektiven Übernutzung als auch in einer Veränderung der Nutzungsüberzeugung liegen. Letzteres scheint vor allem deshalb plausibel, weil die Probanden zu dem Zeitpunkt, zu dem sie ihre anfängliche Nutzungsüberzeugung angeben, noch gar nicht mit den vermeintlichen Ratschlägen und Zufallszahlen konfrontiert waren. Zudem wäre diese Argumentation gut vereinbar mit dem Befund, dass die anfänglich intendierte Gewichtung der Zufallszahl nicht systematisch mit der tatsächlich beobachteten Gewichtung der Zufallszahl zusammenhing. Es könnte also sein, dass die ohnehin schon falsche Einschätzung der Validität im Sinne der anfänglichen irrationalen Nutzungsüberzeugung sich im Verlauf des Experiments noch verstärkt, und zwar durch einen Spill-Over-Effekt, aufgrund dessen die Probanden nur noch bedingt in der Lage waren, zwischen validen und nicht validen Ratschlägen zu differenzieren. Da die vermeintliche Zufallszahl nämlich immer ähnliche Werte annahm wie die Ratschläge des vermeintlich kompetenten und des vermeintlich wenig kompetenten Ratgebers, könnten die Probanden durchaus den Eindruck erhalten haben, dass die Zufallszahlen auch ähnlich valide sein müssten.

Ein solcher Effekt sollte aber spezifisch für das bisher verwendete Innersubjekt-Design sein, bei dem neben den Zufallszahlen stets auch vermeintlich valide Ratschläge dargeboten wurden. Experiment 3 hat daher zum Ziel, diesen Erklärungsansatz zu überprüfen, indem anstelle eines Innersubjekt-Designs ein Zwischensubjekt-Design verwendet wird. Außerdem kann dadurch sichergestellt werden, dass die Übernutzung nicht valider Ratschläge unabhängig von der Art des verwendeten experimentellen Designs auftritt. Wenn sich im Zwischensubjektdesign also ebenfalls eine objektive Übernutzung der Zufallszahl zeigen

lässt, dann kann davon ausgegangen werden, dass es sich bei der Übernutzung von Zufallszahlen tatsächlich um ein genuines und stabiles Phänomen handelt, das auch dann auftreten kann, wenn Personen ausschließlich mit nicht validen Ratschlägen konfrontiert werden. Gleichermaßen kann überprüft werden, ob auch im Zwischensubjektdesign das Ausmaß der Übernutzung über die anfänglich angebende irrationale Nutzungsüberzeugung hinausgeht. Sofern dies der Fall ist, kann der Spill-Over-Effekt als Erklärung für die zusätzliche Übernutzung ausgeschlossen werden. Für Experiment 3 werden folgende Hypothesen formuliert:

Hypothese 1: Die Probanden, denen zwischen Initial- und Finalschätzung eine vermeintliche Zufallszahl dargeboten wird, geben zu Beginn des Experiments im Mittel an, diese Zufallszahl sollte stärker als Null gewichtet werden, um möglichst akkurate Schätzungen zu erreichen. Es liegt also eine irrationale Nutzungsüberzeugung vor (Replikation des Befundes aus Experiment 2).

Hypothese 2: Personen passen ihre Finalschätzung überzufällig in Richtung der Zufallszahl an. Es liegt also eine objektive Übernutzung vor (Replikation der Befunde aus den Experimenten 1 und 2).

Die entscheidende Frage ist nun, ob auch im Zwischensubjekt-Design, in dem ein Spill-Over-Effekt als Erklärung für eine im Verlaufe des Experiments auftretende Fehlwahrnehmung der Zufallszahlen ausgeschlossen ist, eine Übernutzung jenseits der anfänglichen irrationalen Nutzungsüberzeugung auftritt. Wenn der Spill-Over-Effekt die zusätzliche Übernutzung erklären kann, dann sollte das Ausmaß der Übernutzung genau dem Niveau entsprechen, das die Probanden zu Beginn angebe. Tritt hingegen kein Spill-Over-Effekt auf, dann sollte sich auch im Zwischensubjektdesign eine Übernutzung über die anfängliche irrationale Nutzungsüberzeugung hinaus zeigen. Entsprechend werden folgende konkurrierende Hypothesen aufgestellt:

Hypothese 3a: Die effektive Gewichtung der vermeintlichen Zufallszahlen entspricht der von den Probanden als optimal eingeschätzte Gewichtung. Es liegt also keine zusätzliche Übernutzung vor.

Hypothese 3b: Die effektive Gewichtung der vermeintlichen Zufallszahlen übersteigt die von den Probanden als optimal eingeschätzte Gewichtung. Es liegt also eine zusätzliche Übernutzung vor.

Hypothese 4a: Probanden, die vor Beginn des Experiments explizit angeben, die Zufallszahlen sollten ignoriert werden, gewichten die zufallszahl nicht systematisch.

Hypothese 4b: Probanden, die vor Beginn des Experiments explizit angeben, die Zufallszahlen sollten ignoriert werden, passen ihre Finalschätzung überzufällig in Richtung der Zufallszahlen an.

6.2 Stichprobe und Design

An Experiment 3 nahmen 105 Studierende unterschiedlicher Fachrichtungen der Georg-August-Universität Göttingen teil. Vier Personen mussten von den Analysen ausgeschlossen werden, davon drei Personen, weil sie in der Mehrzahl der Fälle bei der Initialschätzung und/oder der Finalschätzung den Wert 0 eingegeben hatten, und eine Person, weil sie sämtlichen Schätzungen Werte zwischen 20.000 und 700.000 Kilometern eingegeben hatte. Diese hohen Eingaben führten dazu, dass bei jedem Durchgang der Sicherungsmechanismus für Extremwerte aktiviert wurde, der dann jeweils per Zufall eine Zahl zwischen 500 und 4.000 als Ratschlag oder Zufallszahl anzeigte. Entsprechend entsprachen die angezeigten Zahlen für diese Person nicht mehr den prozentualen Abweichungen, weshalb keine Vergleichbarkeit zu den übrigen Probanden mehr gegeben war. Unter den verbleibenden 101 Probanden waren 71 weibliche Studierende (70%). Das Durchschnittsalter der Teilnehmer lag bei 22,05 Jahren (*SD* = 2,67 Jahre). Experiment 3 folgt einem einfaktoriellen Design mit der Art des Ratschlags (vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) als Zwischensubjektfaktor.

6.3 Methode

Ebenso wie die vorigen Experimente wurde Experiment 3 computergestützt mittels der Software Presentation® (Version 13.0, www.neurobs.com) programmiert und dargeboten (der entsprechende Programmcode für Experiment 3 ist im digitalen Anhang enthalten). Die Vorgehensweise entspricht der von Experiment 2 mit folgenden beiden Änderungen: die Art des Ratgebers wurde zwischen den Probanden manipuliert, das heißt jeder Proband erhielt in sämtlichen Durchgängen Ratschläge aus derselben Quelle, also entweder in jedem Durchgang einen Ratschlag des vermeintlich kompetenten Ratgebers, in allen Durchgängen Ratschläge des vermeintlich wenig kompetenten Ratgebers oder in allen Durchgängen vermeintliche Zufallszahlen. Die Zuweisung der Probanden zu der jeweiligen experimentellen

Bedingung erfolgte zufällig. In keiner der drei Bedingungen wurden die Probanden darüber informiert, welche möglichen Ratgeber es gab, um jegliche Arten von komparativen Urteilen über die Validität der Ratschläge auszuschließen. Zweitens wurde die Anzahl der Durchgänge von 25 auf 30 erhöht. Die prozentualen Abweichungen lagen wie in den vorigen Experimenten im Bereich zwischen ± 40%.

6.4 Ablauf der einzelnen Durchgänge

Abgesehen davon, dass in Experiment 3 nur noch jeweils eine Art von Ratschlag dargeboten wurde und das Experiment damit auf 30 Durchgänge beschränkt war, liefen alle Durchgänge identisch zu Experiment 2 ab.

6.5 Ergebnisse

6.5.1 Berechnung des Advice Taking und Überprüfung möglicher Störvariablen

Zunächst wurden in allen drei experimentellen Bedingungen die mittleren AT-Werte über alle 30 Durchgänge berechnet. Wie auch in den vorherigen Experimenten wurden nur Durchgänge berücksichtigt, in denen der AT-Wert zwischen -1,5 und +1,5 lag, wodurch 212 der insgesamt 3030 Durchgänge (7%) von der Mittelwertbildung ausgeschlossen wurden. Zunächst wurde über alle Probanden hinweg der AT-Wert mit dem Alter korreliert, wobei sich kein signifikanter Zusammenhang zeigte, $r(101) = .10$, $p = .30$. Mögliche Geschlechtseffekte auf die Gewichtung der Ratschläge wurden in einer ANOVA mit der Art des Ratgebers und dem Geschlecht als Zwischensubjektfaktoren überprüft. Weder Haupteffekt des Geschlechts noch die Interaktion mit der Art des Ratgebers waren statistisch signifikant, $F(1, 95) = 0.78$, $p = .78$, $\eta^2_p = .00$, bzw. $F(2, 95) = 1.07$, $p = .35$, $\eta^2_p = .02$.

6.5.2 Subjektiv optimale Gewichtung der Ratgeber und der Zufallszahl

Zunächst wurde die von den Probanden zu Beginn des Experiments als optimal angegebene Gewichtung analysiert. Im ersten Schritt wurde dabei die Validität der anfänglichen Einschätzungen bezüglich der späteren tatsächlichen Gewichtung für jeden Ratgebertyp untersucht. Deutlicher als in Experiment 2 zeigte sich, dass diejenigen Probanden, die den vermeintlich kompetenten Ratgeber anfänglich stärker gewichten wollten, diesen im Verlauf des Experiments auch tatsächlich stärker gewichteten, $r(33) = .65$, $p < .001$. Gleiches zeigte sich für die vermeintliche Zufallszahl, $r(32) = .59$, $p < .001$. Es hat also den Anschein, dass die anfängliche berichtete intendierte Gewichtung für diese beiden Ratgebertypen zumindest in

gewissem Maße abbildet, wie stark die entsprechenden Ratschläge später gewichtet werden, und dass die Probanden nicht völlig ihren anfänglichen Intentionen zuwider handelten. Lediglich für den vermeintlich wenig kompetenten Ratgeber ergab sich, wie in Experiment 2, kein signifikanter Zusammenhang, $r(35) = .25$, $p = .15$.

Eine ANOVA mit der Art des Ratgebers als Zwischensubjektfaktor zeigte einen Effekt des Ratgebers auf die intendierte Gewichtung, $F(1, 97) = 27.58$, $p < .001$, $\eta^2_p = .36$. Post-hoc-Kontraste zeigten, dass die subjektiv optimale Gewichtung bei Probanden, die glaubten, Ratschläge von einem vermeintlich kompetenten Ratgeber zu erhalten, höher ausfiel als bei den Probanden, die glaubten, Ratschläge eines vermeintlich wenig kompetenten Ratgebers ($M = .52$, $SD = .22$ vs. $M = .25$, $SD = .23$) oder Zufallszahlen ($M = .52$, $SD = .22$ vs. $M = .16$, $SD = .17$) zu erhalten, beide $t(97) > 5.53$, beide $p < .001$, beide $d > 1.12$. Die anfänglich als optimal angegebe Gewichtung unterschied sich hingegen nicht signifikant zwischen der Bedingung mit vermeintlich wenig kompetentem Ratgeber und der Bedingung mit vermeintlicher Zufallszahl, obwohl rein deskriptiv die subjektiv optimale Gewichtung des vermeintlich wenig kompetenten Ratgeber höher ausfiel ($M = .25$, $SD = .23$ vs. $M = .16$, $SD = .17$), $t(97) = 1.68$, $p = .10$, $d = 0.34$. Die subjektiv optimale Gewichtung der vermeintlichen Zufallszahl lag dabei wie beiden vorherigen Experimenten signifikant über Null, $t(32) = 5.35$, $p < .001$, $d = 0.95$. Es lag also auch in Experiment 3 wieder im Mittel eine irrationale Nutzungsüberzeugung bezüglich der vermeintlichen Zufallszahlen vor, weshalb Hypothese 1 angenommen wird.

6.5.3 Gewichtung der Ratschläge und Zufallszahlen

Eine ANOVA mit der Art des Ratgebers als Zwischensubjektfaktor und der Gewichtung des Ratschlags, gemessen am mittleren AT-Wert, als abhängige Variable ergab einen Haupteffekt für die Art des Ratgebers, $F(2, 98) = 9.49$, $p < .001$, $\eta^2_p = .16$. Post-hoc-Kontraste zeigten ferner, dass der vermeintlich kompetente Ratgeber mit einem AT-Wert von .36 ($SD = .21$) stärker gewichtet wurde als sowohl der vermeintlich wenig kompetente Ratgeber mit einem AT-Wert von .19 ($SD = .11$) als auch die Zufallszahl mit einem AT-Wert von .21 ($SD = .21$), beide $t(98) > 3.48$, beide $p < .001$, beide $d > 0.70$. Zwischen der Gewichtung des vermeintlich wenig kompetenten Ratgebers und der Zufallszahl zeigte sich konsistent mit den Ergebnissen der Experimente 1 und 2 kein Unterschied, $t(98) = -0.50$, $p = .62$, $d = 0.10$ (siehe Abbildung 4).

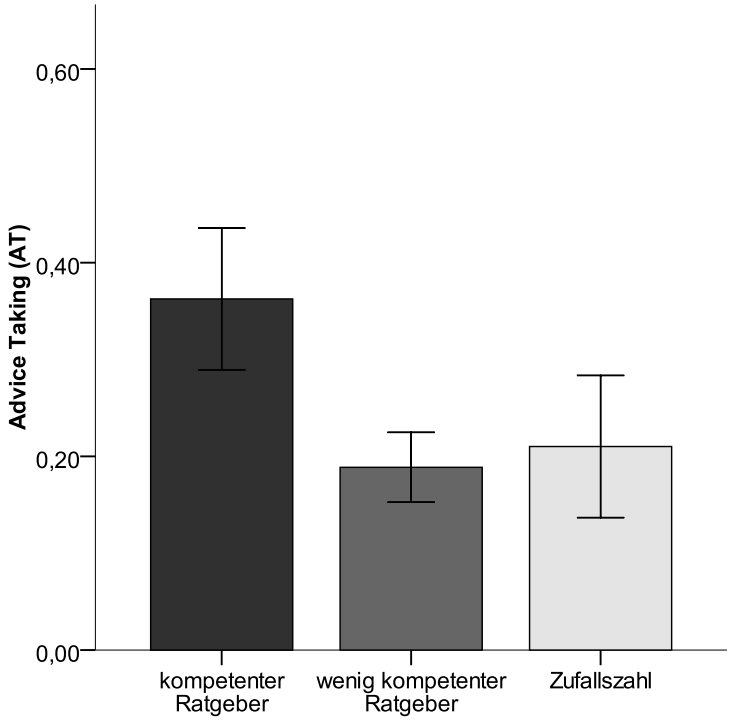


Abbildung 4. Mittlere AT-Werte je nach experimenteller Bedingung in Experiment 3.

Ein *t*-Test gegen Null in der Bedingung mit Zufallszahl zeigte weiterhin, dass die Zufallszahl systematisch gewichtet wurde, $t(32) = 5.83$, $p < .001$, $d = 1.02$. Es lag also eine objektive Übernutzung der Zufallszahl vor. Insgesamt entsprechen die Befunde des Zwischensubjektdesigns damit denen des im Rahmen der Experimente 1 und 2 eingesetzten Innersubjektdesigns, weshalb Hypothese 2 angenommen wird.

6.5.4 Subjektiv optimale Gewichtung der Ratgeber und der Zufallszahl

Als nächstes wurde die tatsächliche Gewichtung der Zufallszahl, gemessen am entsprechenden AT-Wert, mit der von den Probanden zu Beginn des Experiments als optimal angegeben Gewichtung verglichen. Es zeigte sich dabei, dass die tatsächliche Gewichtung sich nicht signifikant von der subjektiv als optimal berichteten unterschied, $t(31) = 1.59$, $p = .12$, $d = 0.57$, obgleich sich deskriptiv eine stärkere tatsächliche Nutzung zeigt ($M = .21$, $SD = .21$ vs. $M = .16$, $SD = .17$). Der beobachtete Mittelwertunterschied ist dabei auf der einen Seite zu klein, um ihn mit der relativ geringen Stichprobengröße absichern zu können. Gleichzeitig ist er aber zu groß, um guten Gewissens die Nullhypothese beizubehalten, dass

keine subjektive Übernutzung vorliegt.[9] In der Summe fällt also der Mittelwertunterschied so aus, dass weder die Annahme der Hypothese 3a noch die der Hypothese 3b gerechtfertigt wäre. Basierend auf der beobachteten Effektstärke von d = .57 liegt jedoch die Vermutung nahe, dass das Ausbleiben eines signifikanten Unterschieds eine Frage der Stichprobengröße ist. Um dennoch ein Antwort auf die Frage zu liefern, ob eine Übernutzung jenseits der anfänglichen irrationalen Nutzungsüberzeugung auch im Zwischensubjekt-Design nachweisbar ist, wurden im folgenden Schritt wurden selektiv diejenigen Probanden (N = 13) betrachtet, die anfangs angaben, die Zufallszahl solle ignoriert werden. Ein t-Test gegen Null zeigte im Einklang mit den Ergebnissen des Experiments 2, dass diese Probanden ihre Finalschätzungen systematisch an die Zufallszahlen anpassten (M = .12, SD = .15), $t(12) = 2.83$, $p < .05$, d =0.79. Da also für diese Probanden eindeutig eine zusätzliche Übernutzung nachgewiesen werden konnte, wird Hypothese 4b angenommen.

6.5.5 Explorative Analyse möglicher Distanz-Effekte auf die Gewichtung

Erneut wurden mögliche Distanzeffekte in einer explorativen Analyse untersucht. Die AT-Werte wurden analog zu den Experimenten 1 und 2 zu 13 Mittelwerten zusammengefasst. Aufgrund einzelner fehlender Werte gingen nur 85 Probanden in die Analyse ein. Eine 3 (*Art des Ratgebers*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 13 (Abweichung von der Initialschätzung: 16%, 18%, 20%, 22%, 24%, 26%, 28%, 30%, 32%, 34%, 36%, 38% sowie 40%) ANOVA mit der Art des Ratgebers als Zwischensubjektfaktor und der Distanz als Innersubjektfaktor ergab einen Haupteffekt für die Distanz, $F(12, 984) = 5.49$, $p < .001$, $\eta^2_p = .06$, der am besten durch einen Kontrast erster Ordnung abgebildet wurde, $F(1, 82) = 34.23$, $p < .001$, $\eta^2_p = .29$. Der Haupteffekt des Ratgebers war erwartungsgemäß ebenfalls signifikant, $F(2, 82) = 12.22$, $p < .001$, $\eta^2_p = .23$, jedoch zeigte sich keine Interaktion der beiden Faktoren, $F(24, 984) = 0.79$, $p = .76$, $\eta^2_p = .02$ (siehe Abbildung 5). Analog zu Experiment 1, wurden Ratschläge und Zufallszahlen weniger stark gewichtet, wenn sie nahe an der Initialschätzung lagen. So betrug der AT-Wert der Ratschläge und Zufallszahlen, die nur 16% oder 18% von der Initialschätzung abwichen, im Mittel .20, während die Gewichtung der übrigen Durchgänge im Mittel bei .27 lag. Da der

[9] Diese Argumentation beruht auf der Konvention, Nullhypothesen auf einem höheren Alpha-Fehler-Niveau (in der Regel per Konvention .20 anstelle von .05) zu testen, um den für diese Hypothesen kritischen Beta-Fehler zu reduzieren.

Interaktionseffekt nicht signifikant war, wurde hier auf eine separate Analyse der drei Ratgebertypen verzichtet.[10]

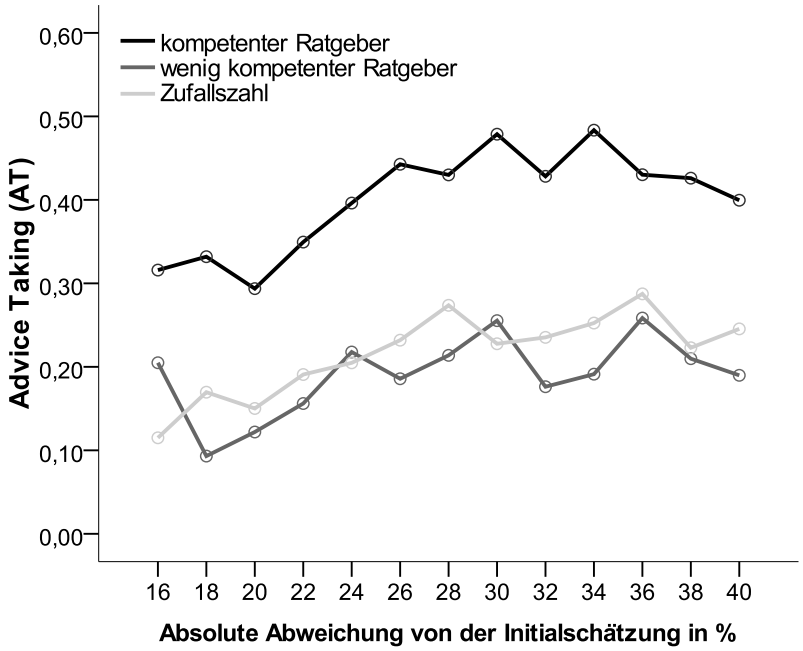


Abbildung 5. Mittlere AT-Werte in Abhängigkeit der Distanz zwischen Initialschätzung und Ratschlag bzw. Zufallszahl nach Art des Ratgebers in Experiment 3.

