GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

# Erkennung und Korrektur von Urteilsfehlern bei eigenen Urteilen und beim Umgang mit Ratschlägen: Ein hierarchischer Sampling-Ansatz

Dissertation

zur Erlangung des mathematisch-naturwissenschaftlichen Doktorgrades

"Doctor rerum naturalium"

der Georg-August-Universität Göttingen

im Promotionsprogramm Biologie

der Georg-August University School of Science (GAUSS)

vorgelegt von

Jacob Christian Rittich

aus Hamburg

Göttingen, 2021

Betreuungsausschuss

PD Dr. Thomas Schultze-Gerlach, Abteilung für Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Michael R. Waldmann, Abteilung für Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Stefan Schulz-Hardt, Abteilung für Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Mitglieder der Prüfungskommission

Referent: PD Dr. Thomas Schultze-Gerlach, Abteilung für Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Korreferent: Prof. Dr. Michael R. Waldmann, Abteilung für Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Stefan Schulz-Hardt, Abteilung für Wirtschafts- und Sozialpsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Margarete Boos, Abteilung für Sozial- und Kommunikationspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. Sascha Schroeder, Abteilung für Pädagogische Psychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Prof. Dr. York Hagmayer, Abteilung für Kognitionswissenschaft und Entscheidungspsychologie, Georg-Elias-Müller-Institut für Psychologie, Georg-August-Universität Göttingen

Tag der mündlichen Prüfung: 15.12.2021

Für meine Mutter

Isabel Rittich

und meinen Vater

Dr. Peter Rittich

**Danksagung**

Mit meiner Promotion geht für mich eine lange Reise zu Ende, bei der ich sehr viel über mich und meine Rolle als Wissenschaftler und Psychologe gelernt habe. Sie hat mir gezeigt, wie häufig wir nur wenig gesichert über psychologische Prozesse und Phänomene wissen und wie wichtig es deshalb ist, mit sauberer theoretischer und methodischer Arbeit etwas mehr über diese Prozesse zu erfahren und die Ergebnisse am Ende auch allgemeinverständlich und transparent aufzuschreiben! Bei dieser Reise haben mich sehr viele wunderbare Menschen begleitet, von denen ich sehr viel über gute Theorienbildung und sauberes methodisches Arbeiten gelernt habe und die mich unterstützt haben, da sich ein solcher Forschungsprozess durchaus zeitweise als sehr frustrierend herausstellen kann.

Mein erster Dank gilt meinem Doktorvater, Mentor und Freund Dr. Thomas Schultze-Gerlach, der mir seit meinem Bachelorstudium bei allen meinen Forschungsvorhaben immer mit Rat und Tat zu Seite stand. Thomas, ohne meine Ausbildung bei Dir und Deine Unterstützung wäre ich sicherlich nicht in der Lage gewesen eine solche Arbeit zu meistern und es fällt mir schwer, die richtigen Worte zu finden. Als Doktorvater hast Du die richtige Mischung aus Gandalf und Samweis Gamdschie in Dir, was ein Herr Frodo gerade auf den letzten Metern vor dem Schicksalsberg sehr zu schätzen weiß (wobei ich mich am Ende doch eher leichten Herzens von meiner Doktorarbeit trennen werde).

Ein ganz besonderer Dank gilt natürlich auch Prof. Stefan Schulz-Hardt, der mir immer mit sehr wertvollem Feedback zur Seite stand und mich in der Umsetzung meiner Ideen bekräftigt und unterstützt hat. Gleiches gilt für Prof. Jack Soll und Prof. Richard Larrick, mit denen ich eng zusammenarbeiten durfte. Ich hatte das große Glück gleich mehrere Mentoren zu finden, von denen ich sehr viel über gute und methodisch saubere Arbeit in der Wissenschaft gelernt habe.

Ich möchte mich auch besonders bei Prof. Michael Waldmann bedanken, dessen Feedback als Kognitionspsychologe für meine Arbeit sehr wichtig war und der sich auch bereit erklärt hat das Zweitgutachten meiner Arbeit zu übernehmen. Bei Prof. York Hagmayer, Prof. Margarete Boos und Prof. Sascha Schroeder möchte ich mich zudem bedanken, dass sie sich bereit erklärt haben meine Prüfungskommission zu vervollständigen.