6.6 Diskussion

Insgesamt zeigen die Ergebnisse von Experiment 3, dass auch dann eine objektive Übernutzung nicht valider Ratschläge auftritt, wenn die Probanden nur Ratschläge aus nur einer Quelle erhalten. Zweitens übertraf das Ausmaß der Übernutzung die anfängliche irrationale Nutzungsüberzeugung. Auch wenn Hypothese 3b nicht statistisch abgesichert werden konnte, spricht das deskriptive Muster gemeinsam mit der Annahme von Hypothese 4b dafür, dass die systematische Übernutzung der Zufallszahlen nicht auf einen Spill-Over-Effekt zurückzuführen ist. Es scheint also für das Auftreten der Übernutzung unerheblich zu sein, ob anstatt den nicht validen Ratschlägen auch in manchen Durchgängen vermeintlich valide Ratschläge dargeboten werden, die phänomenologisch ähnlich erscheinen. Damit kann zum einen der Spill-Over-Effekt als psychologische Erklärung für das Zustandekommen des zusätzlichen Übernutzungseffekts ausgeschlossen werden, zum anderen wurde gezeigt, dass die Übernutzung von Ratschlägen unabhängig von der Art des experimentellen Designs auf-

[10] Auch hier wurden im Anschuss die fehlenden Werte der Probanden durch deren mittlere Gewichtung des Ratgebers ersetzt. Analog zu den vorigen Experimenten blieb das Befundmuster dabei jedoch identisch.

tritt. Auf Basis der Ergebnisse 1 bis 3 lässt sich bisher festhalten, dass Zufallszahlen als Prototyp nicht valider Ratschläge systematisch die Finalschätzungen der Probanden beeinflussen, dass also objektive Übernutzung stattfindet. Diese objektive Übernutzung beruht zum Teil auf einer anfänglichen irrationalen Nutzungsüberzeugung. Jedoch kann diese Nutzungsüberzeugung das Phänomen der Übernutzung von Ratschlägen nicht vollständig erklären. Ein Spill-Over-Effekt, bei dem die Validität der Zufallszahlen aufgrund ihrer Ähnlichkeit zu den vermeintlich echten Ratschlägen zusätzlich überschätzt wird, konnte als Erklärung für die zusätzliche Übernutzung ausgeschlossen werden. Deshalb wird in Experiment 4 ein weiterer Erklärungsmechanismus überprüft, der die zusätzliche Übernutzung im Sinne einer im Laufe des Experiments auftretenden Fehlwahrnehmung der Validität deutet.

7. Experiment 4

7.1 Zielsetzung und Hypothesen

In drei Experimenten konnte nun konsistent nachgewiesen wurde, dass auch Zufallszahlen, also Ratschläge mit einer effektiven Validität von Null, systematisch in die Urteile von Personen einbezogen werden und dass dieser Befund teilweise auf einer anfänglichen irrationalen Nutzungsüberzeugung bezüglich der Zufallszahlen beruht. Es zeigte sich weiterhin eine darüber hinausgehende Übernutzung der Zufallszahlen. Es wäre nun denkbar, dass diese zusätzliche Übernutzung dadurch zustande kommt, dass die Zufallszahlen stets in einem plausiblen Wertebereich schwankten und dadurch eine *Illusion der Validität* entsteht. Eine solche Illusion der Validität kann vor allem dann auftreten, wenn Personen die Validität eines Ratschlags an dessen Plausibilität festmachen.

Bezogen auf die Übernutzung von Zufallszahlen als Ratschläge wäre die durch eine Illusion der Validität überhöhte Nutzungsüberzeugung naturgemäß nicht in dem anfänglichen Urteil widergespiegelt worden, da die Probanden dort ja noch gar nicht mit den vermeintlichen Zufallszahlen konfrontiert waren; es würde sich also um eine erst im Laufe des Experiments entstandene (oder verstärkte) Fehleinschätzung der Validität handeln. Zunächst mag es hier sinnvoll erscheinen, einfach im Verlaufe des Experiments die Gewichtungsintentionen zu erfassen, um eine solche Veränderung zu entdecken. Jedoch besteht dabei das Risiko, dass diese zusätzlichen Abfragen zu Demand-Effekten führen und die Probanden dann verstärkt an ihren Initialschätzungen festhalten. Ebenfalls denkbar wäre, dass durch die Abfrage

der Gewichtung konkrete Implementations-Intentionen (Gollwitzer, 1999; Gollwitzer & Brandstätter, 1997) erzeugt werden. Diese Intentionen könnten dann dazu führen, dass der Zusammenhang zwischen intendierter Gewichtung und tatsächlicher Gewichtung stärker wird. In beiden Fällen würde die Abfrage der Nutzungsüberzeugung also direkt den Grad der Nutzung beeinflussen. Wäre dies der Fall, so böte die wiederholte Abfrage der Nutzungsüberzeugung zwar möglicherweise Potential für eine Interventionsstrategie, würde aber gleichermaßen die Vergleichbarkeit mit den vorherigen Experimenten und die Interpretierbarkeit der Daten gefährden.

Um dennoch zu überprüfen, ob der geringe Schwankungsbereich der Zufallszahl dazu geführt haben könnte, dass die Probanden sie stärker gewichteten als ursprünglich intendiert, wurden in Experiment 4 die Abweichungswerte der Zufallszahl für einen Teil der Probanden so verändert, dass sie gelegentlich gänzlich unplausible Werte annahmen, nämlich besonders niedrige Werte im einstelligen oder zweistelligen Bereich sowie extrem hohe Werte im oberen vierstelligen und fünfstelligen Bereich. Wenn der geringe Schwankungsbereich und die damit verbundenen plausiblen Werte der Zufallszahl tatsächlich eine Rolle bei der Übernutzung nicht valider Ratschläge spielen sollten, dann sollte sich der Übernutzungseffekt durch die Ausprägung der Zufallszahl moderieren lassen. Andernfalls sollte sich zeigen lassen, dass der Schwankungsbereich keinen Einfluss auf die Übernutzung der Zufallszahl, und insbesondere auf die Übernutzung jenseits der anfänglichen irrationalen Nutzungsüberzeugung, ausübt. Insgesamt werden für Experiment 4 zunächst folgende Hypothesen formuliert, die die bisherigen Befunde replizieren:

Hypothese 1: Die Probanden geben zu Beginn des Experiments im Mittel an, die Zufallszahl sollte stärker als Null gewichtet werden, um möglichst akkurate Schätzungen zu erreichen (Replikation der Befunde der Experimente 2 und 3).

Hypothese 2: Personen übernutzen vermeintliche Zufallszahlen, d.h. sie passen ihre Finalschätzung überzufällig in Richtung der Zufallszahl an (Replikation der Befunde der Experimente 1 bis 3).

Um zu bestimmen, ob der Schwankungsbereich der vermeintlichen Zufallszahl über die anfängliche irrationale Nutzungsüberzeugung hinaus zu einer Übernutzung nicht valider Ratschläge führt, werden im Folgenden konkurrierende Hypothesen aufgestellt. Für den Fall,

dass sich tatsächlich ein Teil der Übernutzung über den geringen Schwankungsbereich der Zufallszahlen erklären lässt, sollte sich ein Befundmuster zeigen, das in folgenden Hypothesen abgebildet ist:

Hypothese 3a: Wenn die Zufallszahl Extremwerte annimmt, fällt die objektive Übernutzung geringer aus als wenn die Zufallszahl nur in einem kleinen und plausiblem Wertebereich schwankt.

Hypothese 4a: Wenn die Zufallszahl Extremwerte annimmt, fällt die Übernutzung jenseits der anfänglichen irrationalen Nutzungsüberzeugung geringer aus als wenn die Zufallszahl nur in einem kleinen und plausiblem Wertebereich schwankt.

Hypothese 5a: Probanden, die vor Beginn des Experiments explizit angeben, die Zufallszahlen sollten ignoriert werden, passen ihre Finalschätzung weniger stark in Richtung der Zufallszahl an, wenn diese auch Extremwerte annimmt.

Sollte hingegen die geringe Schwankungsbreite der Zufallszahlen nicht für eine Übernutzung von Zufallszahlen jenseits der anfänglichen irrationalen Nutzungsüberzeugung verantwortlich sein, so sollten sich auch keine Effekte der Schwankungsbreite zeigen. Dies wird durch die folgenden beiden Hypothesen abgebildet:

Hypothese 3b: Ob die Zufallszahl Extremwerte annimmt oder nicht, hat keinen Einfluss auf das Ausmaß der objektiven Übernutzung.

Hypothese 4b: Ob die Zufallszahl Extremwerte annimmt oder nicht, hat keinen Einfluss auf das Ausmaß der Übernutzung jenseits der anfänglichen irrationalen Nutzungsüberzeugung.

Hypothese 5b: Ob die Zufallszahl Extremwerte annimmt oder nicht, hat keinen Einfluss darauf, wie stark Probanden, die vor Beginn des Experiments explizit angeben, die Zufallszahlen sollten ignoriert werden, ihre Finalschätzung in Richtung der Zufallszahl anpassen.

7.2 Stichprobe und Design

An Experiment 4 nahmen 100 Studierende unterschiedlicher Fachrichtungen der Georg-August-Universität Göttingen teil. Eine Person musste von den Analysen ausgeschlossen werden, da sie in der Mehrzahl der Fälle bei der Finalschätzung den Wert 0 eingegeben

hatte. Unter den verbleibenden 99 Probanden waren 75 weibliche Studierende (76%). Das Durchschnittsalter der Teilnehmer lag bei 24,06 Jahren (*SD* = 2,91 Jahre). Experiment 4 folgt einem 3 (*Art des Ratschlags*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 2 (*Ausprägung der Zufallszahl*: mit Extremwerten vs. ohne Extremwerte) Design mit der Art des Ratgebers als Innersubjektfaktor und der Ausprägung der Zufallszahl als Zwischensubjektfaktor.

7.3 Methode

Ebenso wie die vorigen Experimente wurde Experiment 4 computergestützt mittels der Software Presentation® (Version 13.0, www.neurobs.com) programmiert und dargeboten (der entsprechende Programmcode für Experiment 4 ist im digitalen Anhang enthalten). Die Vorgehensweise entspricht der von Experiment 2 mit folgenden beiden Änderungen: erstens wurde in Experiment 4 auf die Kontrollbedingung ohne Ratschlag verzichtet. Die übrigen drei Bedingungen wurden vollständig übernommen und um fünf Durchgänge pro Bedingung erweitert, so dass Experiment 4 pro Ratgeber bzw. Zufallszahl jeweils 30 Durchgänge enthielt. Für die beiden vermeintlichen Ratgeber wurde bei diesen fünf zusätzlichen Durchgängen zweimal die Abweichung von +23% und dreimal die Abweichung von -23% von der Initialschätzung gewählt. Dadurch war einerseits gewährleistet, dass die zusätzlichen Abweichungen im Rahmen der bisherigen Spannbreite lagen. Andererseits war so sichergestellt, dass bezogen auf die nun 30 Durchgänge bei jeweils der Hälfte ein Ratschlag oberhalb der Initialschätzung und in der anderen Hälfte ein Ratschlag unterhalb der Initialschätzung dargeboten wurde.

Zweitens, und verbunden mit der Erweiterung um fünf Durchgänge, wurde die Manipulation der Ausprägung der Zufallszahl als Zwischensubjektfaktor eingeführt. In etwa die Hälfte der Probanden wurde der Bedingung mit Zufallszahl ohne Extremwerte (*N* = 51) sowie Zufallszahl mit Extremwerten (*N* = 48) zugewiesen wurde. Extremwerte der Zufallszahl wurden sowohl in der Übungsaufgabe mit Zufallszahl als auch in den fünf zusätzlichen Durchgängen des eigentlichen Experiments über die Höhe der prozentualen Abweichung von der Initialschätzung zwischen den Bedingungen manipuliert. In der Bedingung ohne Extremwerte wurden dieselben moderaten Abweichungen wie bei den vermeintlichen Ratgebern gewählt. Dadurch ist diese Bedingung vergleichbar mit den Experimenten 1 bis 3. In der Bedingung mit Extremwerten dagegen nahm die Zufallszahl sowohl bei der Beispielaufgabe als auch bei

den fünf zusätzlichen Durchgängen einen Extremwert an. Die exakten Abweichungen von der jeweiligen Initialschätzung betrugen in der Bedingung mit Extremwerten -99% für die Übungsaufgabe mit Zufallszahl (im Gegensatz zu einer Abweichung von -10% in der Bedingung ohne Extremwerte), sowie +450%, -98%, +400%, -95% und +500% für die fünf zusätzlichen Durchgänge des eigentlichen Experiments. In beiden Bedingungen und für jede Art des Ratgebers erschienen die fünf zusätzlichen Durchgänge mit den jeweils beschriebenen Abweichungswerten an 3., 9., 15., 21., und 27. Stelle, so dass im Falle der Zufallszahl mit Extremwerten die Probanden zum frühestmöglichen Zeitpunkt, nämlich bereits bei der Übungsaufgabe, sowie in relativ gleichmäßigen Abständen während des Experiments mit Extremwerten konfrontiert wurden.

7.4 Ablauf der einzelnen Durchgänge

Abgesehen von der Ausweitung auf 30 Durchgänge pro Bedingung und die damit einhergehenden zusätzlichen Abweichungswerte liefen alle Durchgänge identisch zu Experiment 2 ab.

7.5 Ergebnisse

7.5.1 Berechnung des Advice Taking und Überprüfung möglicher Störvariablen

Analog zu den vorigen Experimenten wurden zunächst die mittleren AT-Werte für die einzelnen Ratgeberarten berechnet. Einzelne AT-Werte zwischen -1,5 und +1,5 wurden wie in den Experimenten 1 bis 3 ausgeschlossen, so dass 595 der insgesamt 8910 (7%) Einzelwerte nicht in die Mittelwertbildung eingingen. Die fünf zusätzlichen Durchgänge jeder der drei Ratgeberarten wurden bei der Analyse zunächst außen vor gelassen. Dies war erforderlich, da bei Zufallszahlen mit Extremwerten die entsprechenden AT-Werte nicht mehr mit den übrigen Bedingungen vergleichbar gewesen wären. Der Ausschluss der zusätzlichen Durchgänge hingegen gewährleistete, dass im Mittel das Stimulusmaterial wieder für alle drei Ratgeberarten sowie für beide Varianten der Ausprägung der Zufallszahl identisch war. Ferner ergibt sich so auch die maximale Vergleichbarkeit zu den Experimenten 1 bis 3. Die AT-Werte wurden dann wie in den vorherigen Experimenten mit dem Alter korreliert, das abermals keinen Einfluss auf die Gewichtung der Ratgeber und der Zufallszahl hatte, alle $|r(96)| < .14$, alle $p > .20$. Eine ANOVA mit der Art des Ratgebers als Innersubjektfaktor und der Ausprägung der Zufallszahl sowie dem Geschlecht der Probanden als Zwischensubjekt-

faktor zeigte weder einen signifikanten Haupteffekt des Geschlechts noch eine Interaktion mit der Art des Ratgebers und auch keine Dreifachinteraktion mit der Art des Ratgebers und der Ausprägung der Zufallszahl, alle $F(2, 190) < 2.97$, alle $p > .05$, alle $\eta^2_p < .04$.

7.5.2 Subjektiv optimale Gewichtung der Ratgeber und der Zufallszahl

Zunächst wurde wieder die von den Probanden zu Beginn des Experiments als optimal angegebene Gewichtung analysiert. Im ersten Schritt wurde dabei die Validität der anfänglichen Einschätzungen bezüglich der späteren tatsächlichen Gewichtung für jeden Ratgebertyp untersucht. Deutlicher als in Experiment 2 und analog zu Experiment 3 zeigte sich, dass diejenigen Probanden, die den vermeintlich kompetenten Ratgeber anfänglich stärker gewichten wollten, diesen im Verlauf des Experiments auch tatsächlich stärker gewichteten, $r(99) = .42$, $p < .001$. Gleiches zeigte sich für die vermeintliche Zufallszahl, $r(99) = .36$, $p < .001$. Es hat also erneut den Anschein, dass die anfängliche berichtete intendierte Gewichtung für diese beiden Ratgebertypen zumindest in gewissem Maße abbildet, wie stark die entsprechenden Ratschläge später gewichtet werden, und dass die Probanden nicht völlig ihren anfänglichen Intentionen zuwider handeln. Lediglich für den vermeintlich wenig kompetenten Ratgeber ergab sich, wie in den vorigen Experimenten kein signifikanter Zusammenhang zwischen anfänglich intendierter und tatsächlicher Gewichtung, $r(99) = .15$, $p = .14$.

Eine 3 (*Art des Ratschlags*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 2 (*Ausprägung der Zufallszahl*: mit Extremwerten vs. ohne Extremwerte) ANOVA mit der Art des Ratgebers als Innersubjektfaktor, der Ausprägung der Zufallszahl als Zwischensubjektfaktor und der zu Beginn als optimal angegebenen Gewichtung der Ratschläge als abhängiger Variable zeigte lediglich einen Haupteffekt des Ratgebers, $F(2, 192) = 250.05$, $p < .001$, $\eta^2_p = .72$, während für die Ausprägung der Zufallszahl weder ein Haupteffekt noch eine Interaktion mit der Art des Ratgeber auftraten, beide $F(2, 192) < 0.48$, beide $p > .47$, beide $\eta^2_p < .01$. Wie in den Experimenten 2 und 3 gaben die Probanden zu Beginn an, den vermeintlich kompetenten Ratgeber stärker gewichten zu wollen als den vermeintlich wenig kompetenten ($M = .60$, $SD = .19$ vs. $M = .14$, $SD = .22$) sowie die Zufallszahl ($M = .60$, $SD = .19$ vs. $M = .08$, $SD = .17$), beide $F(1, 96) > 309.10$, beide $p < .001$, beide $\eta^2_p > .72$, und den vermeintlich wenig kompetenten Ratgeber wiederum stärker als die Zufallszahl ($M = .14$, $SD = .22$ vs. $M = .08$, $SD = .17$), $F(1, 96) = 4.05$,

$p < .05$, $\eta^2_p = .44$. Es zeigte sich also auch in Experiment 4, dass die Probanden sensitiv für die erwartete Qualität der Ratschläge waren, wenn sie einschätzten, wie stark diese jeweils in ihr Urteil einbezogen werden sollten. Die anfänglich als optimal angegebene Gewichtung der vermeintlichen Zufallszahl war dabei erneut signifikant von Null verschieden, $t(98) = 4.82$, $p < .001$, $d = .48$. Analog zu den Experimenten 2 und 3 wiesen also auch die Probanden in Experiment 4 zu Beginn des Experiments eine irrationale Nutzungsüberzeugung auf. Hypothese 1 wird daher angenommen.

7.5.3 Gewichtung der Ratschläge und Zufallszahlen

Eine 3 (*Art des Ratschlags*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 2 (*Ausprägung der Zufallszahl*: mit Extremwerten vs. ohne Extremwerte) ANOVA mit der Art des Ratgebers als Innersubjektfaktor, der Ausprägung der Zufallszahl als Zwischensubjektfaktor und der Gewichtung des Ratschlags (gemessen wieder mit dem AT-Wert) als abhängige Variable ergab einen Haupteffekt für die Art des Ratgebers, $F(2, 194) = 199.47$, $p < .001$, $\eta^2_p = .67$. Weder der Haupteffekt der Ausprägung der Zufallszahl noch deren Interaktion mit der Art des Ratgebers waren statistisch signifikant, $F(1, 97) = 2.30$, $p = .13$, $\eta^2_p = .02$, bzw. $F(2, 194) = 1.42$, $p = .24$, $\eta^2_p = .01$. Post-hoc-Kontraste zeigten ferner, dass der vermeintlich kompetente Ratgeber mit einem AT-Wert von .43 ($SD = .20$) stärker gewichtet wurde als sowohl der vermeintlich wenig kompetente Ratgeber mit einem AT-Wert von .12 ($SD = .12$) als auch die Zufallszahl mit einem AT-Wert von .09 ($SD = .11$), beide $F(1, 97) > 210.85$, beide $p < .001$, beide $\eta^2_p > .68$. Konsistent mit den vorigen Befunden zeigte sich kein signifikanter Unterschied hinsichtlich der Gewichtung des vermeintlich wenig kompetenten Ratgebers und der Zufallszahl, $F(1, 97) = 3.62$, $p = .06$, $\eta^2_p = .04$ (siehe Abbildung 6).

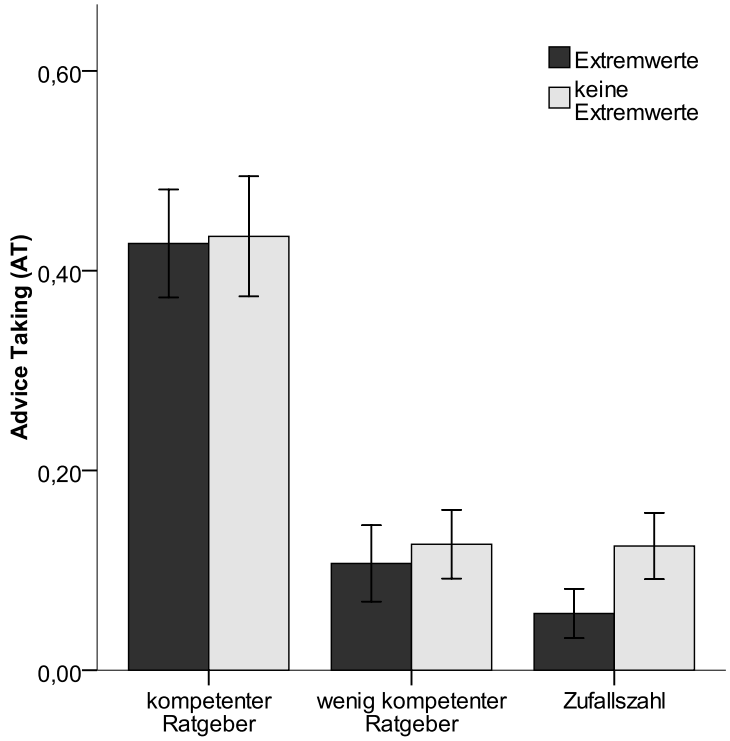


Abbildung 6. Mittlere AT-Werte je nach experimenteller Bedingung in Experiment 3.

Insgesamt zeigt sich also das vertraue Muster, dass vermeintlich kompetente Ratgeber stärker berücksichtigt werden als vermeintlich wenig kompetente Ratgeber und Zufallszahlen, während letztere in etwa gleich stark gewichtet werden. Die Gewichtung der Zufallszahl lag ferner wie auch in den vorigen Experimenten signifikant über Null, $t(98) = 8.41$, $p < .001$, $d = 0.84$, weshalb Hypothese 2 angenommen wird.

7.5.4 Einfluss der Schwankungsbreite der Zufallszahl

Um die konkurrierenden Hypothesen 3 bis 5 zu testen, wurden ausschließlich die Durchgänge berücksichtigt, in denen die Probanden als Ratschlag eine vermeintliche Zufallszahl erhalten hatten. Obwohl in der unter in Abschnitt 7.5.3 berichteten ANOVA weder der Haupteffekt für die Ausprägung der Zufallszahl noch der Interaktionseffekt mit der Art des Ratgebers auf die Gewichtung signifikant waren, zeigte ein Einzelvergleich in Abhängigkeit der Ausprägung der Zufallszahl, dass die Zufallszahl signifikant weniger gewichtet wurde, wenn sie Extremwerte annahm als wenn keine Extremwerte auftraten ($M = .06$, $SD = .09$ vs. $M = .12$, $SD = .12$) , $t(97) = -3.25$, $p < .01$, $d = 0.66$.[11] Separate *t*-Tests gegen Null zeigten, dass die Zufallszahl sowohl in der Bedingung ohne Extremwerte als auch in der Bedingung mit

[11] Dies mag darauf zurückzuführen sein, dass die Varianzen der Gewichtung in der ANOVA größer waren als die des hier berichteten *t*-Tests.

Extremwerten systematisch gewichtet wurde, $t(50) = 7.52$, $p < .001$, $d = 1.05$, bzw. $t(47) = 4.62$, $p < .001$, $d = 0.67$. Die objektive Übernutzung der Zufallszahlen fiel also geringer aus, wenn die Zufallszahlen nicht nur in einem kleinen und plausiblen Bereich schwankten, sondern auch Extremwerte annahmen. Dies spricht zunächst für die Annahme von Hypothese 3a und die Ablehnung der Hypothese 3b.

Als nächstes wurde die tatsächliche Gewichtung der Zufallszahl gemessen am entsprechenden AT-Wert mit der von den Probanden zu Beginn des Experiments als optimal angegeben Gewichtungen der Zufallszahl verglichen. Eine ANOVA mit der Ausprägung der Zufallszahl als Zwischensubjektfaktor und dem Vergleich von subjektiv optimaler und tatsächlicher Gewichtung der Zufallszahl als Innersubjektfaktor zeigte lediglich einen Haupteffekt der Ausprägung der Zufallszahl, $F(1,97) = 4.17$, $p < .05$, $\eta^2_p = .04$. Es zeigte sich jedoch kein Haupteffekt für den Vergleich der subjektiv optimalen Gewichtung mit der tatsächlichen Gewichtung der Zufallszahl ($M = 0.8$, $SD = .17$ vs. $M = .09$, $SD = .11$), $F(1, 97) = 0.38$, $p = .54$, $\eta^2_p = .00$, und auch die Interaktion der beiden Faktoren war nicht signifikant, $F(1, 97) = 1.67$, $p = .20$, $\eta^2_p = .02$. Entsprechend der Vermutung, dass die tatsächliche Gewichtung der Zufallszahlen sich nicht allein durch eine anfängliche irrationale Nutzungsüberzeugung erklären lässt, wäre jedoch zu erwarten gewesen, dass die tatsächliche Gewichtung signifikant stärker ausgeprägt ist als die anfänglich angegebene subjektiv optimale Gewichtung. Mit Hinblick auf die Übernutzung der Zufallszahlen jenseits der anfänglichen irrationalen Nutzungsüberzeugung ließen sich die Ergebnisse aus Experiment 2 also nicht replizieren, da sich kein signifikanter Unterschied zwischen der subjektiv optimalen und der tatsächlichen Gewichtung zeigte. Mit Hinblick auf die Hypothesen 4a und 4b sind die Ergebnisse daher vorerst schwer interpretierbar, da ein Effekt, der nicht auftritt, auch nicht moderiert werden kann. Rein deskriptiv schient das Ergebnismuster eher für die Annahme der Hypothese 4a zu sprechen. So liegt die tatsächliche Gewichtung der Zufallszahl ohne Extremwerte erwartungskonform etwas über der subjektiv als optimal angegebenen ($M = .12$, $SD = .12$ vs. $M = .09$, $SD = .19$), während sich für die Bedingung mit Extremwerten deskriptiv andeutungsweise das gegenteilige Muster zeigt ($M = .06$, $SD = .12$ vs. $M = .07$, $SD = .13$).

Eine genauere Analyse der subjektiv als optimal angegebenen Gewichtung der Probanden liefert die Erklärung für das von Experiment 2 abweichende Befundmuster: Es zeigt sich, dass ein Großteil der Probanden, nämlich 64%, ähnlich wie in den Experimenten 2 und

3, die optimale Gewichtung der Zufallszahl korrekt mit Null angibt, während ein kleinerer Teil der Probanden (8%), extreme und rational nicht erklärbare Werte angibt, z.B. Werte von mindestens ±50%. Diese Extremwerte reichen aber aus, um den Mittelwert der Einschätzungen stark zu verzerren. Es wäre also denkbar, dass das Ausbleiben der Unterschiede zwischen subjektiv optimaler Gewichtung auf der einen Seite und tatsächlicher Gewichtung auf der anderen vor allem darauf zurückzuführen ist, dass aufgrund weniger extremer Ausreißer die subjektiv optimale Gewichtung artifiziell überhöht ist.

Um diese Vermutung zu überprüfen, wurden in einem ersten Schritt alle Probanden ausgeschlossen, deren subjektiv optimale Gewichtung mehr als 2 Standardabweichungen vom Mittelwert entfernt waren. Auf Basis der Standardabweichung von knapp .17 ergab sich ein Ausschluss aus der Analyse für subjektiv optimale Gewichtungen, die größer gleich +42% oder kleiner gleich -26% waren. 9 Personen wurden auf diese Weise von den folgenden Analysen ausgeschlossen, darunter 2 in der Bedingung mit Extremwerten und 7 in der Bedingung ohne Extremwerte; in der Analyse verbleiben daher insgesamt 89 Personen, davon 46 in der Bedingung mit Extremwerten. Auf Basis dieser 89 Personen wurde zunächst wieder eine ANOVA mit der Ausprägung der Zufallszahl als Zwischensubjektfaktor und dem Vergleich von subjektiv optimaler und tatsächlicher Gewichtung der Zufallszahl als Innersubjektfaktor gerechnet. Erwartungsgemäß zeigten sich sowohl ein Haupteffekt des Vergleichs zwischen subjektiv optimaler und tatsächlicher Gewichtung, $F(1, 88) = 4.60$, $p < .05$, $\eta^2_p = .05$, als auch die Interaktion mit der Ausprägung der Zufallszahl, $F(1, 88) = 4.20$, $p < .05$, $\eta^2_p = .05$. Der Haupteffekt der Ausprägung der Zufallszahl war nicht länger signifikant, $F(1,88) = 3.24$, $p = .08$, $\eta^2_p = .04$. Die tatsächliche Gewichtung lag dabei über der als optimal angegebenen ($M = .08$, $SD = .10$ vs. $M = .05$, $SD = .10$).

Separate Analysen der bedingten Haupteffekte des Vergleichs zwischen subjektiv optimaler und tatsächlicher Gewichtung zeigten, dass in der Bedingung ohne Extremwerte die Zufallszahl signifikant stärker gewichtet wurde als die Probanden anfänglich intendierten ($M = .11$, $SD = .11$ vs. $M = .05$, $SD = .09$), $t(43) = 2.64$, $p < .05$, $d = 0.81$, während sich kein solcher Unterschied in der Bedingung mit Extremwerten zeigte ($M = .05$, $SD = .08$ vs. $M = .05$, $SD = .10$), $t(45) = 0.77$, $p = .94$, $d = 0.23$. Diese Befunde decken sich hinsichtlich der Bedingung ohne Extremwerte der Zufallszahl mit den Befunden aus Experiment 2 und dem deskriptiven Muster aus Experiment 3. Da sich in der Bedingung mit Extremwerten kein Unterschied zwi-

schen der subjektiv optimalen und der tatsächlichen Gewichtung der Zufallszahlen zeigt und das Befundmuster im Prinzip dem deskriptiven Muster der entsprechenden Analyse unter Einschuss aller Probanden gleicht, wird Hypothese 4a angenommen, während Hypothese 4b abgelehnt wird.

Für den kritischen Test der konkurrierenden Hypothesen 5a und 5b wurden gesondert diejenigen Probanden berücksichtigt, die anfänglich angaben, die Zufallszahl solle gar nicht berücksichtigt werden (N = 63, davon 33 in der Bedingung mit Extremwerten). Hier zeigt zum einen ein t-Test gegen Null, dass diese Probanden insgesamt ihre Finalschätzungen systematisch an die Zufallszahlen anpassten (M = .07, SD = .10), $t(62) = 5.42$, $p < .001$, d = 0.68. Das Ausmaß, in dem Probanden, die anfänglich angaben, die Zufallszahl sollte zu Null gewichtet werden, ihre Finalschätzungen tatsächlich in Richtung der Zufallszahl anpassten, unterschied sich in Abhängigkeit davon, ob die Zufallszahl Extremwerte annahm oder nicht (M = .05, SD = .08 vs. M = .10, SD = .12), $t(61) = -2.13$, p = .04, d = 0.55. Da bei Vorliegen von Extremwerten also die Übernutzung derjenigen Probanden, die anfänglich nicht intendierten, die Zufallszahl zu gewichten, geringer ausfiel, wird auch Hypothese 5a angenommen.

Insgesamt sprechen die Ergebnisse von Experiment 4 also dafür, dass nicht valide Ratschläge in stärkerem Maße übernutzt werden als die anfängliche irrationale Nutzungsüberzeugung nahelegt, wenn sie – bezogen auf die zu schätzende Größe – in einem plausiblen Wertebereich schwanken. Jedoch zeigten separate t-Tests für die Probanden, die anfänglich die Zufallszahl nicht gewichten wollten, sowohl eine überzufällige Anpassung der Finalschätzungen in Richtung der Zufallszahl für die Bedingung ohne Extremwerte, $t(29) = 4.46$, $p < .001$, d = 0.81, als auch für die Bedingung mit Extremwerten, $t(32) = 3.36$, $p < .01$, d = 0.54. Da also auch die Probanden, die anfänglich angaben, die Zufallszahl solle ignoriert werden, in der Bedingung mit Extremwerten ihre Finalschätzungen immer noch zu ca. 5% in Richtung der Zufallszahlen anpassten, muss davon ausgegangen werden, dass die Übernutzung von Ratschlägen nicht vollständig durch die anfängliche irrationale Nutzungsüberzeugung und die zusätzliche Fehlwahrnehmung der Validität aufgrund des plausiblen Schwankungsbereich der Zufallszahl erklärt werden kann.

7.5.5 Explorative Analyse möglicher Distanz-Effekte auf die Gewichtung

Auch in Experiment 4 wurde explorativ untersucht, in wie weit die Entfernung zwischen Initialschätzung und Ratschlag sich auf die Gewichtung des Ratschlags auswirkt. Dazu

wurden wieder die AT-Werte der Durchgänge mit derselben Abweichung unter Vernachlässigung der Richtung gemittelt. Die Durchgänge mit extremen Abweichungen der Zufallszahl wurden dabei nicht berücksichtigt. Um Vergleichbarkeit zwischen den Bedingungen herzustellen, wurden deshalb auch die korrespondierenden Durchgänge der Zufallszahl ohne Extremwerte sowie der beiden vermeintlichen Ratgeber außen vor gelassen. In der Konsequenz wurden also für jeden Ratgebertyp 25 Trials berücksichtigt, aus denen 13 Mittelwerte berechnet wurden. Die so berechneten AT-Werte gingen dann in eine 3 (*Art des Ratgebers*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 2 (*Ausprägung der Zufallszahl*: mit Extremwerten vs. ohne Extremwerte) × 13 (Abweichung von der Initialschätzung: 16%, 18%, 20%, 22%, 24%, 26%, 28%, 30%, 32%, 34%, 36%, 38% sowie 40%) ANOVA mit der Ausprägung der Zufallszahl als Zwischensubjektfaktor und den übrigen beiden Faktoren als Innersubjektfaktoren ein. Aufgrund fehlender Einzelwerte gingen in diese Analyse nur 76 der 99 Probanden ein, davon 36 in der Bedingung mit Extremwerten der Zufallszahl. Erwartungsgemäß zeigte sich ein Haupteffekt des Ratgebers, $F(2, 148) = 167.97$, $p < .001$, $\eta^2_p = .69$. Zusätzlich zeigten sich ein Haupteffekt der Distanz zwischen Ratschlag und Initialschätzung, $F(12, 888) = 5.40$, $p < .001$, $\eta^2_p = .07$, der am besten durch einen linearen Kontrast erklärt wurde, $F(1, 74) = 29.01$, $p < .001$, $\eta^2_p = .28$, sowie eine Interaktion zwischen Ratgeber und Distanz, $F(24, 1776) = 3.60$, $p < .001$, $\eta^2_p = .05$. Darüber hinaus zeigte die Analyse keine weiteren signifikanten Effekte, alle $F < 2.14$, alle $p > .12$, alle $\eta^2_p < .02$.

Der Haupteffekt der Distanz resultiert wie in den Experimenten 1 und 3 daraus, dass insbesondere Ratschläge und Zufallszahlen mit geringer Abweichung weniger stark berücksichtigt werden. So wurden Ratschläge mit Entfernungen von 16% oder 18% im Mittel zu .17 gewichtet, während die übrigen Ratschläge im Mittel zu .22 gewichtet wurden. Aufgrund der Interaktion zwischen Ratgeber und Distanz wurden wie in Experiment 2 separate Analysen für die drei Arten des Ratgebers durchgeführt. Diese zeigten, dass Distanzeffekte nur für den vermeintlich kompetenten und den vermeintlich wenig kompetenten Ratgeber auftraten, $F(12, 1032) = 7.92$, $p < .001$, $\eta^2_p = .08$, bzw. $F(12, 984) = 3.974$, $p < .001$, $\eta^2_p = .05$. Dabei konnte der Distanz-Effekt des kompetenten Ratgebers sowohl durch einen linearen Kontrast als auch durch einen zweiter Ordnung erklärt werden, $F(1, 86) = 14.38$, $p < .001$, $\eta^2_p = .14$, bzw. $F(1, 86) = 21.04$, $p < .001$, $\eta^2_p = .20$, während für den vermeintlich wenig kompetenten

Ratgeber ein Kontrast erster Ordnung die beste Passung bot, $F(1, 82) = 18.44$, $p < .001$, $\eta^2_p = .18$. Ferner zeigte sich bei beiden Ratgebern konsistent, dass die Ratschläge, die besonders nah an der Initialschätzung lagen weniger stark gewichtet wurden als Ratschläge, die weiter von der Initialschätzung entfernt waren. Für die Zufallszahl hingegen zeigte sich kein signifikanter Distanz-Effekt, $F(12, 1032) = 1.35$, $p = .19$, $\eta^2_p = .02$. Die aufgrund der Stichprobengröße etwas powerstärkere Analyse der Distanz-Effekte in Experiment 4 suggeriert also, dass Distanz-Effekte nur bei den vermeintlichen Ratgebern auftreten (siehe auch Abbildung 7).[12]

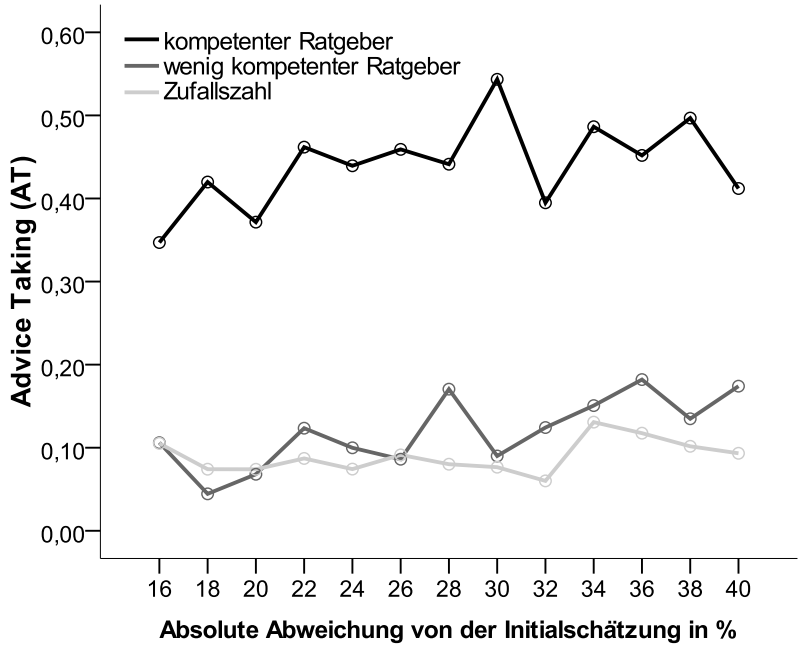


Abbildung 7. Mittlere AT-Werte in Abhängigkeit der Distanz zwischen Initialschätzung und Ratschlag bzw. Zufallszahl nach Art des Ratgebers in Experiment 3.

7.6 Diskussion

Die Ergebnisse von Experiment 4 replizieren zunächst erneut den zentralen Befund, dass Zufallszahlen systematisch übernutzt werden. In Experiment 4 sollte ferner geprüft werden, ob diese Übernutzung teilweise durch den Schwankungsbereich der nicht validen Ratschläge erklärt werden kann – also dadurch, dass sich die Zufallszahlen in einem plausiblen Wertebereich bewegen, so dass ihre Zufälligkeit zwar abstrakt bekannt, jedoch nicht augenfällig ist. Diese Annahme konnte bestätigt werden. Wenn die Zufallszahlen Extremwerte außerhalb eines plausiblen Bereichs annehmen, zeigte sich, dass die Übernutzung der Zu-

[12] Eine Analyse, in der die fehlenden Werte einer Versuchsperson durch deren jeweilige mittlere Gewichtung des betreffenden Ratgebers ersetzt wurden, um lieferte auf Basis des vollständigen Datensatzes dasselbe Befundmuster.

fallszahl mit ca. 6% deutlich niedriger ausfiel als die Gewichtung der Zufallszahl ohne Extremwerte mit 12%. Dies lässt darauf schließen, dass die systematische Gewichtung von Zufallszahlen zu einem gewissen Teil darauf beruht, wie plausibel die realisierten Werte anmuten. Dieser Befund ist vor allem deshalb interessant, weil die Validität eines Ratschlags und dessen Plausibilität in ganz besonderer Weise zusammenhängen. So ist es häufig möglich, von unplausiblen Werten darauf zu schließen, dass ein bestimmter Ratschlag wenig valide ist und besser ignoriert werden sollte. Beispielsweise kann man mit relativer Sicherheit behaupten, dass der Ratschlag, Rom und Athen lägen genau 12 Kilometer voneinander entfernt, wegen seiner mangelnden Plausibilität als wenig valide identifiziert werden kann; genauso kann man schlussfolgern, dass eine Finanzprognose, die einen Leitzinssatz von -12% vorhersagt, extrem unplausibel und daher wenig valide ist.[13] Der Schluss von hoher Plausibilität auf hohe Validität dagegen ist nicht automatisch gültig, das heißt, auch plausible Ratschläge können effektiv eine Validität von Null aufweisen. In der Realität ist es sogar sehr wahrscheinlich, dass auch wenig valide Ratschläge einem plausiblen Wertebereich entstammen, schon allein deshalb, weil der Ratgeber antizipieren kann, dass ein unplausibel erscheinender Ratschlag nicht ernst genommen werden wird. So bewegen sich auch Finanzprognosen in der Regel in plausiblen Bereich, sind aber dennoch oft nicht besser als der Zufall (z.B. Simon, 1989; Spiwoks, Bedke & Hein, 2008). Das bedeutet, dass Personen, die anhand der Plausibilität auf die Validität eines Ratschlags schließen, in der Realität möglicherweise besonders anfällig für die Übernutzung wenig valider Ratschläge sind.

Eine mögliche Erklärung für die Befunde von Experiment 4 besteht also darin, dass durch die Plausibilität der Werte eine Illusion der Validität entsteht, dass also die Probanden noch stärker als zu Beginn des Experiments der Überzeugung sind, die Gewichtung der Zufallszahl könne zu einer Verbesserung ihrer Finalschätzungen beitragen. Diese Illusion der Validität könnte insbesondere dadurch Zustande kommen, dass die Probanden bei der Einschätzung der Validität die Plausibilität der Werte im Sinne einer einfachen Heuristik als Hinweisreiz nutzen.

Die Ergebnisse von Experiment 4 zeigten weiterhin, dass auch diejenigen Probanden, die zu Beginn des Experiments keine irrationale Nutzungsüberzeugung aufwiesen und die

[13] Ich gehe hier davon aus, dass die Zentralbank es auch in Zukunft für sinnvoll hält, das Verleihen von Kapital zu vergüten, und dass sich die Idee, den Kreditnehmer dafür zu bezahlen, dass er sich Geld leiht, nicht durchsetzt.

auch nicht der Illusion der Validität erliegen sollten, weil bisweilen Extremwerte auftraten, die Zufallszahl immer noch zu ca. 5% gewichteten. Daraus lässt sich ableiten, dass die systematische Übernutzung von Zufallszahlen durch eine anfängliche irrationale Nutzungsüberzeugung und die Illusion der der Validität aufgrund der Plausibilität der Werte nicht vollständig erklärt werden kann. Eine Übernutzung von nicht validen Ratschlägen ist selbst dann noch beobachtbar, wenn diese beiden Mechanismen ausgeschaltet werden. Es ist also denkbar, dass auch eine subjektive Übernutzung der Zufallszahl auftritt. Eine solche subjektive Übernutzung müsste dann wiederum durch einen psychologischen Mechanismus hervorgerufen werden, der nicht auf der Wahrnehmung der Validität der Zufallszahlen basiert. Ein solcher Mechanismus wird in Experiment 5 postuliert und überprüft.