Insbesondere von den Kolleg:innen aus meiner Abteilung habe ich viel Unterstützung und Rückhalt erfahren. Vorneweg möchte ich mich bei meinem Freund Matthias Lippold bedanken, mit dem ich während meiner Promotion die meiste Zeit verbracht habe. Genauso möchte ich mich bei Dr. Christian Treffenstädt, Dr. Johanna Prüfer, Jana Reichhold und Johannes Rollwage bedanken, mit all denen ich eine sehr schöne Zeit als Doktorand verbinde. An dieser Stelle möchte ich mich auch bei Ronja Demel für ihre Unterstützung bedanken, obwohl sie zwei Stockwerke unter unserer Abteilung promoviert hat, aber glücklicherweise ihr Home-Office nur eine Straße weiter von meinem eingerichtet hat und gerade auf den letzten Metern vor der Abgabe eine ganz wichtige Unterstützung war.

In dieser Zeit habe ich noch viele andere Menschen kennengelernt, bei denen ich mich bedanken möchte. Durch die Zusammenarbeit an anderen Forschungsprojekten habe ich viel von Dr. Tanja Gerlach, Prof. Susann Fiedler, Prof. Andreas Mojzisch, und Dr. Markus Germar gelernt. Eine ganz wichtige Unterstützung waren zudem alle studentischen Hilfskräfte sowie Bachelor- und Masterabsolvent:innen, die mich bei meinen Forschungsprojekten (insbesondere bei der Datenerhebung) tatkräftig unterstützt haben, nämlich Birk Oelhoff, Natalie Bleijlevens, Peter Neumann, Ulrike Herms, Eva Zielke, Judith Diele, Paula Tönshoff , Franziska Teschner, Lea Rehbein, André Rinn, Tjark Rode, Jana Schildknecht, Tobias Peters, und Diana Festor.

Sehr wichtig ist es mir auch an dieser Stelle die Institutionen zu nennen, die mir überhaupt die Möglichkeit gegeben haben in dieser Form an meinen eigenen Projekten zu arbeiten und mich dafür weiterzubilden. Die finanzielle und ideelle Förderung der Studienstiftung des deutschen Volkes hat es mir erlaubt, mich vollkommen auf meine Ideen zu konzentrieren und dafür auch meine beiden Mentoren in den USA zu besuchen. Aber auch als Mitglied des Leibniz ScienceCampus ‘Primate Cognition’ und assoziiertes Mitglied des Graduiertenkolleg 2070 zu sozialen Beziehungen habe ich von den Kursangeboten, Konferenzförderungen und dem Austausch mit anderen Forscher:innen profitiert.

Abschließend möchte ich mich bei meinen Freunden und Freundinnen, meinen Eltern, meinen Geschwistern und meiner Nichte und meinen Neffen bedanken, die mein Leben so sehr bereichern. Ihr hättet alle mehr als eine eigene Seite Danksagung verdient. Aber das mache ich lieber persönlich.

**Inhaltsverzeichnis**

## 1. Einleitung

In unserem Leben müssen wir immer wieder Entscheidungen unter Unsicherheit treffen, die weitreichende Konsequenzen für uns und unsere Umwelt mit sich bringen. Auch vermeintlich kleinere Entscheidungen können in der Summe von großer gesellschaftlicher Tragweite sein. Dies lässt sich derzeit besonders deutlich anhand des Infektionsgeschehens der COVID-19-Pandemie beobachten: Das Tragen von Masken, Reduzierung von Kontakten, das Einhalten von Abstands- und Hygieneregeln und nicht zuletzt die Durchführung einer Schutzimpfung sind alles Entscheidungen von Einzelpersonen, die in der Summe erheblich zur Pandemiebekämpfung beitragen. Bei all diesen Entscheidungen ist es wichtig, dass wir diese auf Basis guter *Urteile* treffen, das heißt, dass die Sachverhalte, die diesen Entscheidungen zugrunde liegen, möglichst korrekt eingeschätzt werden.