8. Experiment 5

8.1 Zielsetzung und Hypothesen

Die Experimente 2 bis 4 haben konsistent gezeigt, dass die Fehlwahrnehmung der Validität nicht valider Ratschläge anteilig erklären kann, warum Zufallszahlen systematisch in das Finalurteil einbezogen werden. Sowohl anfängliche irrationale Nutzungsüberzeugungen als auch die darüber hinausgehende Überschätzung der Validität aufgrund plausibler Wertebereiche führen dazu, dass nicht valide Ratschläge objektiv übernutzt werden. Dieser Erklärungsansatz ähnelt in Grundzügen demjenigen, der auch bei Multi-Cue-Judgments zu einer überhöhten Nutzung weniger valider Hinweisreize führt (siehe Abschnitt 3.1). Da eine Übernutzung der Zufallszahlen aber auch dann noch nachweisbar ist, wenn weder eine anfängliche Fehleinschätzung der Validität noch ein plausibler Schwankungsbereich der Zufallszahl vorliegen, stellt sich die Frage, welche weiteren psychologischen Mechanismen zu einer systematischen Übernutzung nicht valider Ratschlägen beitragen können. Experiment 5 soll diese Frage zumindest in Grundzügen klären, indem ein weiterer möglicher Erklärungsansatz untersucht wird.

Da eine Vielzahl denkbarer Erklärungsansätze existiert und eine erschöpfende Testung aller denkbaren Ansätze den Rahmen bei weitem sprengen würde, wird in Experiment 5 nur derjenige Erklärungsansatz getestet, der vor dem Hintergrund der Ausführungen im Theorieteil am vielversprechendsten erscheint und sich somit auch theoretisch am besten begründen lässt, nämlich eine Kombination aus *positivem Hypothesentesten* und *selektiver*

Verfügbarkeit zielrelevanter Wissensinhalte. Es handelt sich dabei um einen Ansatz aus der sozialen Kognitionsforschung, der eine kritische Rolle bei Vergleichsprozessen in der Urteilsbildung spielt (Mussweiler, 2003). Darunter fallen soziale Vergleiche (Mussweiler, 2001), Kategorisierungen bei der Personenwahrnehmung (Mussweiler, 2003; Mussweiler & Bodenhausen, 2002), aber auch der Anker-Effekt (Mussweiler & Strack, 1999, 2001; Mussweiler et al., 2000). Nach der Selective-Accessibility-Hypothese von Mussweiler und Strack (1999, 2001; siehe auch Mussweiler, 2003) beginnt ein Vergleichsprozess zunächst mit einer oberflächlichen Prüfung der Ähnlichkeit eines zu bestimmenden Objekts oder Sachverhalts mit einer bestimmten Kategorie oder einem Referenzwert. Im Falle der sozialen Vergleiche wird beispielsweise die Ähnlichkeit zwischen sich selbst und einer anderen Person als Referenzmaßstab geprüft, bei der Kategorisierung die Ähnlichkeit zwischen einem Objekt und dem prototypischen Vertreter der jeweiligen Referenzkategorie. Das Ergebnis dieses oberflächlichen Vergleichs bestimmt dann, ob im Folgenden die Hypothese geprüft wird, dass Ziel und Referenz ähnlich sind oder ob sie unähnlich sind. Beim Anker-Effekt besteht der kritische Vergleich in der Prüfung der Ähnlichkeit zwischen der zu schätzenden Zielgröße und dem Anker. Dieser Vergleich wird durch die komparative Frage ausgelöst, ob die gesuchte Größe höher oder niedriger als der Anker ist. Hier ist die Besonderheit, dass durch die Art der komparativen Frage stets die Hypothese entsteht, dass Anker und Zielgröße ähnlich sind (Mussweiler, 2003). Die Hypothese, die sich aus dem anfänglichen oberflächlichen Vergleichsprozess ergibt, wird dann im Sinne positiven Hypothesentestens nur einseitig geprüft, das heißt, es werden vorwiegend Wissensinhalte abgerufen oder Argumente generiert, die diese Hypothese stützen. Wissensinhalte oder Argumente, die eine Alternativhypothese stützen, bleiben hingegen unberücksichtigt. Bei dem nun folgenden Urteil, das die Person abgibt, setzt der zweite Teilprozess ein, nämlich selektive Verfügbarkeit von Wissensinhalten. Grundidee ist hierbei, dass die Person nun ihre zielrelevanten Wissensinhalte abruft, um ein akkurates Urteil zu fällen. Hier sind nun bereits durch das positive Hypothesentesten selektiv diejenigen Wissensinhalte im Sinne semantischen Primings voraktiviert, die die fokale Hypothese stützen, also die Hypothese, die durch den anfänglichen Vergleich entsteht. Da diese voraktivierten Wissensinhalte mit höherer Wahrscheinlichkeit abgerufen werden können als Wissensinhalte, die noch nicht voraktiviert wurden, haben sie im Mittel auch einen maßgeblichen Einfluss auf die Urteilsbildung. Im Ergebnis wird letztlich das Urteil in Richtung der fokalen Hypothese verzerrt, das heißt also das Urteil bestätigt oder extremisiert sogar das Ergeb-

nis des anfänglichen Vergleichs. Im Falle des Anker-Effekts wird das Urteil in Richtung des Ankers verzerrt.

Da positives Hypothesentesten und selektive Verfügbarkeit also bei einer Vielzahl unterschiedlicher Urteilsprozesse zu wirken scheinen, scheint es plausibel, dass diese Prozesse auch im Judge-Advisor-Paradigma auftreten können. Dies soll an einem fiktiven Beispiel verdeutlicht werden: Eine Person soll im Judge-Advisor-Paradigma die Entfernung zwischen London und Rom schätzen. Sie gibt zunächst ihre Initialschätzung ab, zum Beispiel 1200 Kilometer. Konfrontiert man diese Person dann mit einem nicht validen Ratschlag in Form einer Zufallszahl, die den Wert 1500 annimmt, dann sollte diese Person nach der Selective-Accessibility-Hpothese unwillkürlich überlegen, ob die zu schätzende Größe dem Wert von 1500 entsprechen könnte. Allerdings bezieht sich die Selective-Accessibility-Hypothese bisher vorrangig auf Situationen, in denen noch kein Initialurteil vorliegt. Deswegen ist es unklar, ob Personen im Judge-Advisor-Paradigma tatsächlich die fokale Hypothese generieren, dass der Zielwert exakt dem Ratschlag entspricht, oder ob, ganz allgemein die Hypothese generiert wird, der Zielwert müsse in Richtung des Ratschlags liegen. Durch positives Hypothesentesten würden dann aber, unabhängig von der konkreten Ausprägung der fokalen Hypothese selektiv diejenigen Wissensinhalte aktiviert, die die fokale Hypothese stützen. Bei der Abgabe des Finalurteils wären dann diese Wissensinhalte wiederum voraktiviert und leichter verfügbar, d.h. die Person hätte im konkreten Beispiel vor allem Argumente dafür verfügbar, weshalb der wahre Wert höher liegen sollte als ihre Schätzung. Folglich wird sie ihre Schätzung nach oben und damit in Richtung der Zufallszahl korrigieren.[14]

Eine gut realisierbare Methode, einen vermittelnden Mechanismus nachzuweisen, besteht darin, diesen Prozess gezielt anzusteuern, das heißt, ihn durch experimentelle Manipulationen gezielt zu imitieren oder aber auszuschalten. Im konkreten Fall der Übernutzung von Ratschlägen kommen dabei zwei Teilprozesse in Frage, die separat voneinander untersucht werden können, die sich aber unterschiedlich gut für die gerade beschriebene Art der Überprüfung eignen. Die selektive Aktivierung von Wissensinhalten kann im Prinzip wie ein

[14] Inhaltlich wäre nach dieser Argumentation das Judge-Advisor-Paradigma am ehesten mit dem Anker-Paradigma vergleichbar. Allerdings bestehen zwei gravierende Unterschiede zwischen den beiden Paradigmen. Erstens fehlt im Judge-Advisor-Paradigma das für das Anker-Paradigma typische komparative Urteil, und zweitens liegt bereits ein Urteil vor, wenn der Ratschlag präsentiert wird, weshalb die Situation deutlich komplexer ist. Aufgrund dieser Unterschiede kann per definitionem nicht von einem Anker-Effekt im klassischen Sinne gesprochen werden. Jedoch sind diese Voraussetzungen für die postulierten Prozesse gegeben, da im Judge-Advisor-Paradigma sowohl ein Referenzwert als auch ein nachfolgendes Urteil vorliegen.

semantisches Priming verstanden werden (Mussweiler, 2003). Das bedeutet, dass es sich hier um einen automatischen Prozess handelt, dem schwer entgegenzuwirken ist (Mussweiler & Strack, 1999). Positives Hypothesentesten hingegen kann durch bestimmte Strategien angeregt oder verhindert werden. Eine solche Strategie wird in Experiment 5 eingesetzt, um zu überprüfen, ob selektives Hypothesentesten an der Entstehung der Übernutzung von nicht validen Ratschlägen beteiligt ist.

Eine Möglichkeit, positivem Hypothesentesten einer bestimmten Hypothese entgegenzuwirken, besteht darin, denselben Prozess für die gegenteilige Hypothese anzuregen, zum Beispiel durch den Einsatz der Explanation- und Counter-Explanation-Strategie (Anderson & Sechler, 1986). Unter Explanation versteht man dabei, dass Gründe dafür überlegt und aufgeschrieben werden, dass eine bestimmte Hypothese zutrifft. Diese Explanation-Strategie erzeugt oder verstärkt artifiziell positives Testen einer bestimmten Hypothese, da nur Gründe angeführt werden sollen, die der Hypothese entsprechen, nicht jedoch solche, die ihr widersprechen. Anderson und Sechler baten zum Beispiel einige Probanden, gezielt Gründe dafür zu nennen, dass besonders risikofreudige Personen gute Feuerwehrmänner abgeben. Personen, die diese Explanation-Strategie anwandten, hielten danach risikobereite Bewerber auf eine Position als Feuerwehrmann in der Tat für besser geeignet als Personen, die nicht speziell instruiert wurden.

Dieser Logik entsprechend dient Counter-Explanation dazu, positives Hypothesentesten zu vermeiden, indem zusätzlich negatives Hypothesentesten induziert wird. Es wird also gezielt instruiert, auch Gründe gegen eine bestimmte Hypothese zu nennen. Das bedeutet, dass Counter-Explanation immer dann hilfreich ist, wenn eine Person von sich aus einem bestimmten Bias unterliegt, der durch positives Hypothesentesten zustande kommt, und dann instruiert wird, selektiv die Gegenhypothese zu testen. Formal werden dann separat Hypothese und Gegenhypothese jeweils selektiv getestet, wodurch in der Summe ausgewogenes Hypothesentesten entsteht (z.B. Kadous, Krische & Sedor, 2002; Sanna, Schwarz & Stocker, 2002).

Um diesen Mechanismus im Judge-Advisor-Paradigma zu manipulieren, werden die Probanden gebeten, entweder Gründe dafür zu nennen, warum ihre erste Schätzung zu hoch sein könnte oder warum sie zu niedrig sein könnte. Sofern die Explanation-Strategie wirkt, sollten die Probanden grundsätzlich im ersten Fall ihre Finalschätzungen nach unten

und im zweiten Fall nach oben korrigieren, weil sie zu positivem Hypothesentesten in die jeweilige Richtung angeregt werden. Wenn diese Explanation-Instruktion in einer Situation gegeben wird, in der kein Ratschlag vorliegt, dann sollte sich durch die Instruktion der vermutete Prozess, der der Übernutzung nicht valider Ratschläge zugrunde liegt, nämlich positives Hypothesentesten, imitieren lassen. Mit anderen Worten: Die Aufforderung, Gründe dafür zu nennen, warum die erste Schätzung zu hoch sein könnte, sollte prinzipiell dieselbe Wirkung haben wir die Vorgabe eines Ratschlags bzw. einer Zufallszahl, der bzw. die von der Initialschätzung der Person nach unten abweicht.

Ein differenzierteres Bild sollte sich dagegen zeigen, wenn zusätzlich zu der Instruktion ein nicht valider Ratschlag dargeboten wird. Unter der Annahme, dass die subjektive Übernutzung von Ratschlägen tatsächlich durch selektive Aktivierung und positives Hypothesentesten entsteht, sollte dieser Prozess immer dann ausgeschaltet oder zumindest abgeschwächt werden, wenn die Instruktion eine Anpassung in eine Richtung nahelegt, der Ratschlag aber eine Anpassung in die jeweils andere, wenn also die Implikationen der Instruktion und die des Ratschlags zueinander *inkonsistent* sind. Instruktion und Ratschlag sind also zum Beispiel dann inkonsistent, wenn der Judge zunächst als Initialschätzung eine Entfernung von 1000km angibt, dann Gründe dafür anführen soll, warum diese Schätzung zu hoch ist und anschließend als Ratschlag die Entfernung von 1500km angezeigt wird, was eine Anpassung der Finalschätzung nach oben suggeriert. Im Ergebnis würde man hier erwarten, dass die Übernutzung nicht valider Ratschläge im Vergleich zu Durchgängen ohne vorherige Instruktion reduziert ist. Wenn dagegen die Implikationen der Instruktion und des Ratschlags zueinander *konsistent* sind, weil beide eine Anpassung der Finalschätzung in dieselbe Richtung nahelegen, dann sollte zweimal derselbe Prozess positiven Hypothesentestens angeregt werden. Im Ergebnis sollte dann im Vergleich zu Bedingungen, in denen nur ein entsprechender Ratschlag oder nur eine entsprechende Instruktion dargeboten wird, kein oder nur ein geringer Unterschied hinsichtlich der Anpassung der Finalschätzung auftreten, weil sowohl die Instruktion als auch der Ratschlag denselben Prozess anregen. Im Folgenden wird immer dann von Explanation gesprochen, wenn Instruktion und Ratschlag zueinander konsistent sind und in der Summe positives Hypothesentesten resultieren sollte; sind sie dagegen zueinander inkonsistent, dann sollten sowohl positives als auch negatives (und deshalb in der Summe ausgewogenes) Hypothesentesten eintreten, was als Counter-Explanation bezeichnet wird.

Wenn die experimentelle Manipulation tatsächlich in der Lage ist, positives Hypothesentesten im Judge-Advisor-Paradigma zu imitieren, sollte sich ein Haupteffekt der Explanation-Instruktion zeigen. Dass dieser Prozess überhaupt angeregt werden kann, ist die Voraussetzung dafür, dass er in den Bedingungen mit Ratgebern einer Übernutzung entgegenwirken kann. Die optimale Bedingung, um dies zu überprüfen ist die Bedingung ohne Ratschläge, da hier die Wirkung der Instruktion nicht durch das Zusammenspiel mit Ratschlägen verzerrt wird. Diese Wirksamkeitskontrolle wird formuliert als:

Hypothese 1: Die Explanation-Strategie hat einen Einfluss auf die Finalschätzung in der Bedingung ohne Ratgeber, d.h. Personen passen ihre Finalschätzung verglichen mit Durchgängen ohne Explanation systematisch in die Richtung an, die durch die jeweilige Instruktion nahegelegt wird.

Ferner wird angenommen, dass sich der Übernutzungseffekt auch in Experiment 5 erneut replizieren lässt. Dies wird wie folgt formuliert:

Hypothese 2: Personen passen ihre Finalschätzung überzufällig in Richtung der vermeintlichen Zufallszahl an, d. h. es liegt objektive Übernutzung vor (Replikation der Befunde aus den Experimenten 1 bis 4).

Die Wirkung der experimentellen Manipulation vorausgesetzt sollte Counter-Explanation (inkonsistente Bedingung) dem Übernutzungs-Effekt entgegenwirken. Ob der Effekt vollkommen neutralisiert wird, ist allerdings aus zwei Gründen fraglich: erstens ist unklar, wie stark die Wirkung der Counter-Explanation im Vergleich zur Wirkung der Zufallszahlen ausfällt, und zweitens kommt, wie in den Experimenten 2 bis 4 demonstriert, ein Teil der Gewichtung der Zufallszahlen durch eine anfängliche irrationale Nutzungsüberzeugung und möglicherweise eine im Verlauf des Experiments entstehende Illusion der Validität zustande. In jedem Falle aber sollte Counter-Explanation den Effekt des selektiven Hypothesentestens reduzieren. Im Fall einer Explanation-Instruktion (konsistente Bedingung), sollten sich hingegen keine oder nur geringe Unterschiede im Vergleich zu Durchgängen ohne Counter-Explanation zeigen.

Ein interessanter Aspekt positiven Hypothesentesten und selektiver Verfügbarkeit besteht darin, dass die Plausibilität oder Validität des Referenzwertes keine Rolle spielt (Mussweiler & Strack, 1999). Das impliziert, dass jeder Ratschlag gleichermaßen positives

Testen der fokalen Hypothese anstößt, der Zielwert liege in Richtung dieses Ratschlags, und zwar unabhängig davon, ob der Ratschlag von einem Experten, einem Laien oder eben einem Zufallsgenerator stammt. Träfe dies zu, dann sollte sich nicht nur für die vermeintliche Zufallszahl, sondern auch für die beiden vermeintlichen Ratgeber zeigen, dass Counter-Explanation das Ausmaß der Nutzung der Ratschläge reduziert, während Explanation den Grad der Nutzung nicht oder nur in geringem Maße erhöht. Der entscheidende Unterschied zwischen den Ratgebertypen wäre dann, dass zusätzlich zu der – für alle Ratgeber gleichen – unwillkürlichen Nutzung durch positives Hypothesentesten und selektive Verfügbarkeit von Wissensinhalten eine intentionale der Gewichtung in Abhängigkeit der wahrgenommenen Validität vorgenommen wird. In der Summe sollte sich also folgendes zeigen lassen:

Hypothese 3: Counter-Explanation reduziert die Nutzung von Ratschlägen im Vergleich zu Kontrolldurchgängen ohne gesonderte Instruktion, während die Gewichtung der Ratschläge bei Explanation verglichen mit den Kontrolldurchgängen größer oder gleich ausfällt.

Hypothese 4: Counter-Explanation reduziert die Übernutzung von Zufallszahlen im Vergleich zu Kontrolldurchgängen ohne gesonderte Instruktion, während die Gewichtung der Zufallszahlen bei Explanation verglichen mit den Kontrolldurchgängen größer oder gleich ausfällt.

8.2 Stichprobe und Design

An Experiment 5 nahmen 123 Studierende unterschiedlicher Fachrichtungen der Georg-August-Universität Göttingen teil, davon 70 weibliche Studierende (57%). Das Durchschnittsalter der Teilnehmer lag bei 23,71 Jahren (*SD* = 2,72 Jahre). Experiment 5 folgt einem 4 (*Art des Ratgebers*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl vs. kein Ratschlag) × 3 (*Instruktion:* Explanation vs. Counter-Explanation vs. keine Instruktion) – Design mit der Art des Ratgebers als Innersubjektfaktor und der Konsistenz zwischen Ratschlag und Überlegung als Zwischensubjektfaktor[15].

[15] Die Darstellung als 4 × 3 – Design ist streng genommen vereinfacht und an die inhaltliche Bedeutung der Kombination der manipulierten Variablen geknüpft. Rein formal entspricht das zugrunde liegende Design einem 4 (Art des Ratgebers vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl vs. kein Ratschlag) × 3 (Richtung der Instruktion: zu hoch vs. zu niedrig vs. keine

8.3 Methode

Ebenso wie die vorigen Experimente wurde Experiment 5 computergestützt mittels der Software Presentation® (Version 13.0, www.neurobs.com) programmiert und dargeboten (der entsprechende Programmcode für Experiment 5 ist im digitalen Anhang enthalten). Die Vorgehensweise entspricht der von Experiment 3 mit folgenden beiden Änderungen: der Zwischensubjektfaktor „Art des Ratgebers“ wurde um eine Kontrollbedingung ohne Ratschläge erweitert. Diese Kontrollbedingung entspricht derjenigen aus Experiment 1, das heißt anstelle eines Ratschlag erschien der Hinweis, die Probanden hätten nun Gelegenheit, ihre erste Schätzung zu überdenken. Zweitens wurde die Anzahl der Durchgänge von 30 auf 90 erhöht. In 30 der 90 Durchgänge wurden die Probanden instruiert, nach ihrer ersten Schätzung und vor Erscheinen des Ratschlags Gründe dafür anzugeben, warum diese Schätzung zu hoch sein könnte; in weiteren 30 Durchgängen sollten sie Gründe dafür angeben, warum die erste Schätzung zu niedrig sein könnte; die übrigen 30 Durchgängen dienten als Kontrolldurchgänge, dort wurde also keine entsprechende Instruktion dargeboten. Sowohl bei den Durchgängen mit Frage als auch in den Kontrolldurchgängen, wichen die dargebotenen Ratschläge in der Hälfte der Durchgänge nach oben und in der anderen Hälfte nach unten ab. Die jeweiligen prozentualen Abweichungen waren identisch zu denen der vorherigen Experimente, lagen damit also wieder im Bereich von ± 40%. Die Reihenfolge der Darbietung der 90 Trials wurde randomisiert, war aber für alle Probanden identisch.

Eine weitere kleinere Änderung betraf die Darbietung der Zufallszahl. Um den Anschein der Zufälligkeit im Vergleich zu den vorigen Experimenten zu verstärken, wurde die Zufallszahl immer so dargeboten, dass vor Erscheinen der konkreten Zahl ein Schnelldurchlauf aller Vielfachen von 10 im Bereich von der Zahlen von bis 9990 gezeigt wurde. Auf diese Weise sollte jeglicher Verdacht vermieden werden, dass diese Zahlen nicht wirklich per Zufall ermittelt werden.

8.4 Ablauf der einzelnen Durchgänge

Die Durchgänge von Experiment 5 liefen mit folgenden Ausnahmen identisch zu Experiment 4 ab: in den Durchgängen mit Explanation erschien nach Eingabe der Initialschätzung die Aufforderung, Gründe zu überlegen, warum die erste Schätzung zu hoch bzw. zu

Instruktion) x 2 (Richtung der Abweichung des Ratschlags: höher als Initialschätzung vs. niedriger als Initialschätzung) – Design. In beiden Fällen wird vernachlässigt, dass formal ein „nested Design“ vorliegt, da in der Kontrollbedingung ohne Ratschläge der letzte Faktor nicht existieren kann.

niedrig sein könnte. Zudem wurden die Probanden gebeten, diese Überlegungen in ein dafür vorgesehenes Eingabefeld einzugeben. Je nach experimenteller Bedingung erschien nach der Eingabe der Gründe der jeweilige Ratschlag bzw. die Zufallszahl oder die Aufforderung, die ursprüngliche Schätzung zu überdenken.

8.5 Ergebnisse

8.5.1 Berechnung des Advice Taking und Überprüfung möglicher Störvariablen

Zunächst wurden in allen Kombinationen der experimentellen Bedingungen, in denen entweder ein Ratschlag oder eine Zufallszahl dargeboten wurde, die mittleren AT-Werte über alle 30 Durchgänge berechnet. Analog zu den vorigen Experimenten wurden nur AT-Werte einbezogen, die kleiner -1,5 oder größer 1,5 waren. Dadurch wurden 292 der 8280 Einzelwerte (ca. 4%) ausgeschlossen. Der mittlere AT-Wert über alle 90 Durchgänge wurde mit dem Alter der Probanden korreliert, wobei sich wie in den vorigen Experimenten kein signifikanter Zusammenhang zeigte $r(91) = .12$, $p = .26$. Eine ANOVA mit der Art des Ratgebers und dem Geschlecht der Probanden als Zwischensubjektfaktoren und der Instruktion als Zwischensubjektfaktor zeigte ferner weder einen Haupteffekt des Geschlechts, noch eine Interaktion des Geschlechts mit der Instruktion oder mit Instruktion und Art des Ratgebers, alle $F < 0.92$, alle $p > .59$, alle $\eta^2_p < .03$. Es zeigte sich jedoch eine Interaktion zwischen dem Geschlecht der Probanden und der Art des Ratgebers, $F(2, 86) = 3.26$, $p = .04$, $\eta^2_p = .07$, die jedoch inhaltlich nicht erklärbar ist, da sich männliche und weibliche Probanden bei einem separaten Vergleich nach Art des Ratgebers hinsichtlich ihrer Gewichtung nicht signifikant unterschieden, alle $|t(28 \text{ bzw. } 29)| < 1.70$, alle $p > .10$, alle $d < 0.63$. Dass die Unterschiede statistisch nicht signifikant waren, mag vorranging ein Resultat der relativ geringen Stichprobengröße der jeweiligen Einzelvergleiche sein. Da männliche und weibliche Probanden jedoch über die drei Bedingungen hinweg gleichverteilt waren, χ^2 (2, $N = 92$) = .02, $p = .99$, wurde der Geschlechtseffekt im weiteren Verlauf vernachlässigt.

8.5.2 Wirksamkeitskontrolle der Instruktion

Zunächst wurde im Sinne einer Wirksamkeitskontrolle geprüft, ob durch die Instruktionen eine Veränderung der Finalschätzung erreicht werden konnte. Dazu wurden in der Kontrollbedingung ohne Ratgeber zunächst für die einzelnen Durchgänge die prozentualen Veränderungen zwischen Final- und Initialschätzung berechnet. Analog zu den Experimenten

1 und 2 wurden Extremwerte, die höchstwahrscheinlich aus Fehleingaben resultierten, ausgeschlossen. Berücksichtigt wurden wie in den vorigen Experimenten Werte, die zwischen -0.6 und +0.6 lagen. Aus den Einzelwerten wurden dann drei Mittelwerte gebildet, davon ein Mittelwert für die Durchgänge, in denen Gründe dafür genannt werden sollten, dass die Initialschätzung zu niedrig war, ein Mittelwert für die Durchgänge, in denen Gründe dafür genannt werden sollten, dass die Initialschätzung zu hoch war, und ein Mittelwert für die Durchgänge ohne Instruktion. Eine ANOVA mit der Richtung der Instruktion (zu niedrig vs. zu hoch vs. keine Instruktion) als Innersubjektfaktor zeigte Unterschiede zwischen den drei Bedingungen hinsichtlich der mittleren prozentualen Veränderung zwischen Initial- und Finalschätzung, $F(2, 60) = 17.73$, $p < .001$, $\eta^2_p = .37$, wobei sich alle drei Mittelwerte jeweils von den anderen beiden unterschieden, alle $F(1, 30) > 5.53$, alle ps $< .03$, alle $\eta^2_p > .15$. Separate t-Tests gegen Null zeigten, dass Probanden ihre Finalschätzung systematisch nach oben anpassten, wenn dies durch die Explanation-Instruktion nahegelegt wurde, und zwar im Mittel um 6%, $t(30) = 5.35$, $p < .001$, $d = 0.96$. Wurden die Probanden hingegen gebeten, Gründe dafür zu nennen, warum die erste Schätzung zu hoch sein könnte, so resultierte eine Anpassung um -3%, das heißt, die Finalschätzungen wurden systematisch nach unten angepasst, $t(30) = -2.70$, $p = .01$, $d = 0.48$. Im Gegensatz dazu fand in den Durchgängen ohne Explanation-Instruktion keine systematische Veränderung zwischen Initial- und Finalschätzung statt, $t(30) = -0.61$, $p = .55$, $d = 0.11$. Die Ergebnisse dieser Analysen zeigen, dass die Instruktion wirksam ist und in Abwesenheit von Ratschlägen zu einer systematischen Veränderung der Finalschätzungen führt. Hypothese 1 wird somit angenommen.

8.5.3 Subjektiv optimale Gewichtung der Ratgeber und der Zufallszahl

Analog zu den vorherigen Experimenten wurde die von den Probanden zu Beginn des Experiments als optimal angegeben Gewichtung analysiert. Zunächst wurde dabei die Validität der anfänglichen Einschätzungen bezüglich der späteren tatsächlichen Gewichtung für jeden Ratgebertyp untersucht. Wie in den Experimenten 3 und 4 zeigte sich, dass diejenigen Probanden, die den vermeintlich kompetenten Ratgeber anfänglich stärker gewichten wollten, diesen im Verlauf des Experiments auch tatsächlich stärker gewichteten, $r(31) = .42$, $p = .02$. Gleiches zeigte sich für die vermeintliche Zufallszahl, $r(30) = .53$, $p < .01$. Anders als in den vorigen Experimenten zeigte sich ein vergleichbarer Zusammenhang für den vermeintlich wenig kompetenten Ratgeber, $r(31) = .52$, $p < .01$. Es konnte also erneut gezeigt werden,

dass die Probanden bei der Gewichtung der Ratschläge in gewissem Maße im Sinne ihrer anfänglichen Intentionen handeln. Jedoch war der Zusammenhang auch in Experiment 5 nur moderat, da die anfängliche Nutzungsintention je nach Ratgeber nur zwischen 16% und 28% der Varianz aufklärt. Das wiederum könnte ein Hinweis darauf sein, dass die tatsächliche Gewichtung zu einem substantiellen Teil von Faktoren abhängt, die die Probanden nicht antizipieren können (wie z.B. die im Laufe des Experiments auftretende Illusion der Validität) oder im Laufe des Experiments nicht kontrollieren können (z.B. positives Hypothesentesten und selektive Verfügbarkeit).

Eine ANOVA mit der Art des Ratschlags als Zwischensubjektfaktor und der subjektiv optimalen Gewichtung als abhängiger Variable zeigte einen Haupteffekt des Ratgebers, $F(2, 89) = 52.83$, $p < .001$, $\eta^2_p = .55$. Wie auf der Grundlage der Ergebnisse der Experimente 2 bis 4 erwartet, gaben die Probanden zu Beginn an, den vermeintlich kompetenten Ratgeber stärker gewichten zu wollen als den vermeintlich wenig kompetenten Ratgeber ($M = .59$, $SD = .17$ vs. $M = .22$, $SD = .20$) sowie die Zufallszahl ($M = .59$, $SD = .17$ vs. $M = .14$, $SD = .18$), beide $t(87) > 7.99$, beide $p < .001$, beide $d > 1.71$. Die Probanden gaben ferner an, den vermeintlich wenig kompetenten Ratgeber stärker gewichten zu wollen als die Zufallszahl ($M = .22$, $SD = .20$ vs. $M = .14$, $SD = .18$), jedoch war der Unterschied statistisch nicht signifikant, $t(87) = 1.75$, $p = .08$, $d = 0.38$. Insgesamt zeigte sich aber auch in Experiment 5, dass die Probanden sensitiv für die erwartete Qualität der Ratschläge waren, wenn sie einschätzten, wie stark diese jeweils in ihr Urteil einbezogen werden sollten.

Abschließend wurde noch, wie in den vorherigen Experimenten die von den Probanden zu Beginn des Experiments als optimal angegebene Gewichtung der Zufallszahl mit der tatsächlichen Gewichtung verglichen. Zum Zwecke maximaler Vergleichbarkeit wurden dabei nur die Durchgänge ohne Explanation- oder Counter-Explanation-Instruktion verwendet. Hier zeigte sich zunächst kein Unterschied zwischen der anfänglich intendierten und der tatsächlichen Gewichtung der Zufallszahl, $t(28) = -0.90$, $p = .38$, $d = 0.34$.[16] Eine ANOVA mit dem Vergleich zwischen anfänglicher Nutzungsintention und tatsächlicher Gewichtung der Zufallszahl als Innersubjektfaktor und dem Vorliegen einer irrationalen Nutzungsüberzeugung als Zwischensubjektfaktor zeigte zunächst einen Haupteffekt des Vorliegens einer irrationa-

[16] Die Ergebnisse sind identisch, wenn die anfängliche Nutzungsintention mit dem AT-Wert über alle 90 Durchgänge verglichen wird.

len Nutzungsüberzeugung, $F(1, 26) = 25.29$, $p < .001$, $\eta^2_p = .49$, der aufgrund der Korrelation zwischen intendierter und tatsächlicher Nutzung zu erwarten war. Der Haupteffekt des Vergleichs zwischen intendierter und tatsächlicher Nutzung der Zufallszahl war analog zu dem zuvor berichteten *t*-Test nicht signifikant, $F(1, 26) = 0.34$, $p = .56$, $\eta^2_p = .01$. Es zeigte sich aber auch Interaktion der beiden Faktoren, $F(1, 26) = 9.24$, $p < .01$, $\eta^2_p = .26$. Diese Interaktion kam dadurch Zustande, dass diejenigen Probanden, die anfänglich angaben, die Zufallszahl vollständig ignorieren zu wollen, sie letztendlich signifikant stärker als Null nutzten ($M = .06$ $SD = .08$), $t(11) = 2.47$, $p = .03$, $d = 1.49$, während für die Probanden, die anfangs eine irrationale Nutzungsüberzeugung aufwiesen, die tatsächliche Gewichtung schwächer ausfiel als anfänglich intendiert ($M = .14$ $SD = .09$ vs. $M = .16$ $SD = .09$), $t(11) = -2.32$, $p = .03$, $d = 1.20$. Dieses Muster deutet erneut darauf hin, dass die anfängliche Nutzungsintention nur bedingt in der Lage ist, die tatsächliche Nutzung der Zufallszahlen vorherzusagen.

8.5.4 Gewichtung der Ratschläge und Zufallszahlen

Die mittleren AT-Werte gingen zunächst als abhängige Variable in eine 3 (*Art des Ratgebers*: kompetent vs. wenig kompetent vs. Zufallszahl) × 3 (*Instruktion*: Explanation vs. Counter-Explanation vs. keine Instruktion) ANOVA ein. Da für die Kontrollbedingung ohne Ratgeber keine AT-Werte berechnet werden können, wurde sie in dieser Analyse nicht berücksichtigt. Es zeigte sich zunächst ein Haupteffekt für die Art des Ratgebers, $F(2, 89) = 50.65$, $p < .001$, $\eta^2_p = .53$. Post-hoc-Kontraste zeigten, dass wie in den vorigen Experimenten der vermeintlich kompetente Ratgeber mit einem mittleren AT-Wert von .39 ($SD = .11$) stärker gewichtet wurde als sowohl der vermeintlich wenig kompetente mit einem AT-Wert von .14 ($SD = .12$), $t(89) = 8.65$, $p < .001$, $d = 1.83$, sowie die Zufallszahl mit einem AT-Wert von .12 ($SD = .12$), $t(89) = 9.20$, $p < .001$, $d = 1.95$, während sich letztere nicht voneinander unterschieden, $t(89) = 0.62$, $p = .54$, $d = .13$. Erwartungsgemäß war die Gewichtung der vermeintlichen Zufallszahl signifikant von Null verschieden, $t(29) = 5.81$, $p < .001$, $d = 1.06$, d.h. es lag erneut eine objektive Übernutzung der Zufallszahl vor, weshalb Hypothese 2 angenommen wird.

Die ANOVA ergab weiterhin einen Haupteffekt der Instruktion, $F(2, 178) = 11.58$, $p < .001$, $\eta^2_p = .12$. Post-hoc-Kontraste zeigten hier, dass Ratschläge bei Vorliegen einer Explanation nicht stärker gewichtet wurden als in der Bedingung ohne Instruktion ($M = .26$, $SD = .20$ vs. $M = .24$, $SD = .20$), $F(1, 89) = 2.72$, $p = .10$, $\eta^2_p = .03$. Jedoch wurden Ratschläge bei Coun-

ter-Explanation weniger stark gewichtet als in der Bedingung ohne Instruktion (*M* = .17, *SD* = .22 vs. *M* = .24, *SD* = .20), *F* (1, 89) = 10.46, *p* < .001, η^2_p = .11, und in Explanation-Bedingung (*M* = .17, *SD* = .22 vs. *M* = .26, *SD* = .20), *F* (1, 89) = 14.63, *p* < .001, η^2_p = .14. Die Interaktion zwischen der Art des Ratgebers und der Instruktion war nicht signifikant, *F* (4, 178) = 1.22, *p* = .30, η^2_p = .03 (siehe Abbildung 8). Im Ergebnis wird daher Hypothese 3 angenommen.

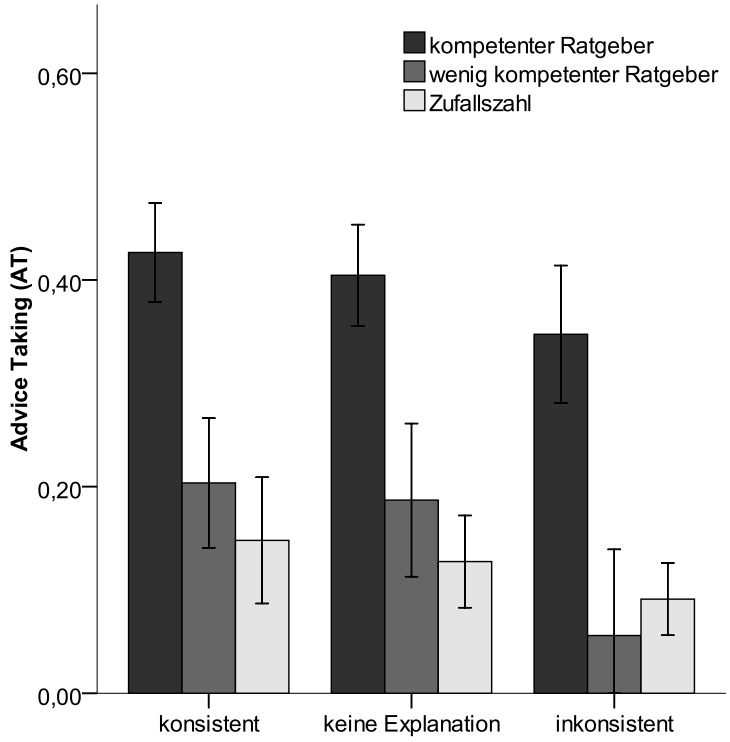


Abbildung 8. Mittlere AT-Werte je nach experimenteller Bedingung in Experiment 5.

Die Wirkung des Innersubjektfaktors Instruktion wurde im nächsten Schritt in einer ANOVA separat für die Bedingung mit Zufallszahl betrachtet. Es zeigten sich signifikante Unterschiede zwischen den drei Bedingungen, *F*(2, 58) = 5.13, *p* = .01, η^2_p = .15. Post-hoc-Kontraste zeigten, dass die Gewichtung der Zufallszahl bei Counter-Explanation geringer ausfiel als in den Durchgängen ohne Instruktion (*M* = .09, *SD* = .09 vs. *M* = .13, *SD* = .12), *F*(1, 29) = 5.64, *p* = .02, η^2_p = .16, und den Durchgängen mit Explanation (*M* = .09, *SD* = .09 vs. *M* = .15, *SD* = .16), *F*(1, 29) = 6.85, *p* = .01, η^2_p = .19, während sich letztere nicht signifikant voneinander unterschieden (*M* = .13, *SD* = .12 vs. *M* = .15, *SD* = .16), *F*(1, 29) = 1.59, *p* = .22, η^2_p = .05. Die Zufallszahl wurde jedoch auch in den inkonsistenten Durchgängen noch systematisch berücksichtigt, *t*(29) = 5.33, *p* < .001, *d* = 0.97. In der Summe bestätigen diese Ergebnisse Hypothese 4, da durch Counter-Explanation das Ausmaß der Übernutzung der Zufallszah-

len reduziert werden konnte, während Explanation keine Verstärkung der Übernutzung bewirkte.

8.5.5 Explorative Analyse möglicher Distanz-Effekte auf die Gewichtung

Auch im Experiment 5 wurden mögliche Distanzeffekte in einer explorativen Analyse untersucht. Die AT-Werte wurden wir in den vorigen Experimenten zu 13 Mittelwerten zusammengefasst. Die Bedingung ohne Ratschläge wurde für die Analyse vernachlässigt, da ohne Ratschläge auch kein Distanzeffekt auftreten kann. Aufgrund einzelner fehlender Werte gingen 75 der 92 Probanden, die entweder einen Ratschlag oder eine Zufallszahl erhalten hatten, in die Analyse ein. Eine 3 (*Art des Ratgebers*: vermeintlich kompetenter Ratgeber vs. vermeintlich wenig kompetenter Ratgeber vs. vermeintliche Zufallszahl) × 3 (*Instruktion*: Explanation vs. Counter-Explanation vs. keine Instruktion) × 13 (Abweichung von der Initialschätzung: 16%, 18%, 20%, 22%, 24%, 26%, 28%, 30%, 32%, 34%, 36%, 38% sowie 40%) ANOVA mit der Art des Ratgebers als Zwischensubjektfaktor sowie der Konsistenz und der Distanz als Innersubjektfaktoren ergab analog zu den bereits berichteten Analysen sowohl einen Haupteffekt des Ratgebers als auch der Instruktion, $F(2, 72) = 40.93$, $p < .001$, $\eta^2_p = .53$, bzw. $F(2, 144) = 11.13$, $p < .001$, $\eta^2_p = .13$. Es zeigte sich wie in den vorigen Experimenten auch ein Haupteffekt für die Distanz, $F(12, 864) = 4.25$, $p < .001$, $\eta^2_p = .06$, der am besten durch einen Kontrast erster Ordnung abgebildet wurde, $F(1, 72) = 16.12$, $p < .001$, $\eta^2_p = .18$. Auch in Experiment 5 wurden Ratschläge, die besonders nahe an der Initialschätzung lagen, weniger stark gewichtet wurden als solche, die weiter entfernt lagen. So wurden Ratschläge, die 16% oder 18% von der Initialschätzung abwichen, im Mittel zu .18 gewichtet, während die übrigen Ratschläge im Mittel zu .23 gewichtet wurden. Wie in Experiment 2 und 3 zeigte sich auch eine Interaktion zwischen Distanz und Ratgeber, $F(24, 864) = 1.66$, $p = .02$, $\eta^2_p = .04$. Aufgrund dieser Interaktion wurden separate Analysen des Distanzeffekts für jeden Ratgebertyp durchgeführt. Für den vermeintlich kompetenten Ratgeber sowie den vermeintlich wenig kompetenten Ratgeber zeigte sich ein Haupteffekt der Distanz, $F(12, 312) = 4.49$, $p < .001$, $\eta^2_p = .15$, bzw. $F(12, 288) = 3.20$, $p < .001$, $\eta^2_p = .12$, die jeweils am besten durch Kontraste erster Ordnung erklärt wurden, $F(1, 26) = 16.67$, $p < .001$, $\eta^2_p = .39$, bzw. $F(1, 24) = 17.05$, $p < .001$, $\eta^2_p = .42$. Für die Zufallszahl zeigte sich kein Effekt der Distanz, $F(2. 44) = 0.88$, $p = .57$, $\eta^2_p = .04$ (siehe Abbildung 9).

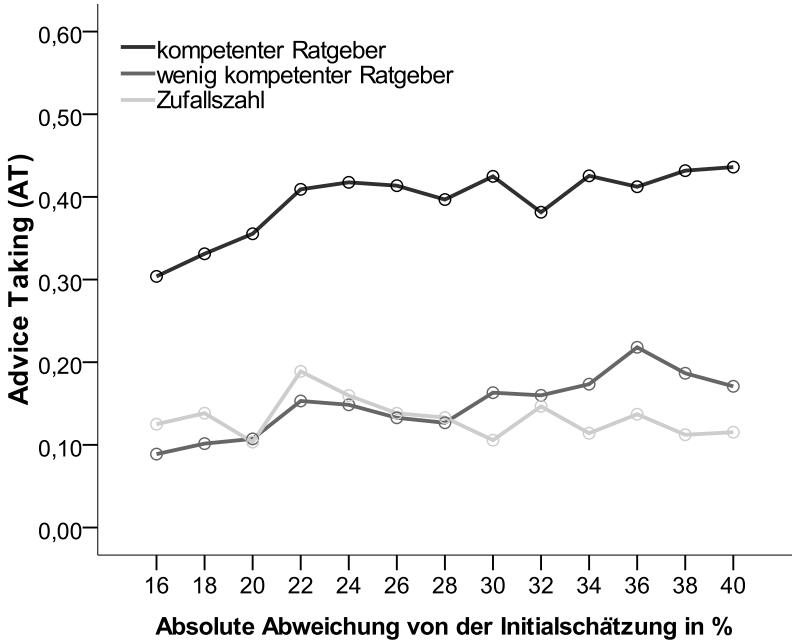


Abbildung 9. Mittlere AT-Werte in Abhängigkeit der Distanz zwischen Initialschätzung und Ratschlag bzw. Zufallszahl nach Art des Ratgebers in Experiment 5.

Schließlich zeigte sich noch eine Interaktion zwischen Konsistenz und Distanz, $F(24, 1728) = 2.48$, $p < .001$, $\eta^2_p = .03$. Separate Analysen nach Art der Instruktion zeigten, dass sowohl für Durchgänge mit Counter-Explanation als auch für Durchgänge ohne Instruktion ein Distanzeffekt auftrat, $F(12, 1020) = 7.28$, $p < .001$, $\eta^2_p = .08$, bzw. $F(12, 996) = 3.14$, $p < .001$, $\eta^2_p = .04$, die jeweils am besten durch einen Kontrast erster Ordnung beschrieben wurden, $F(1, 85) = 30.10$, $p < .001$ bzw. $F(1, 83) = 11.18$, $p < .001$, während in der Explanation-Bedingung kein Distanzeffekt auftrat, $F(12, 996) = 0.82$, $p = .63$, $\eta^2_p = .01$. Warum der Distanzeffekt in der Explanation-Bedingung nicht auftrat, ist jedoch unklar. Die Dreifachinteraktion zwischen Instruktion, Ratgeber und Distanz war nicht signifikant, $F(48, 1728) = 0.03$, $p = .91$, $\eta^2_p = .02$.