Unter dem Begriff *Urteil* wird in dieser Arbeit eine subjektive Einschätzung verstanden, die sich meist zu dem jetzigen oder zu einem zukünftigen Zeitpunkt nach objektiven Standards hinsichtlich ihres Zutreffens beziehungsweise ihrer Genauigkeit bewerten lässt (Pachur & Bröder, 2013). Eine besonders wichtige Rolle nehmen die *quantitativen Urteile* bei der Entscheidungsfindung ein, die auch im Zentrum dieser Arbeit stehen. Bei quantitativen Urteilen erfolgt eine Einschätzung über einen Urteilsgegenstand auf einer kontinuierlichen Dimension. Die Relevanz quantitativer Urteile wird wiederum anhand der derzeitigen Pandemie deutlich, bei der zum Beispiel Urteile bezüglich benötigter Krankenhausbetten, Ansteckungsrisiken und wirtschaftlicher Kosten getroffen werden müssen. Bei unseren individuellen Urteilen handelt es sich meist um sehr alltägliche Urteile, wie das Abschätzen der benötigen Menge verschiedener Einkaufswaren beim Wocheneinkauf. Doch auch unsere individuellen Urteile können weitreichende Konsequenzen mit sich bringen. Ein aktuelles Beispiel sind die eingangs erwähnten Impfentscheidungen, die unter anderem individuell auf Basis der geschätzten Risiken aufgrund von Nebenwirkungen und den Risiken einer Infektion mit dem Virus getroffen werden. Leider kommt es bei solchen Urteilen häufig zu Fehlern, die sich teilweise nur als ärgerlich gestalten, wenn beispielsweise zu wenige oder zu viele Tomaten gekauft werden, aber schwerwiegende Konsequenzen haben, wenn beispielsweise bei Impfentscheidungen die Risiken einer Impfung massiv überschätzt und die Risiken einer Infektion massiv unterschätzt werden. Deshalb ist es wichtig, dass Urteilsfehler möglichst frühzeitig erkannt

und korrigiert werden. Ziel meiner Arbeit ist es zu untersuchen, inwiefern Menschen hierzu in der Lage sind und inwiefern Prozesse der Fehlerkorrektur mit Hilfe von gezielten Interventionen verbessert werden können.

Sowohl die Erkennung als auch die Korrektur von Fehlern sind wichtige Bestandteile der Urteils- und Entscheidungsforschung. Bisher werden diese jedoch weitestgehend losgelöst voneinander untersucht. Erkenntnisse zu der Erkennung von Fehlern kommen primär aus der Literatur zu *Metakognitionen* (Kognition über Kognitionen) bei der Urteilsbildung. In diesem Forschungsbereich wird untersucht, wie gut Menschen ihre eigenen Urteilsprozesse, beziehungsweise die Qualität ihrer Urteile, bewerten können. Diese Prozesse werden vornehmlich anhand der subjektiven Urteilssicherheit (*Confidence*) untersucht (z.B., Juslin et al., 2007; Klayman et al., 2006). Bei quantitativen Urteilen sind solche Sicherheitsurteile vergleichbar mit der subjektiven Einschätzung der *Größe* der eigenen Fehler (Yaniv & Foster, 1997) und werden teilweise auch direkt so operationalisiert oder lassen sich mit wenigen zusätzlichen Annahmen in solche übersetzen (Soll et al., 2021). Das Problem für Urteilskorrekturen bei quantitativen Urteilen ist jedoch, dass neben der *Größe* von Fehlern auch eine adäquate Einschätzung der *Richtung* von Fehlern notwendig ist, damit Menschen ihre Urteile auch in die entsprechend entgegengesetzte Richtung korrigieren können (Wilson & Brekke, 1994). So reicht es beispielsweise in der medizinischen Praxis nicht allein aus zu wissen, dass Patient:innen bei einer medikamentösen Behandlung falsch eingestellt sind. Das medizinische Personal muss zusätzlich wissen, ob die jeweilige Dosierung zu niedrig oder zu hoch war, um entsprechende Korrekturen vorzunehmen. Somit liefern die Arbeiten zur Urteilssicherheit zwar einige wichtige Erkenntnisse, beispielsweise, inwiefern Menschen sich bei ihren Urteilen sicherer sind als objektiv gerechtfertigt ist (*Overconfidence*) und wie Over- oder auch Underconfidence von Aufgaben- und Abfrageformat der Sicherheitsurteile abhängen (Juslin et al., 2007; Klayman et al., 2006; Koriat, 2012a). Es lassen sich hieraus jedoch keine direkten Implikationen für die Korrektur von Fehlern ableiten (zumindest nicht im individuellen Urteilskontext; vgl. Yaniv, 1997). Ein Ziel meiner Arbeit ist es daher zu untersuchen, inwiefern Menschen die *Richtung* ihrer Fehler erkennen können, beziehungsweise die Richtung, in der sie ihre eigenen Urteile korrigieren sollten. Eine solche Fähigkeit bezeichne ich als *Weisheitsbewusstsein*, welches später detaillierter theoretisch abgeleitet wird.