8.6 Diskussion

Experiment 5 verfolgte das Ziel, einen potentiellen Erklärungsmechanismus für das Phänomen der Übernutzung nicht valider Ratschläge zu untersuchen, der über die Fehlwahrnehmung der Validität hinausgeht und damit eine subjektive Übernutzung darstellt, nämlich positives Hypothesentesten und selektive Verfügbarkeit von Wissensinhalten. Dazu wurde eine Strategie eingesetzt, die den Prozess des positiven Hypothesentestens beeinflusst, nämlich Explanation bzw. Counter-Explanation. Wenn selektiv Gründe dafür genannt

wurden, dass die erste Schätzung zu niedrig bzw. zu hoch war, und danach eine Zufallszahl dargeboten wurde, die eine Anpassung der Schätzung entgegen der Instruktion nahelegte, wenn also Instruktion und Zufallszahl zueinander inkonsistent waren (Counter-Explanation), fiel der Übernutzungseffekt erwartungsgemäß geringer aus als in den Durchgängen ohne Instruktion. Dies ist ein starker Hinweis darauf, dass positives Hypothesentesten bei der Entstehung der Übernutzung nicht valider Ratschläge eine Rolle spielt, weil durch Counter-Explanation der entsprechende Prozess ausgeschaltet oder zumindest in seiner Stärke reduziert werden konnte. Legten Zufallszahl und Instruktion hingegen eine Anpassung der Finalschätzung in die gleiche Richtung nahe (Explanation), zeigte sich kein nennenswert stärkerer Übernutzungseffekt. Das ist weiterer Hinweis darauf, dass die Übernutzung anteilig durch positives Hypothesentesten entsteht, da die Nicht-Additivität in den konsistenten Durchgängen darauf hindeutet, dass sowohl Explanation als auch Ratschlag denselben Prozess ansteuern. Es ist denkbar, dass durch die zweimalige Anregung des positiven Hypothesentestens der entsprechende Effekt zumindest in geringem Maße verstärkt wird. Rein deskriptiv stützen die Ergebnisse diese Annahme, jedoch konnte der Unterschied nicht statistisch abgesichert werden, was möglicherweise eine Frage der Teststärke ist. In der Summe liefern die Ergebnisse also erste Hinweise darauf, dass der Übernutzung nicht valider Ratschläge ähnliche Prozesse zugrunde liegen wie dem Anker-Effekt, nämlich selektive Aktivierung von Wissensinhalten und positives Hypothesentesten.

Interessanterweise zeigte sich ein auch bei den vermeintlich echten Ratgebern, dass durch Counter-Explanation das Ausmaß reduziert wurde, zu dem Ratschläge gewichtet wurden, während die Gewichtung der Ratgeber in den konsistenten Durchgängen nicht wesentlich zunahm. Dies hat zwei Implikationen. Erstens disqualifiziert es Counter-Explanation als potentielle Interventionsstrategie, da nicht valide Ratschläge im gleichen Maße betroffen sind wie valide Ratschläge. Da aber gerade die validen Ratschläge in der Regel unternutzt werden (Bonaccio & Dalal, 2006; Yaniv, 2004b), würde Counter-Explanation hier zu einer Verschärfung der Problematik führen und das Ausmaß der Unternutzung verstärken. Auch ein differenzierter Einsatz von Counter-Explanation scheint schwer realisierbar, da bei vielen Ratschlägen in der Regel nicht direkt absehbar ist, ob sie valide sind oder nicht (sonst bräuchte man ja auch keine Intervention), und deshalb fraglich ist, wann Counter-Explanation eingesetzt werden sollte und wann nicht. Zweitens ist es ein Hinweis darauf sein, dass ein Teil der Gewichtung valider Ratschläge auch durch selektive Aktivierung und

positives Hypothesentesten entsteht. Dies entspricht auch dem Grundgedanken der Selective-Accessibility-Hypothese (Mussweiler, 2003), da für das Wirken dieser Prozesse irrelevant ist, ob der Referenzwert valide ist oder nicht (Mussweiler & Strack, 1999). Es wäre also denkbar, dass durch positives Hypothesentesten und selektive Verfügbarkeit zunächst jeder Ratschlag zu einem gewissen Grad in das Finalurteil einbezogen wird, und zwar unwillkürlich. Zu dieser unwillkürlichen Sockelnutzung kommt dann noch eine intentionale Gewichtung hinzu, die dann auch sensitiv für die wahrgenommene Validität der Ratschläge sein sollte. Anders ausgedrückt könnte man vermuten, dass die Gewichtung eines Ratschlags mindestens zwei Komponenten beinhaltet, nämlich die intentionale Gewichtung in Abhängigkeit der wahrgenommenen Validität des Ratschlags und die automatisch ablaufende Gewichtung als Resultat positiven Hypothesentestens und selektiver Verfügbarkeit. Der erste Teilprozess wird dabei auch durch die Ergebnisse der vorigen Experimente gestützt, da auch dort zumindest der kompetente Ratgeber und die Zufallszahl immer dann stärker in das Finalurteil integriert wurden, wenn die Probanden auch zu Beginn angeben, diese Ratschläge sollten stärker gewichtet werden.

Träfe die Annahme zu, dass diese beiden Teilprozesse bei der Nutzung von Ratschlägen involviert sind, dann könnte man außerdem den Schluss ziehen, dass die intentionale Gewichtung valider Ratschläge in Wahrheit geringer ist als die tatsächlich beobachtete, und zwar um den Betrag, der durch positives Hypothesentesten und selektive Verfügbarkeit erklärt wird. Das heißt also, dass valide Ratschläge im Sinne intentionaler Gewichtung noch stärker diskontiert werden als bisher angenommen. Mit anderen Worten liegt die tatsächliche Gewichtung valider Ratschläge unter der objektiv optimalen Gewichtung, obwohl die Finalurteile von Personen unwillkürlich bereits in Richtung des Ratschlags angepasst werden.

Ich merke an dieser Stelle aber kritisch an, dass Experiment 5 auch trotz des sehr konsistenten Ergebnismusters aus zwei Gründen keinen erschöpfenden Test der vermittelnden Mechanismen liefern kann. Einerseits können nämlich die postulierten zugrundeliegenden Prozesse entweder nur schwer direkt angesteuert werden, z.B. im Falle selektiver Aktivierung oder selektiver Verfügbarkeit von Wissensinhalten, und andererseits wirken die Methoden, die den anderen Prozess ansteuern, nämlich positives Hypothesentesten, relativ unspezifisch und können auch eine Vielzahl anderer Mechanismen beeinflussen. Vielmehr liefern die Ergebnisse also einen ersten Anhaltspunkt darauf, dass die Übernutzung nicht

valider Ratschläge, soweit sie nicht auf Fehlwahrnehmungen der Validität beruht, durch die Prozesse erklärt werden können, die auch dem Anker-Effekt zugrunde liegen. Damit stellen meine Ergebnisse aber zumindest eine plausible Arbeitshypothese für die weitere Erforschung und systematische Überprüfung der vermittelnden Mechanismen dar, die der Übernutzung von Ratschlägen zugrunde liegen.

9. Abschließende Diskussion

In insgesamt fünf Experimenten konnte erstmals eindeutige Evidenz für die systematische Übernutzung nicht valider Ratschläge erbracht werden. Probanden, die die Rolle des Judges im Judge-Advisor-Paradigma innehatten, passten ihre Finalschätzungen systematisch in Richtung vermeintlicher Zufallszahlen an, obwohl ausdrücklich darauf hingewiesen wurde, dass diese Zahlen keinerlei Informationswert bezüglich der zu schätzenden Größe hätten. Dieses Verhalten trat auf, obwohl sich die Probanden ansonsten näherungsweise rational verhielten. So gaben sie zu Beginn des jeweiligen Experiments an, der vermeintlich kompetente Ratgeber solle stärker gewichtet werden als der vermeintlich wenig kompetente sowie die Zufallszahl, und der wenig kompetente wiederum stärker als die Zufallszahl. Diesen anfänglichen Einschätzungen folgten die Probanden auch weitestgehend. Einerseits gingen höhere intendierte Gewichte mit höheren tatsächlichen Gewichten einher, wobei der Zusammenhang insgesamt moderat war, was ein Hinweis darauf sein könnte, dass sich die Einschätzung der Validität im Laufe des Experiments noch änderte. Andererseits gewichteten die Probanden den vermeintlich kompetente Ratgeber stärker als vermeintlich wenig kompetente Ratgeber oder vermeintliche Zufallszahlen, was in der Summe darauf hinweist, dass die Probanden gegenüber der wahrgenommenen Validität der Ratschläge sensitiv waren und entsprechend handelten. Sowohl dieses in Grundzügen rationale Verhalten als auch die systematische Gewichtung der Zufallszahl zeigten sich konsistent in allen 5 Experimenten, so dass davon ausgegangen werden kann, dass die Übernutzung nicht valider Ratschläge ein robustes Phänomen ist. Die systematische Übernutzung nicht valider Ratschläge stellt damit den Gegenpol zum vielfach dokumentierten Phänomen der Unternutzung von Ratschlägen dar und ergänzt somit das bisherige Verständnis vom Umgang mit Ratschlägen. Nicht nur werden valide Ratschläge in zu geringem Maße berücksichtigt (Bonaccio & Dalal, 2006), sondern auch der gegenteilige Fall kann eintreten, nämlich, dass insbesondere wenig oder gar nicht valide Ratschläge zu stark berücksichtigt werden.

9.1 Ursachen der Übernutzung nicht valider Ratschläge

In meiner Dissertation habe ich erste Hinweise auf die Prozesse geliefert, die einer systematischen Übernutzung nicht valider Ratschläge zugrunde liegen können. Die Befunde sprechen dafür, dass die Übernutzung mindestens zwei Komponenten beinhaltet, nämlich eine intentionale Nutzung aufgrund der Fehlwahrnehmung der Validität einerseits und positives Hypothesentesten sowie daraus resultierende selektive Verfügbarkeit von Wissensinhalten andererseits. Im Bezug auf den ersten psychologischen Prozess, also die Fehlwahrnehmung der Validität, konnte zunächst gezeigt werden, dass einige Personen bereits zu Beginn des Experiments die irrationale Nutzungsüberzeugung aufweisen, eine Gewichtung der Zufallszahl führe zu besseren Schätzungen, weshalb diese Zufallszahlen systematisch in die Finalurteile integriert werden sollten (Experiment 2). Diese Personen unterliegen dann zwar einer Fehleinschätzung der Validität, handeln gemäß dieser Einschätzung aber subjektiv rational, wenn sie die Zufallszahl in dem Maß berücksichtigen, das sie anfänglich als optimal angeben. Allerdings zeigt sich auch, dass die anfängliche Nutzungsintention nur moderat mit der tatsächlichen Gewichtung der Zufallszahl zusammenhängt. Zwar gewichten Probanden, die ursprünglich höhere Gewichte der Zufallszahl als optimal angeben, die Zufallszahl dann später auch stärker, aber die anfängliche Nutzungsintention kann ausgehend von den beobachteten Korrelationen nur 25 bis 30% der Varianz der tatsächlichen Gewichtung aufklären. Dies lässt Spielraum sowohl für die Existenz weiterer Prozesse, aufgrund der sich im Laufe des Experiments die Nutzungsüberzeugung ändert, als auch für nicht-intentionale Prozesse, die zu einer Gewichtung der nicht valider Ratschläge führen.

So wird die Fehlwahrnehmung der Validität anscheinend zusätzlich dadurch verstärkt, dass die vermeintlichen Zufallszahlen in einem plausiblen Wertebereich schwanken. Zwar ist diese Vermutung noch teilweise spekulativ, da im Laufe der Experimente keine weiteren Abfragen der wahrgenommenen Validität der Ratschläge bzw. der Gewichtungsintentionen erfolgten. Jedoch scheint diese Erklärung plausibel, die Zufallszahlen insgesamt weniger stark gewichtet wurden, wenn sie bisweilen unplausible Extremwerte annahmen (Experiment 4). Möglicherweise spielt also die Plausibilität des Wertebereichs, aus dem ein vermeintlicher Ratschlag stammt, eine Rolle bei der Wahrnehmung der entsprechenden Validität. Während bei unplausiblen Werten in der Regel auf geringe Validität geschlossen werden kann, ist der Umkehrschluss nicht zulässig. Dennoch deuten die Ergebnisse von Experiment 4

an, dass die Probanden diesen Umkehrschluss gezogen haben, dass also in der Konsequenz eine Illusion der Validität entstand. Brisant ist dieser Aspekt der Fehleinschätzung der Validität vor allem deshalb, weil Ratschläge in der Realität in den seltensten Fällen unplausible Werte annehmen werden. Zum Beispiel wird kein Finanzberater prognostizieren, dass der DAX am Monatsende bei 50.000 Zählern steht[17], genauso wie kein Arzt seinen Patienten empfehlen wird, täglich minus drei Tabletten Aspirin zu sich zu nehmen. In so fern stellen Zufallszahlen, die innerhalb eines plausiblen Wertebereichs schwanken, eine natürlich vereinfachte, aber dennoch akkurate Annäherung an nicht valide Ratschläge in realen Entscheidungssituationen dar. Das Befundmuster steht zudem Einklang mit Belegen dafür, dass Personen die Zufälligkeit von Ereignissen an der wahrgenommen Variabilität festgemacht wird, wobei mit abnehmender Variabilität auch die subjektiv wahrgenommene Zufälligkeit abnimmt (Kahneman & Tyerksy, 1972; Hastie & Dawes, 2001). Kahneman und Tversky (1972) argumentieren dabei, dass ein Ereignis repräsentativ für ein Zufallsereignis sein muss, um auch als zufällig wahrgenommen zu werden. Dies würde erklären, warum die Zufallszahlen weniger stark gewichtet wurden, wenn sie bisweilen Extremwerte annahmen, also eine viel höhere Variabilität aufwiesen.

Die Fehlwahrnehmung der Validität aufgrund einer anfänglichen irrationalen Nutzungsüberzeugung einerseits und der im Laufe des Experiments entstehenden Illusion der Validität andererseits führen dabei zu einer objektiven, nicht aber einer subjektiven Übernutzung nicht valider Ratschläge. Allerdings zeigte sich auch dann noch ein Übernutzungseffekt, wenn die jeweiligen Probanden zu Beginn des Experimentes angaben, die Zufallszahl sollte nicht berücksichtigt werden, und wenn gleichzeitig die Zufallszahlen auch Extremwerte annahmen. Dieser zusätzliche Übernutzungseffekt könnte in Abgrenzung zur Fehlwahrnehmung der Validität tatsächlich eine subjektive Übernutzung darstellen. Die Ergebnisse von Experiment 5 stützen diese Annahme und liefern erste Anhaltspunkte für den zugrunde liegenden Mechanismus, nämlich positives Testen der fokalen Hypothese, dass die Zielgröße in Richtung des dargebotenen Ratschlags liegt, und daraus resultierend zunächst selektive Aktivierung und später bessere Verfügbarkeit derjenigen Wissensinhalte, die diese fokale Hypothese stützen (Mussweiler, 2003; Mussweiler et al., 2000; Mussweiler & Strack, 1999, 2001). Wie bereits erläutert, handelt es sich dabei um die Prozesse, die dem Selective-

[17] Seinen absolut höchsten Stand im Handelsverlauf (Allzeithoch) erreichte der DAX am 13. Juli 2007 mit 8.151,57 Punkten. Den bisher höchsten Tagesschlusswert erzielte er am 16. Juli 2007 bei 8.105,69 Punkten.

Accessiblity-Modell (Mussweiler, 2003) zugrunde liegen, das davon ausgeht, dass positives Hypothesentesten durch den Vergleich mit einem Referenzwert automatisch und unwillkürlich angestoßen wird, und dass die selektive Verfügbarkeit von Wissensinhalten unbewusst abläuft. Zumindest einer der beiden Teilprozesse, nämlich positives Hypothesentesten konnte direkt über gezielte Instruktionen angesteuert werden, einseitig Gründe dafür zu nennen, dass der wahre Wert der Zielgröße entweder höher liegen könnte als die Initialschätzung oder dass er niedriger liegen könnte. Da die Effekte dieser Instruktion einerseits die Nutzung von Ratschlägen immer dann reduzierte, wenn Ratschlag und Instruktion gegenläufig waren, jedoch keine deutliche Verstärkung der Gewichtung resultierte, wenn sie zueinander konsistent waren, ist es plausibel anzunehmen, dass derselbe Prozess für die jeweiligen Anpassungen verantwortlich ist. Andernfalls hätte man Additivität der Wirkungen von Ratschlag und Instruktion auch in den konsistenten Durchgängen erwartet. Bei der Bildung der Finalschätzung erhalten also allem Anschein nach die bereits voraktivierten Inhalte ein stärkeres Gewicht, so dass im Ergebnis die Finalschätzung entsprechend dieser Inhalte angepasst wird.

In der Summe kann also festgehalten werden, dass die objektive Übernutzung nicht valider Ratschläge mindestens zwei Komponenten aufweist, nämlich eine intentionale Nutzung aufgrund anfänglicher irrationaler Nutzungsüberzeugungen und einer Illusion der Validität sowie eine nicht-intentionale und damit subjektive Übernutzung aufgrund positiven Hypothesentestens und selektiver Verfügbarkeit. Diese konzeptuelle Zweiteilung ist insbesondere für die systematische Untersuchung von Moderatorvariablen und die Entwicklung von Interventionsstrategien interessant, weil die beiden Teilkomponenten voraussichtlich durch völlig unterschiedliche Methoden angesteuert werden müssen.

9.2 Vergleich zwischen Zufallszahl und wenig kompetentem Ratgeber

Zusätzlich zu den bisher beschriebenen Befunden zeigt sich ein weiteres interessantes Muster in den Daten, wenn man die Gewichtung der vermeintlichen Zufallszahl und die Gewichtung des vermeintlich wenig kompetenten Rategebers vergleicht. In allen fünf Experimenten werden beide in etwa gleich stark gewichtet, was sich zum einen darin äußert, dass keine signifikanten Unterschiede zwischen den jeweiligen AT-Werten auftreten, und dass auch deskriptiv keine der beiden Ratgeberarten in allen fünf Experimenten konsistent stärker gewichtet wird als die jeweils andere. Dass auch wenig kompetente Ratgeber zu einem substantiellen Teil, meist zwischen 10% und 25% gewichtet werden, findet sich häufig in der

Literatur zum Judge-Advisor-Paradigma (Harvey & Fischer, 1997; Yaniv, 2004b; Yaniv & Kleinberger, 2000). Die Ergebnisse der hier dargestellten Ergebnisse lassen nun aufgrund der vergleichbaren Gewichtung von Zufallszahl und wenig kompetenten Ratgeber zwei mögliche Schlüsse zu. Die eine Möglichkeit besteht darin, dass auf Basis unterschiedlicher Überlegungen und unterschiedlicher Prozesse, die jeweils spezifisch für einen dieser beiden Ratgebertypen sind, phänomenologisch ein ähnliches Ergebnismuster resultiert. In diesem Falle können die Erkenntnisse zur Übernutzung von Zufallszahlen nicht auf den wenig kompetenten Ratgeber übertragen werden. Alternativ dazu könnte es aber auch sein, dass ähnliche Prozesse und Überlegungen zur Nutzung beider Ratgeber beitragen. In diesem Falle läge einerseits die Vermutung nahe, dass die Probanden auch den vermeintlich wenig kompetenten Ratgeber deutlich stärker gewichtet haben als sie ursprünglich intendierten. Dies lässt sich theoretisch anhand der ursprünglich intendierten Gewichtung überprüfen. Da die intendierte Gewichtung des vermeintlich wenig kompetenten Ratgebers bei allen vier Experimenten in etwa der tatsächlich beobachteten entspricht, spricht dieser Vergleich eher dagegen, dass hier eine subjektive Übernutzung vorlag. Jedoch wies gerade für den vermeintlich wenig kompetenten Ratgeber, anders als für den vermeintlich kompetenten Ratgeber und die Zufallszahl, die intendierte Gewichtung in der Mehrzahl der Experimente keinen systematischen Zusammenhang mit der tatsächlichen Gewichtung auf. Daher ist unklar, ob in diesem Falle die intendierte Gewichtung aussagekräftig ist. Dafür, dass auch der vermeintlich wenig kompetente Ratgeber übernutzt wurde, spricht, dass dessen Gewichtung durch den Einsatz der Counter-Explanation auf ein von Null nicht länger signifikant verschiedenes Niveau reduziert wurde, was nahelegt, dass die Gewichtung des wenig kompetenten Ratgebers mehr noch als die Zufallszahl auf positiven Hypothesentesten und selektiver Verfügbarkeit basierte.

Sollte es tatsächlich der Fall sein, dass der wenig kompetente Ratgeber hier übernutzt wurde, dann hätte dies interessante Implikationen für die Forschung zum Umgang mit Ratschlägen. Bisher ging man nämlich davon aus, dass eine Nutzung selbst extrem wenig kompetenter Ratgeber, auch als „token ammount“ bezeichnet, vor allem auf der Einhaltung sozialer Normen beruht, z.B. um den Ratgeber nicht zu brüskieren (Bonaccio & Dalal, 2006; Harvey & Fischer, 1997). Anhand der vorliegenden Daten kann man als plausible Alternativerklärung für diesen „token ammount“ positives Hypothesentestens und selektive Verfügbarkeit postulieren. Natürlich können soziale Beweggründe bei der Nutzung von Ratschlägen auch

eine Rolle spielen, insbesondere wenn Judge und Advisor real interagieren und der Advisor erfährt, wie stark sein Ratschlag berücksichtigt wird. Allerdings findet eine solche reale Interaktion in den Studien, die eine substantielle Gewichtung auch wenig kompetenter Ratgeber finden, nicht statt, da die Ratschläge wie in den hier berichteten Experimenten nur aus der Darbietung einzelner Werte oder Wertebereiche bestehen (Harvey & Fischer, 1997; Yaniv, 2004b; Yaniv & Kleinberger, 2000). Daher scheinen zumindest für solche Situationen positives Hypothesentesten und selektive Verfügbarkeit die plausiblere Erklärung für die relativ hohe Gewichtung wenig valider Ratschläge zu sein.

9.3 Distanzeffekte auf die Gewichtung der Ratschläge

Obwohl dazu keine spezifischen Hypothesen formuliert wurden, boten die hier berichteten Experimente die Möglichkeit, im Rahmen explorativer Analysen mögliche Effekte der Distanz zwischen Ratschlag und Initialschätzung auf die Gewichtung des jeweiligen Ratschlags zu untersuchen. Hier zeigte sich, wenn auch mit leichten Inkonsistenzen, ein spannendes Muster. Erstens schienen Ratschläge immer dann weniger stark gewichtet zu werden, wenn sie besonders nah an der Initialschätzung lagen. Zweitens trat dieser Effekt bei fast allen Experimenten nur für die beiden vermeintlichen Ratgeber auf, während sich bei der Zufallszahl mehrheitlich kein solcher Distanzeffekt zeigte.

Bezüglich des Distanzeffekts bei den vermeintlichen Ratgebern besteht die Möglichkeit, dass die Unterschiede zwischen Initialschätzung und Ratschlag für besonders geringe prozentuale Abweichungen so klein ausfielen, dass zumindest ein Teil der Probanden Initialschätzung und Ratschlag als hinreichend ähnlich zu ihrer Initialschätzung empfanden und daher nicht mehr adjustierten. Beispielsweise hätte eine Abweichung von 16% bei einer Initialschätzung von 500km einen Ratschlag in Höhe von 580km produziert. Ein solcher Ratschlag könnte nun dazu führen, dass sich der Judge bezüglich der Akkuratheit seiner Initialschätzung bestätigt fühlt, also durch den Ratschlag sozial validiert wird (Festinger, 1954). Ähnliche Befunde zeigen sich beispielsweise in der Einstellungsforschung, wo Konsens häufig mit Akkuratheit gleichgesetzt wird (Chaiken & Stangor, 1987). Das hätte dann zur Konsequenz, dass sich der Judge bezüglich seiner Initialschätzung bestätigt fühlt und deshalb wenig Anlass dazu hat, diese Schätzung noch zu verändern.

Besonders spannend ist der hier beobachtete Distanzeffekt deshalb, weil er dem bisher einzigen Befund zum Einfluss der Distanz zwischen Ratschlag und Initialschätzung auf die

Gewichtung von Ratschlägen widerspricht. So fand Yaniv (2004b), dass Ratschläge mit zunehmender Entfernung zwischen Ratschlag und Initialschätzung weniger stark gewichtet wurden. Hingegen suggerieren die hier berichteten explorativen Analysen, dass die Gewichtung entweder mit zunehmender Distanz steigt, oder aber mit Ausnahme besonders naher Schätzungen auf einem ähnlichen Niveau bleibt. Dieser Widerspruch kann zunächst natürlich eine Folge der experimentellen Manipulation sein. Während in der vorliegenden Arbeit die Distanz schrittweise in einem eher kleinen Beriech, nämlich bis maximal 40% Abweichung, untersucht wurde, wurde in der Studie von Yaniv die Distanz in jeweils drei Stufen mit vermutlich stärkeren Abweichungen variiert (wie stark die Abweichungen bei Yaniv relativ zu den Initialschätzungen waren, lässt sich nicht rekonstruieren). Möglicherweise zeigt sich eine abnehmende Gewichtung also erst dann, wenn die Ratschläge deutlich weiter als 40% von der Initialschätzung abweichen. Eine detaillierte Untersuchung der Distanzeffekte über eine weite Bandbreite von Entfernungen und unter Berücksichtigung der Möglichkeit, dass besonders geringe Abweichungen zu sozialer Validierung führen könnten, scheint daher vielversprechend. Letzteres würde dabei implizieren, dass zusätzlich zu den bloßen Initial- und Finalschätzungen auch jeweils die subjektive Sicherheit des Judges, dass die jeweilige Schätzung akkurat ist, erfasst werden sollte. Sollte tatsächlich soziale Validierung dafür verantwortlich sein, dass Ratschläge, die nur wenig von der Initialschätzung abweichen, weniger stark gewichtet werden, dann sollte sich bei diesen Ratschlägen gleichermaßen eine Erhöhung der subjektiven Sicherheit zeigen.

Der zweite spannende Befund, nämlich, dass Distanzeffekte selektiv nur für die vermeintlichen Ratgeber auftraten, gibt Hinweise darauf, dass sich der vermeintlich wenig kompetente Ratgeber und die Zufallszahl in der Wahrnehmung der Probanden unterscheiden müssen, obwohl sie im Mittel etwa gleich stark gewichtet wurden. Worin diese Unterschiede bestehen, kann hier nicht geklärt werden. Jedoch erscheint plausibel, dass ein sozialer Validierungseffekt nur dann eintritt, wenn die Quelle der Validierung menschlich ist, während Zufallszahlen nicht zur Validierung dienen können.

9.4 Beschränkungen und Implikationen für weitere Forschung

Die Ergebnisse der Experimente, die ich im Rahmen meiner Dissertation präsentiert habe, bieten einige Ansatzpunkte für neue Fragestellungen und weiterführende Forschung, die zum Teil auch aus den methodischen Beschränkungen der Experimente resultieren. So

wurden hier zugunsten der experimentellen Kontrolle keine echten Ratschläge, sondern prozentuale Abweichungen von den Initialschätzungen dargeboten, die in jeweils der Hälfte der Fälle unter bzw. über dem Ratschlag lagen. Als Konsequenz war selbst bei einer deutlichen Nutzung vermeintlicher Zufallszahlen nicht zu erwarten, dass die Akkuratheit der Probanden dadurch leiden würde. In realen Situationen scheint es jedoch plausibel, dass eine Übernutzung wenig oder nicht valider Ratschläge ebenso zu einer Beeinträchtigung der Urteilsqualität führen kann wie die Diskontierung valider Ratschläge (Yaniv, 2004a, 2004b). Folglich wäre es äußerst wichtig, in Folgeuntersuchungen entweder echte Ratschläge mit geringer Validität darzubieten oder die nicht validen Ratschläge so zu gestalten, dass deren systematische Übernutzung sich auch potentiell auf die Akkuratheit der Finalschätzungen auswirken kann. Letzteres könnte man beispielweise umsetzen, indem die Ratschläge aus einer Normalverteilung um den wahren Wert gezogen werden, wobei entweder die Varianz so groß gewählt wird, dass eine Berücksichtigung der Ratschläge durch den Judge im Mittel zu weniger akkuraten Finalschätzungen führt, oder indem zusätzlich zu einer Varianz der Ratschläge noch ein Bias eingebaut wird, so dass die Berücksichtigung der Ratschläge die Finalschätzungen systematisch vom wahren Wert wegführt.[18] Sollte sich auch dann zeigen, dass wenig oder nicht valide Ratschläge übernutzt werden und dass dies tatsächlich zu weniger akkuraten Finalschätzungen führt, dann wäre dringend geboten, effektive Interventionen zu entwickeln, die in der Lage sind, einer Übernutzung von Ratschlägen entgegenzuwirken. Gleichzeitig aber müssten sie so differenziert wirken, dass es nicht zu einer verstärkten Diskontierung valider Ratschläge kommt. Die in Experiment 5 eingesetzte Strategie der Counter-Explanation wäre in diesem Sinne also kontraproduktiv, dass sie zwar eine Übernutzung nicht valider Ratschläge reduzieren kann, aber in gleichem Maße auch die Gewichtung derjenigen Ratschläge vermindert, die ohnehin schon zu wenig berücksichtigt werden.

Eine zweite Einschränkung der berichteten Experimente besteht darin, dass Judge und Advisor nicht direkt interagierten. Eine solche Interaktion wäre aber notwendig, wenn man überprüfen möchte, in wie weit sozialer Druck oder soziale Normen einen Einfluss auf die Nutzung wenig oder nicht valider Ratschläge nehmen. Wie bereits gesagt, erscheint die Erklärung der Übernutzung nicht valider Ratschläge über sozialen Druck oder soziale Normen

[18] Selbstverständlich können auch andere Verteilungen als eine Normalverteilung für die Generierung nicht oder wenig valider Ratschläge genutzt werden. Jedoch bietet sich eine Normalverteilung insbesondere wegen des im Folgenden Abschnitt beschriebenen normativen Modells an.

(Harvey & Fischer, 1997) in Situationen ohne soziale Interaktion als eher unwahrscheinlich. Jedoch ist denkbar, dass sich durch solche Prozesse der Übernutzungseffekt noch verstärkt, wenn Judge und Advisor real interagieren. Dies zu überprüfen scheint mir besonders deshalb wichtig, weil in realen Entscheidungssituationen die Interaktion zwischen Judge und Advisor die Regel ist und der Advisor häufig direkt beobachten kann, ob bzw. in welchem Ausmaß sein Ratschlag berücksichtigt wird. Die systematische Untersuchung solcher Situationen könnte also über die beiden hier demonstrierten Komponenten weitere Gründe für eine Übernutzung von Ratschlägen aufdecken.

Die dritte Beschränkung besteht darin, dass nur ein Aufgabentyp verwendet wurde, denn es ist zumindest theoretisch möglich, dass die Befunde der hier beschriebenen Experimente aufgabenspezifisch sind. Diese Möglichkeit erscheint zwar eher unwahrscheinlich, da mit Hinblick auf den vermeintlich kompetenten und den vermeintlich wenig kompetenten Ratgeber die hier berichteten Ergebnisse sehr gut mit früheren Befunden vergleichbar sind, die gänzlich andere Aufgaben verwendeten (z.B. Harvey & Fischer, 1997; Soll & Larrik, 2009; Yaniv, 2004b; Yaniv & Kleinberger, 2000, Yaniv & Milyawsky, 2007). Dennoch wäre eine Replikation mit anderen Aufgaben wünschenswert. Dies würde zudem die Möglichkeit öffnen zu untersuchen, ob das Ausmaß der Übernutzung in Abhängigkeit bestimmter Aufgabencharakteristika variiert. Gerade bezüglich der Übernutzung aufgrund von positivem Hypothesentesten und selektiver Aktivierung wäre denkbar, dass mit sinkender Expertise der Probanden der Grad der Übernutzung steigt, da sich ein ähnliches Muster auch beim Anker-Effekt zeigt (Wilson et al., 1996).

Die vierte Beschränkung betrifft die Erfassung der subjektiven Nutzungsüberzeugung. Die subjektiv optimale Gewichtung wurde nur zu Beginn des Experiments abgefragt, weshalb unklar ist, in welchem Ausmaß sie sich im Verlaufe des Experiments änderte. Das bedeutet auch, dass das Ausmaß, in dem die Illusion der Validität die Nutzungsintention beeinflusst hat, nicht quantifiziert werden konnte, da hierzu die Entwicklung der wahrgenommenen Validität über den Verlauf des Experiments hätte beobachtet werden müssen. In der Konsequenz wäre es also notwendig, in Folgestudien die Nutzungsintention der Probanden nicht nur zu Beginn des Experiments zu erfassen.

Eine letzte Beschränkung besteht darin, dass in den hier dargestellten Experimenten stets Zufallszahlen benutzt wurden, um die Übernutzung von Ratschlägen zu überprüfen.

Dies war erforderlich, da Zufallszahlen wegen ihrer Nullvalidität einen einfachen objektiven Nachweis der Übernutzung erlauben. Analog zu früheren Arbeiten, die stets eine bestimmte Ausgestaltung von Ratschlägen nutzten, um die Unternutzung von Ratschlägen anhand von Plausibilitätsüberlegungen zu demonstrieren (Harvey & Fischer, 1997; Soll & Larrik, 2006, 2009; Yaniv, 2004b; Yaniv & Kleinberger, 2000, Yaniv & Milyawsky, 2007), wurde hier der Plausibilitätsschluss gezogen, dass jede Gewichtung, die von Null abweicht, eine Übernutzung darstellt. Damit bleiben aber zunächst mögliche Übernutzungseffekte solcher Ratschläge unberücksichtigt, die zwar eine niedrige, aber dennoch von Null verschiedene Validität aufweisen. Gerade diese Ratschläge zu untersuchen, wäre für zukünftige Forschung aus zwei Gründen relevant. Erstens sind solche Ratschläge in bestimmten realen Anwendungsbereichen ökologisch valider als Zufallszahlen mit einer Nullvalidität, das heißt, auch wenig kompetente Ratgeber können in der Praxis zumindest in geringem Maße überzufällig gute Ratschläge erteilen. Zweitens wäre eine objektive Betrachtung des Ausmaßes der Über- oder Unternutzung von Ratschlägen in Abhängigkeit ihrer Validität spannend. Wie die hier dargestellten Ergebnisse andeuten, könnte sich bei der Nutzung von Ratschlägen das Phänomen der Unternutzung mit abnehmender Validität graduell über eine normativ korrekte Nutzung bis hin zur Übernutzung entwickeln. Das Gesamtbild könnte also gerade darin bestehen, dass die faktischen Gewichtungen die durch die Variabilität der Validitäten nahegelegte Spannweite nicht ausnutzen. Eine solche systematische Untersuchung setzt jedoch voraus, dass für jede Ausprägung der Validität der Ratschläge normativ die korrekte Gewichtung bestimmt werden kann, zum Beispiel mit Hilfe eines normativen Modells. Bisher jedoch existiert kein solches Modell. Eine erste Idee für ein solches normatives Modell, das die objektiv optimale Gewichtung eines Ratschlags in Abhängigkeit von bestimmten Parametern festlegt, werde ich im folgenden Abschnitt darstellen.

Die hier präsentierten Befunde liefern auch über die direkten Konsequenzen aus den methodischen Beschränkungen hinaus einen fruchtbaren Nährboden für weitere Forschung. Die naheliegende Implikation besteht natürlich darin, Moderatoren und Mediatoren des Übernutzungseffekts zu untersuchen, um noch verlässlichere Aussagen über die vermittelnden Mechanismen und die Bedingungen seines Auftretens machen zu können. Dabei kann die Annahme, dass die Übernutzung von Ratschlägen aus intentionaler Nutzung aufgrund Fehlwahrnehmungen der Validität und nicht-intentionaler Gewichtung im Sinne des Selective-Accessibility-Modells (Mussweiler, 2003) resultiert, als Ausgangspunkt dienen. Ei-

nen wichtigen Beitrag kann die Folgeforschung aber für die Theoriebildung sozialer Entscheidungsprozesse leisten. Bisher gibt es nämlich nur ein einziges Modell, das sich mit der Nutzung von Ratschlägen befasst (Jungermann, 1999), und dieses Modell ist so spezifisch, dass es für den Großteil der Judge-Advisor-Situationen schlichtweg nicht anwendbar ist. Das bisherige Verständnis vom Umgang mit Ratschlägen baut maßgeblich auf der Idee auf, dass Ratschläge generell diskontiert würden (Bonaccio & Dalal, 2006, Harvey & Fischer, 1997, Yaniv, 2004a, 2004b), und obwohl diese Grundidee bisher nicht im Rahmen eines Modell spezifiziert wurde, legt sie die grundsätzliche Vermutung nahe, die Gewichtung von Ratschlägen bestehe aus der normativ korrekten Gewichtung anhand der Validität eines Ratschlags abzüglich der Diskontierung. Dem entgegen suggerieren die hier dargestellten Ergebnisse, dass erstens die Diskontierung nicht immer stattfindet, sondern unter gewissen Bedingungen auch ins Gegenteil umschlagen kann, und dass zweitens eine anscheinend unbewusste Adjustierung in Richtung eines Ratschlags stattfindet, die einem Anker-Effekt ähnelt. Die hier präsentierten Ergebnisse erweitern also die Erkenntnisgrundlage für ein theoretisches Modell zum Umgang mit Ratschlägen.

9.5 Ein normatives Modell zur Bestimmung des optimalen Gewichts

Eine wichtige Voraussetzung, um die Phänomene der Übernutzung und der Unternutzung von Ratschlägen zukünftig besser zu verstehen und vor allem deren ökologische Validität kritischer zu prüfen, ist ein genaues Verständnis davon, wie stark ein Ratschlag aufgrund seiner Qualität tatsächlich gewichtet werden sollte. Für wenige ganz spezifische Situationen kann dies über Plausibilitätsüberlegungen bestimmt werden, zum Beispiel wenn Judge und Advisor exakt gleich kompetent sind (Soll & Larrik, 2006, 2009) oder wenn – wie in dieser Arbeit – der Ratschlag per definitionem eine Validität von Null aufweist. Für den Großteil der Interaktionen zwischen Judge und Advisor in der Realität sind diese Überlegungen aber nicht anwendbar. Deshalb möchte ich diese Arbeit mit einer Idee für ein normatives Modell abschließen, mit dessen Hilfe man die normativ korrekte Gewichtung eines Ratschlags bestimmen kann.[19]

[19] Dieses Modell stellt eine Zusammenarbeit mit Johannes Schmidt-Hieber dar, dem ich auch an dieser Stelle noch einmal ausdrücklich für seine wertvolle Unterstützung und seine Geduld danken möchte.

9.5.1 Herleitung des Modells

Die Grundannahme des Modells ist, dass ein Ratschlag dann optimal gewichtet wird, wenn er zu einer möglichst akkuraten Finalschätzung führt. Das entscheidende Kriterium ist also allein die Genauigkeit der Schätzung, während soziale Aspekte in diesem Modell vernachlässigt werden. Generell gilt, dass die Gewichtung eines Ratgebers davon abhängt, wie kompetent er bezüglich der abzugebenden Schätzungen im Verhältnis zum Judge ist. Um also die optimale Gewichtung zu bestimmen, muss zunächst bekannt sein, wie akkurat die Schätzungen von Judge und der Advisor bei der jeweiligen Aufgabe sind. Das Modell bildet dabei die Kompetenz von Judge und Advisor als voneinander unabhängige Normalverteilungen ab.[20] Die Annahme der Normalverteilung stützt sich dabei zunächst auf theoretische Überlegungen von Steiner (1966) sowie von Einhorn, Hogarth und Klempner (1977), die bei der Formalisierung ihrer Modelle von Gruppenschätzungen davon ausgingen, dass die Schätzungen der einzelnen Gruppenmitglieder normalverteilt sind. Ob die Individualschätzungen einer Person tatsächlich normalverteilt sind, kann anhand bisheriger Forschung zwar nicht eindeutig belegt werden, jedoch gibt es Hinweise darauf, dass die Verteilung der individuellen Schätzungen zumindest symmetrisch sein muss. So konnte Stroop (1932) zeigen, dass die Schätzungen eines Individuums akkurater werden, wenn es dieselbe Aufgabe mehrfach bearbeitet und dann einfach der Mittelwert über die jeweiligen Schätzungen gebildet wird. Mit zunehmender Anzahl an Schätzungen derselben Aufgabe wurde der Mittelwert immer akkurater, bis bei ca. 200 Durchgängen der Mittelwert fast exakt dem wahren Wert entsprach. Eine ähnliche Beobachtung findet sich bei Herzog und Hertwig (2009), die zeigen können, dass die Zweifache Schätzung desselben Sachverhalts durch Mittelwertbildung der beiden Schätzungen im Mittel akkurater ist als eine einfache Schätzung.

Seien also zwei individuelle Schätzungen X_1 und X_2 unabhängig normalverteilt mit Varianzen σ^2_1 und σ^2_2 sowie Bias (Erwartungswert) b_1 und b_2. Der Bias b ist definiert als Abweichung dessen, was eine Person im Mittel schätzt, von einem wahren Wert t, also

, (1)

[20] Die Unabhängigkeit der Schätzungen von Judge und Advisor wird in der bisherigen Forschung häufig als gegeben vorausgesetzt (für Ausnahmen siehe Yaniv, 2004b). Daher wird auch in diesem Modell zunächst Unabhängigkeit der Schätzungen unterstellt. Bei späterem Bedarf kann das Modell jedoch auch so erweitert werden, dass Abhängigkeit der Schätzungen Berücksichtigung findet.

wobei der Bias ***b*** also angibt, in welchem Ausmaß eine Person systematisch zu hohe oder zu niedrige Schätzungen abgibt. Die Varianz σ^2 hingegen repräsentiert den unsystematischen Fehler der Person, also eine Art fehlende Reliabilität der Schätzungen. Sei X_1 die individuelle Schätzung des Judge und X_2 die Schätzung des Advisors, dann ergibt sich:

(2)

Die Finalschätzung ***Y*** kann dann wiederum als gewichtetes Mittel dieser beiden Variablen aufgefasst werden:

(3)

Gesucht ist nun die optimale Gewichtung von X_1und X_2, die dazu führt, dass die Kombination der beiden Schätzungen ***Y*** zu möglichst akkuraten Finalschätzungen führt. Dabei ist entscheidend, wie weit die kombinierten Schätzungen im Mittel vom wahren Wert abweichen. Um das Problem der optimalen Gewichtung zu lösen, bietet es sich an die mittlere quadratische Abweichung von ***Y*** bezogen auf den wahren Wert ***t*** zu minimieren, also den ***MSE*** (Mean Sqaured Error). Zwar können theoretisch auch andere Optimalitätskriterien als der ***MSE*** herangezogen werden, jedoch bietet der ***MSE*** den großen Vorteil, dass er sich einfach additiv zerlegen lässt. Der ***MSE*** des gewichteten Mittelwerts ***Y*** aus zwei Schätzungen X_1 und X_2 lässt sich demnach zerlegen in die Fehlerstreuung der ersten Schätzung, die Fehlerstreuung der zweiten Schätzung und den Bias der kombinierten Schätzung:

(4)

Die Gewichte w_1 und w_2 sollen jetzt so gewählt werden, dass ***MSE(Y)*** minimal wird. Zu diesem Zweck wird folgende Annahme eingeführt: Die Gewichte der beiden Schätzungen ergeben in der Summe 1. Diese Annahme ist konsistent mit der Beobachtung, dass Ratschläge in ca. 95% der Fälle zwischen 0 und 1 gewichtet werden (z.B. Gino, 2008; Gino et al., 2008; Yaniv, 2004). Damit lässt sich das Gewicht der zweiten Schätzung wie folgt darstellen.