Neben der Forschung zu Metakognitionen gibt es mehrere Arbeiten, die sich damit beschäftigen, mit welchen Strategien Urteilsfehler am besten reduziert werden können. Diese Strategien machen sich zumeist ein statistisches Prinzip zunutze, das weitläufig als die *Weisheit der Vielen* (*Wisdom of Crowd*) bekannt geworden ist (Herzog et al., 2019; Larrick et al., 2012; Surowiecki, 2004). Die Weisheit der Vielen bezeichnet den Effekt, dass sich unsystematische Fehler über mehrere Urteile mit einfachen Integrationsstrategien ausmitteln lassen. Eine besonders beliebte Strategie bei quantitativen Urteilen ist die Bildung eines einfachen arithmetischen Mittels über mehrere Urteile. Schätzt beispielsweise eine Person mit einem Urteil eine Zielgröße zu niedrig ein, kann mit dem Mittelwert der Fehler durch eine andere Person korrigiert werden, die dieselbe Zielgröße zu hoch eingeschätzt hat. Es reicht sogar aus, nur zwei Urteile zu mitteln, um von dem Effekt zu profitieren (Soll & Larrick, 2009). Darüber hinaus können sich Einzelpersonen diesen Effekt teilweise sogar zunutze machen, indem sie selbst zwei unterschiedliche Urteile zum selben Sachverhalt abgeben und diese mitteln (z.B., Herzog & Hertwig, 2009, 2014b; Steegen et al., 2014; Stroop, 1932; Vul & Pashler, 2008; für einen Überblick Herzog & Hertwig, 2014a). Diese Variante des Effekts wird auch in der neueren Literatur als die *Weisheit der Vielen in einer Person* (*Wisdom of Many in One Mind*) bezeichnet (Herzog & Hertwig, 2009).

Ein großes Problem der Weisheit der Vielen ist, dass Menschen dieses Potential nur unzureichend nutzen und zwar sowohl dann, wenn sie Urteile von unterschiedlichen Personen integrieren (Fischer & Harvey, 1999; Soll & Larrick, 2009; Yaniv & Choshen-Hillel, 2012a), als auch dann, wenn sie zwei eigene Urteile integrieren (Fraundorf & Benjamin, 2014; Herzog & Hertwig, 2014a, 2014b; Müller-Trede, 2011). Eines der möglichen Probleme ist hierbei, dass die Wirksamkeit einer Mittelwertsbildung von Menschen nicht oder nur unzureichend intuitiv verstanden wird und Menschen sehr häufig davon ausgehen, dass der Mittelwert nur zu durchschnittlichen und nicht zu überdurchschnittlichen Ergebnissen führt (Larrick & Soll, 2006). Das zweite Ziel meiner Arbeit ist daher, auf Basis eines Weisheitsbewusstseins eine neue Intervention zu entwickeln und zu testen. Diese Intervention soll Menschen in die Lage versetzen, ihre Urteile *selbstständig* zu verbessern. Somit wären Menschen nicht auf eine extern implementierte Strategie wie die Mittelwertsbildung angewiesen, deren Wert nicht oder nur unzureichend erkannt wird und die aus diesem oder anderen Gründen erfahrungsgemäß nur unzureichend angewendet

wird. Zudem wirkt die Bildung eines Mittelwerts genau dann besonders gut, wenn Urteile *unterschiedlicher* Personen verwendet werden (Herzog & Hertwig, 2014a; Vul & Pashler, 2008), die nicht immer in ausreichender Zahl verfügbar sind. Da Menschen jedoch auch aktiv Ratschläge einholen und hiervon stark profitieren (Yaniv, 2004a), soll die von mir vorgeschlagene Intervention sowohl in individuellen als auch in sozialen Kontexten (hier: beim Umgang mit Ratschlägen) zum Einsatz kommen können.