(5)

Setzt man (5) in (4) ein, so ist ***MSE(Y)*** für gegebene Erwartungswerte und Varianzen der beiden Verteilungen nur noch von w_1 abhängig.

(6)

Durch Auflösung des Bias-Terms erhält man:

(7)

Um den Wert von w_1 zu ermitteln, für den ***MSE(Y)*** minimal wird, wird (7) nach w_1 abgeleitet.

(8)

Da sämtliche Parameter und Variablen, von denen ***MSE(Y)*** abhängt, positiv sind und der ***MSE*** ein quadratisches Abweichungsmaß ist, kann das Minimum von (7) bestimmt werden, indem (8) gleich Null gesetzt wird.

(9)

Durch Umformung von (9) ergeben sich:

(10)

(11)

Stellt man (11) nach w_1 um, so erhält man das Gewicht w_1, Für das der ***MSE(Y)*** in Abhängigkeit der Parameter der Individualschätzungen X_1 und X_2 minimal ist, nämlich:

(12)

Wegen (5) ergibt sich automatisch:

(13)

(13) ist äquivalent zu:

(14)

Durch diese einfachen Formeln kann also die normativ korrekte Gewichtung eines Advisors bestimmt werden, sofern die entsprechenden Modellparameter bekannt sind. Das bedeutet, es muss bekannt sein, ob und in welchem Ausmaß Judge und Advisor einem Bias unterliegen, und wie reliabel deren Schätzungen jeweils sind. Da diese Daten in der Regel aber nicht verfügbar sind, müssen sie anhand vorheriger Performanz von Judge und Advisor

geschätzt werden. Wichtig bei der Verwendung dieses Modells ist, dass es nur Aussagen darüber macht, wie stark ein Ratgeber im Mittel gewichtet werden sollte. Es kann dabei durchaus vorkommen, dass bei einzelnen Durchgängen andere Gewichtungen zu einer akkurateren Schätzung geführt hätten als diejenigen, die das Modell ermittelt hat, jedoch würden diese alternativen Gewichtungen, wenn man sie über mehrere Durchgänge konstant hielte, in einer geringeren Akkuratheit resultieren. Mit anderen Worten sollte also der Judge bei jedem einzelnen Durchgang genau die optimale Gewichtung wählen, die das Modell berechnet, wohlweißlich der Tatsache, dass damit in einem bestimmten Einzelfall eine suboptimale Gewichtung gewählt werden kann, dass jedoch auf lange Sicht, diese Gewichtung ohne zusätzliche Informationen, die in diesem Modell nicht berücksichtigt sind, nicht zu schlagen ist.

9.5.2 Zwei Beispielrechnungen für das normative Modell

Im Folgenden soll anhand zweier einfacher Beispiele zum einen deutlich gemacht werden, dass das hier dargestellte normative Modell denjenigen Fall korrekt abbildet, der bisher im Rahmen im Rahmen einer Plausibilitätsannahme dazu diente, eine Unternutzung von Ratschlägen abzubilden, nämlich gleiche Kompetenz von Judge und Advisor (Soll & Larrick, 2006, 2009). Zweitens soll demonstriert werden, dass für eine Konstellation der Kompetenz von Judge und Advisor, für die anhand eines Plausibilitätsschlusses keine normative Gewichtung abgeleitet werden kann, eine exakte Aussage darüber trifft, wie der Ratgeber gewichtet werden soll, um eine möglichst akkurate Finalschätzung abzugeben.

Im ersten Beispiel sind Judge und Advisor gleich kompetent. Dies ist gleichbedeutend damit, dass sie sich sowohl in Hinsicht auf den Bias als auch in Hinsicht auf den unsystematischen Fehler gleichen. Welche Werte dabei konkret für diese beiden Parameter gewählt werden, ist unerheblich, solange sie für beide Verteilungen identisch sind. Für das entsprechende Beispiel soll daher der Einfachheit halber sowohl der Bias als auch der unsystematische Fehler jeweils 1 betragen. Setzt man die Werte in Gleichung (12) ein, so ergibt sich das optimale Gewicht der Initialschätzung:

Das Modell macht also genau die Vorhersage, die man bereits auf Basis des Plausibilitätsschlusses erwartet hätte, nämlich eine Gewichtung des Judges und des Advisors zu je-

weils 50%. Wie Abbildung 10 zeigt, nimmt der ***MSE*** der Finalschätzung ***Y*** für die Gewichtung von 50% in der Tat den niedrigsten Wert an, und je weiter sich die Gewichtung von diesem Optimum entfernt, desto stärker steigt die Ungenauigkeit der resultierenden Finalschätzung.

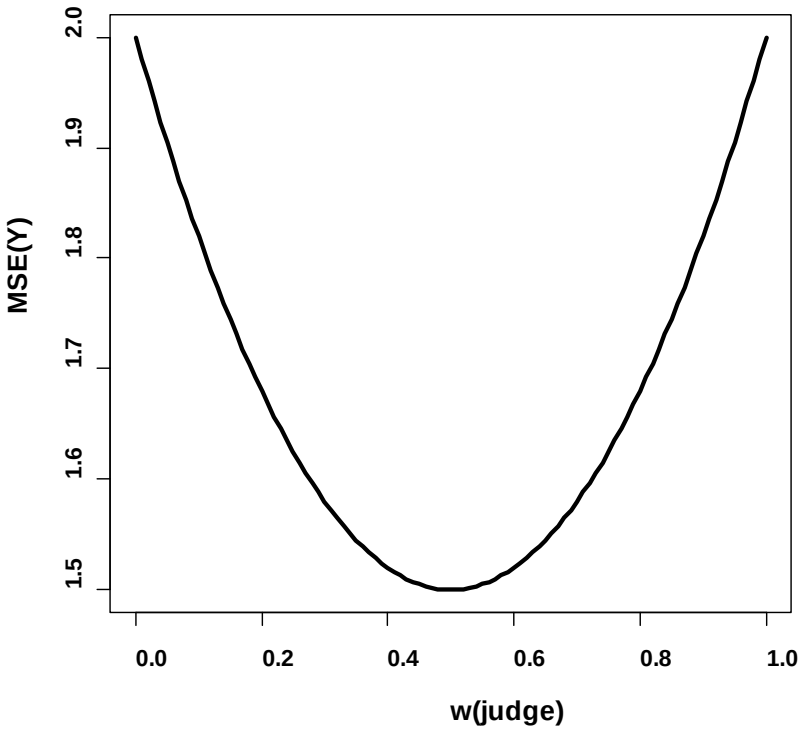


Abbildung 10. Mittlerer quadratischer Fehler der Finalschätzung in Abhängigkeit der Gewichtung der Initialschätzung des Judges bei gleicher Kompetenz von Judge und Advisor.

Nachdem das Modell also gegenüber den bisher in der Forschung verwendeten Plausibilitätsannahmen keinen Nachteil hat, soll im zweiten Beispiel die Bestimmung des optimalen Gewichts für eine Situationen demonstriert werden, für die eine Plausibilitätsüberlegung keine exakte Antwort bieten kann. Für dieses Beispiel sei der Bias des Judges 0,5 und seine unsystematische Varianz sei 1, das heißt der Judge überschätzt mit gewisser Reliabilität den wahren Wert im Ausmaß 0,5. Der Advisor hingegen unterliegt keinem Bias, unterliegt jedoch einem unsystematischen Fehler von 2. Durch Einsetzen in Gleichung (12) erhält man auch für diese Konstellation das optimale Gewicht des Judges:

Für diesen Fall, also einen im Vergleich zum Advisor relativ reliablen Judge mit einem geringen Bias von 0,5 und einen ungebiasten Advisor mit einem im Vergleich zum Judge doppelt so großen unsystematischem Fehler der Größe 2, sollte der Judge zu 76% und damit der Advisor zu 24% gewichtet werden, um im Mittel die bestmöglichen Finalschätzungen zu erzielen. Auch für diesen Fall führt jede Veränderung des Gewichtes im Mittel zu einer Verschlechterung der Akkuratheit der Finalschätzung.

9.5.3 Erweiterbarkeit des Modells auf mehrere Ratgeber

Das normative Modell ist in seiner hier dargestellten Form anwendbar auf Situationen, in denen ein Judge und genau ein Advisor unabhängige Schätzungen abgeben. Das bedeutet, dass das Modell dann nicht anwendbar ist, wenn der Judge von mehr als einem Advisor Ratschläge empfängt. Es kann jedoch so verallgemeinert werden, dass für eine beliebige Anzahl an Advisors deren jeweils optimale Gewichtung bestimmbar ist. Dazu muss zunächst für die Situation mit einem Judge und beliebig vielen Ratgebern der ***MSE*** der Finalschätzung ***Y*** definiert werden. Dabei soll die Anzahl der insgesamt abgegebenen Schätzungen, also Initialschätzung sowie jeder einzelne Ratschlag, als ***i*** bezeichnet werden:

Anstelle eines einzelnen Wertes für ein Gewicht w_1 würde dann eine Anzahl von Gewichten w_1 bis w_{i-1} bestimmt werden, die analog zum Basismodell in der Summe 1 ergeben müssen (das Gewicht der ***i***-ten Schätzung ist dann durch die Rahmenbedingung, dass die Summe der Gewichte 1 ergeben muss, bereits definiert). Um den ***MSE(Y)*** zu minimieren, müssten dann ferner partielle Ableitungen für jedes der ***i-1*** Gewichte durchgeführt werden, deren Resultat dann ein ***i-1***-Tupel mit den jeweils optimalen Gewichten ist. Analog zur Basisvariante des normativen Modells führen die so bestimmten Gewichte für die Initialschätzung und die Ratschläge im Mittel zu den akkuratesten Finalschätzungen.

9.5.4 Einsatzmöglichkeiten in der Forschung zum Umgang mit Ratschlägen

Der wohl offensichtlichste Vorteil des normativen Modells, das ich hier beschrieben habe, besteht darin, dass genau bestimmt werden kann, ob Ratschläge – gemessen an den Vorgaben des Modells – zu stark oder zu wenig berücksichtigt werden. Daraus ergibt sich direkt die erste Einsatzmöglichkeit des Modells, nämlich die genaue Bestimmung des Aus-

maßes, zu dem valide Ratschläge unternutzt und wenig valide Ratschläge übernutzt werden. Für Laborexperimente ist es dabei sogar möglich, Ratschläge zu simulieren, indem man Werte aus einer Normalverteilung zieht. Die beiden Parametern dieser Verteilung, nämlich Erwartungswert und Varianz, entsprechen dabei genau den beiden Parametern, die das normative Modell für die Berechnung der optimalen Gewichtung benötigt. Durch Simulation der Ratschläge auf diese Art wären also bereits zwei der vier notwendigen Parameter bekannt und müssten nicht aus den Daten geschätzt werden. Sofern dann der Judge hinreichend viele Schätzungen für Aufgaben desselben Typs abgibt, um dessen Bias und unsystematische Fehlervarianz zu schätzen, kann dann die von dem Modell berechnete optimale Gewichtung mit der durchschnittlichen tatsächlichen Gewichtung im Sinne des AT-Werts vergleichen werden, um das Ausmaß der Fehlnutzung zu berechnen.

Eine zweite und weniger offensichtliche Einsatzmöglichkeit des normativen Modells besteht darin, es als Unterstützungsmaßnahme für die Urteilsbildung einzusetzen. Sofern für einen Judge und einen Advisor bereits hinreichend viele Schätzungen vorliegen, um die Parameter des Modells zu schätzen, kann dem Judge mitgeteilt werden, wie stark er die Ratschläge gewichten sollte. Hier kann dann nicht nur überprüft werden, ob sich dadurch tatsächlich die Urteilsqualität verbessert, sondern vor allem, ob bzw. in welchem Ausmaß Judges willens und in der Lage sind, diese Hilfestellung anzunehmen. Falls sie das nicht oder nur in unzureichendem Maße tun, könnte weiterhin untersucht werden, aus welchem Grund die Hilfestellung nicht oder zu gering wird.

9.6 Abschließende Bemerkung

In meiner Dissertation habe ich erstmals empirische Evidenz für die systematische Übernutzung nicht valider Ratschläge erbracht. Ein solcher Erstbefund wirft naturgemäß mindestens ebenso viele Fragen auf, wie er beantwortet. Tritt eine solche Übernutzung tatsächlich in der Realität auf? Wenn ja, wie sehr beeinträchtigt sie die Entscheidungs- oder Urteilsqualität? In welchen Situationen ist sie besonders stark ausgeprägt, wann eher schwach? Sind bestimmte Personen anfälliger für eine Übernutzung als andere? Wie kann man die Übernutzung schlechter Ratschläge gezielt vermeiden oder zumindest reduzieren?

Vor allem aber zeigen meine Befunde, dass soziale Einflüsse berücksichtigt werden müssen, wenn menschliches Urteilen und Entscheiden untersucht werden. Menschen fällen Urteile und Entscheidungen nicht in einem Vakuum, sie tun dies in einem sozialen Kontext,

innerhalb dessen die Vorschläge und Meinungen anderer Personen einen nicht zu vernachlässigenden Einfluss auf Urteilen und Entscheiden ausüben. Die Einflüsse dieses sozialen Kontexts können allem Anschein nach nicht vollständig ignoriert werden, selbst wenn dies von einem rationalen Standpunkt her erforderlich wäre. Wenn wir also ein besseres Verständnis davon erlangen wollen, wie Menschen urteilen und entschieden, dann müssen wir auch mehr darüber lernen, wie Menschen mit Ratschlägen umgehen.

Literatur

Anderson, C. A., Sechler, E. S. (1986). Effects of Explanation and Counterexplanation on the Development and Use of Social Theories. *Journal of Personality and Social Psychology, 50,* 24-34.

Ariely, D., Loewenstein, G., & Prelec, D. (2003). Coherent arbitrariness: Stable demand curves without stable preferences. *Quarterly Journal of Economics, 118,* 73-105.

Asch, S. E. (1956). Studies of independence and conformity. A minority of one against a unanimous majority. *Psychological Monographs, 70* (9, Whole No. 416).

Bazerman, M. H., & Neale, M. A. (1992). *Negotiating Rationally*. New York, NY: Free Press.

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101,* 127-151.

Brehmer, B., & Hagafors, R. (1986). The use of experts in complex decision-making: A paradigm for the study of staff work, *Organizational Behavior and Human Decision Processes, 38*, 181-195.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34,* 571-582.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243,* 1668-1674.

Einhorn, H. J., Hogarth, R. M., & Klempner, E. (1977). Quality of group judgment. *Psychological Bulletin*, *84*, 158-172.

Feng, B., & MacGeorge, E. L. (2006). Predicting receptiveness to advice: Characteristics of the problem, the advice-giver, and the recipient. *Southern Communication Journal, 71,* 67-85.

Festinger, L. (1954). A theory of social comparison processes. *Human Relations, 7,* 117–140.

Gardner, P.H., & Berry, D.C. (1995). The effect of different forms of advice on the control of a simulated complex system. *Applied Cognitive Psychology, 9,* S55-S79.

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

Kahneman, D., & Tversky, A. (Eds.). (2000). *Choices, values and frames*. New York : Cambridge University Press and the Russell Sage Foundation.

Gino, F. (2008). Do we listen to advice just because we paid for it? The impact of advice cost on its use. *Organizational Behavior and Human Decision Processes, 107,* 234-245.

Gino, F., & Schweitzer, M. E. (2008). Blinded by anger or feeling the love: how emotions influence advice taking. *Journal of Applied Psychology, 93,* 1165-1173.

Gino, F., Shang, J., & Croson, R. (2009). The impact of information from similar or different advisors on judgment. *Organizational Behavior and Human Decision Processes, 108,* 287-302.

Gollwitzer, P. M. (1999). Implementation intentions: Strong effects of simple plans. *American Psychologist, 54,* 493-503.

Gollwitzer, P. M., & Brandstätter, V. (1997). Implementation intentions and effective goal pursuit. *Journal of Personality and Social Psychology, 73,* 186-199.

Gomez-Beldarrain, M., Harries, C., Garcia-Monco, J. C., Ballus, E., & Grafman, J. (2004). Patients with right fontal lesions are unable to assess and use advice to make predictive judgments. *Journal of Cognitive Neuroscience, 16*, 74-89.

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes, 70,* 117-133.

Harvey, N. & Harries, C. (2004). Effects of judges' forecasting on their later combination of forecasts for the some outcomes. *International Journal of Forecasting, 20,* 391-409.

Harvey, N., Harries, C., & Fischer, I. (2000). Using advice and assessing its quality. *Organizational Behavior and Human Decision Processes, 81,* 252-273.

Hastie, R. (1986). Experimental evidence on group accuracy. In B. Grofman & G. Owen (Eds.), *Decision Research* (Vol. 2). Greewich, CT: JAI Press.

Hastie, R., & Dawes, R. M. (2001). *Rational choice in an uncertain world: The psychology of judgment and decision making.* Thousand Oaks, CA: Sage.

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science, 20,* 231-237.

Kadous, K., Krische, S., & Sedor, L. (2006). Using Counter-Explanation to Limit Analysts' Forecast Optimism. *The Accounting Review*, *81,* 377-397

Kahneman, D., & Tversky, A. (1972). Subjective Probability: A judgment of representativeness. *Cognitive Psychology, 3,* 430-454.

LeBreton, J. M., Ployheart, R. E., & Ladd, R. T. (2004). A Monte Carlo comparison of relative importance methodologies. *Organizational Research Methods, 7,* 258–282.

Lim, J.S., & O'Connor, M. (1995). Judgmental adjustment of initial forecasts: Its effectiveness and biases. *Journal of Behavioral Decision-Making, 8,* 149-168.

Mussweiler, T. (2001). "Seek and ye shall find": Antecedents of assimilation and contrast in social comparison. *European Journal of Social Psychology, 31,* 499–509.

Mussweiler, T. (2003). Comparison processes in social judgment: Mechanisms and consequences. *Psychological Review, 110,* 472-489.

Mussweiler, T., & Bodenhausen, G. (2002). I know you are but what am I? Self-evaluative consequences of judging in-group and out-group members. *Journal of Personality and Social Psychology, 82,* 19–32.

Mussweiler, T., Pfeiffer, T., & Strack, F. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality and Social Psychology Bulletin, 26, 1142-1150.*

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35,* 136-164.

Mussweiler, T., & Strack, F. (2001). The semantics of anchoring. *Organizational Behavior and Human Decision Processes, 86,* 234-255.

Nortchcraft, G. B., & Neale, M. A. (1987). Experts, amateurs, and real estate: An anchoring and adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes, 39,* 84-97.

Satterthwaite, F. E. (1946). An approximate distribution of estimates of variance components. *Biometrics Bulletin, 2,* 110–114.

Sanna, L.J., Schwarz, N., & Stocker, S.L. (2002). When debiasing backfires: Accessible content and accessibility experiences in debiasing hindsight. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28,* 497–502.

Schrah, G. E., Dalal, R. S., & Sniezek, J. A. (2006).No decision maker is an island: Integrating expert advice with information acquisition. *Journal of Behavioral Decision-Making, 19,* 43–60.

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2010). *Why Groups Perform Better than Individuals at Quantitative Judgment Tasks: The Effects of Group Learning and Expertise Feedback.* Manuscript submitted for publication.

Simon, D. P. (1989). The rationality of federal funds rate expectations: Evidence from a survey. *Journal of Money, Credit and Banking, 21,* 388-393.

Soll, J. B., & Larrick, R. (2009). Strategies for revising judgment: how (and how well) people use others' opinions. *Journal of Experimental Psychology: Learning, Memory and Cognition, 35,* 780-805.

Spiwoks, M., Bedke, N., Hein, O. (2008). Forecasting the past: The case of US interest rate forecasts. *Financial Marketing and Portfolio Management, 22,* 357-379.

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes, 62,* 159-174.

Sniezek, J.A., Schrah, G.E., & Dalal, R.S. (2004). Improving judgment with prepaid expert advice. *Journal of Behavioral Decision Making, 17,* 173-190.

Stasser, G., & Dietz-Uhler, B. (2001). Collective choice, judgment and problem solving. In M. A. Hogg & S. Tindale (Eds.), *Blackwell handbook of social psychology: Group processes* (pp. 31-55). Oxford, UK: Blackwell.

Stanovich, K. E. (1999). *Who is Rational? Studies of Individual Differences in Reasoning.* Mahwah, NJ: Erlbaum.

Steiner, I. D. (1966). Models for inferring relationships between groups size and group potential group productivity. *Behavioral Sciences, 11,* 273-283.

Stroop, J. R. (1932). Is the judgment of the group better than that of the average member of the group? *Journal of Experimental Psychology, 15,* 550-562.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185,* 1124–1131.

Wilson, T. D., Houston, C. E., Etling, K. M., & Brekke, N. (1996). A new look at anchoring effects: Basic anchoring and its antecedents. *Journal of Experimental Psychology: General, 125,* 387-402.

Winer, B. J., Brown, D. R., & Michels, K. M. (1991). *Statistical principles in experimental design.* New York, NY: McGraw-Hill.

Yaniv, I. (2004a). The benefit of additional opinions. *Current Directions in Psychological Science, 13,* 75-78.

Yaniv, I. (2004b). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes, 93,* 1-13.

Yaniv, I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes, 83,* 260-281.

Yaniv, I., & Milyawsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes, 103,* 104-210.

Lebenslauf

Dipl.-Psych. Dipl.-Betriebswirt (BA) Thomas Schultze

Georg-August-Universität Göttingen

Georg-Elias-Müller-Institut für Psychologie

Abteilung für Wirtschafts- und Sozialpsychologie

Goßlerstraße 14

37073 Göttingen

Persönliches

Geburtsdatum:	28. Februar, 1979
Geburtsort:	Darmstadt
Staatsangehörigkeit:	Deutsch

Beruflicher Werdegang

1998	Abitur, Franziskaner-Gymnasium Kreuzburg, Großkrotzenburg
1998-1999	Zivildienst, evangelische Kirchengemeinde am Limes, Hanau-Großauheim
1999 – 2002	Studium der Betriebswirtschaftslehre an der Berufsakademie Berlin
2002 – 2004	Grundstudium in Psychologie an der Martin-Luther-Universität Halle-Wittenberg
2004 – 2007	Hauptstudium in Psychologie an der Georg-August-Universität Göttingen
2005 – 2007	Stipendium der Studienstiftung des deutschen Volkes
Seit 2007	Promotion im Fach Psychologie an der Georg-August-Universität Göttingen