### 1.1. Vorgehen

Zur Ableitung des Weisheitsbewusstseins und darauf basierender Interventionen habe ich ein neues Modell entwickelt: das *generalisierte Sampling-Modell* (GSM). Das GSM übernimmt weitestgehend etablierte Annahmen aus der Literatur zu Metakognitionen bei der Urteilsbildung und zur Weisheit der Vielen. Die zentrale Grundannahme dieses Ansatzes ist, dass Menschen als *intuitive Statistiker:innen* verstanden werden (Peterson & Beach, 1967), die ihre Urteile auf Basis (meist) zufällig gezogener Informationsstichproben treffen (Fiedler, 2000; Fiedler & Juslin, 2006b). Hierbei handelt es sich um eine einflussreiche Metapher der Urteils- und Entscheidungsforschung, welche in der einen oder anderen Form vielen (meta-)kognitiven Ansätzen zugrunde liegt (z.B., Juslin et al., 2007; Koriat, 2012a; für einen Überblick siehe Fiedler & Juslin, 2006a), die kurz als *Sampling*-Modelle bezeichnet werden können (Fiedler & Juslin, 2006b). In diesen Arbeiten wird angenommen, dass Menschen zur Urteils- und Entscheidungsfindung Stichproben von fehlerbehafteten Informationen ziehen und diese „Daten“ (wie Statistiker:innen) angemessen zusammenfassen können. Als Beispiel könnte eine fiktive Person aus Schleswig-Holstein dienen, deren Aufgabe es ist, die Bevölkerungszahl der Stadt Celle in Niedersachsen zu schätzen, deren wahrer Wert ihr nicht bekannt ist. Die Person könnte hierfür ihr bekannte Zahlen von ähnlichen Städten aus Schleswig-Holstein heranziehen, wie Elmshorn (ca. 50.000), Norderstedt (ca. 80.000) und Neumünster (ca. 80.000)[1]. Die angemessene statische Zusammenfassung wäre in diesem Fall die Zahlen zu mitteln, zumindest sofern keine weiteren Informationen über die Ähnlichkeit der Städte zu Celle bekannt sind (vgl. Juslin et al., 2003). Für die Zusammenfassung der Informationsstichprobe wird damit eine durchaus

[1] Für alle kommenden Städte-Beispiele wurden die „Liste der größten Städte in Schleswig-Holstein“ (2021), die „Liste der größten Städte in Niedersachsen“ (2021) und die „Liste der größten Städte in Bayern“ (2021) aus Wikipedia herangezogen.

angemessene Integrationsstrategien eingesetzt, die, wie eingangs erwähnt, auch zur Nutzung der Weisheit der Vielen empfohlen wird (Herzog et al., 2019). Die Person würde also 70.000 als Bevölkerungszahl für Celle angeben, was auch ungefähr der tatsächlichen Bevölkerungszahl von Celle entspricht (der wahre Wert ist 69.399). Mit einer solchen Strategie können jedoch auch Fehler entstehen, wenn beispielsweise anstatt Neumünster Mölln (ca. 20.000) zusammen mit den anderen beiden Städten gemittelt wird, was in einer Gesamtschätzung von 50.000 anstatt 70.000 resultieren würde.

Der wesentliche Fokus der Sampling-Ansätze liegt darauf, wie die Auswahl von Informationen bei Urteilen zu Fehlern führt (z.B. Mölln statt Neumünster). Die Auswahl dieser Informationen wird als stochastischer Prozess modelliert (das *Sampling*). Je nach Sampling-Ansatz und Fragestellung gibt es durchaus Unterschiede, was die Informationen beinhalten, die gezogen werden. So können Menschen, wie in dem Beispiel, andere ähnliche Urteilsgegenstände aus der gleichen Domäne heranziehen, sogenannte *Exemplare* (z.B. Juslin et al., 2007). Andere Sampling-Ansätze verwenden einen breiteren Informationsbegriff, der unterschiedliche Informationen, wie Faktenwissen, visuelle Informationen (z.B. ein Stadtbild von Celle), und andere Überlegungen („Ich habe noch nie von Celle gehört, also muss es sehr klein sein“) umfasst (z.B. Koriat, 2012a). Solche Informationen können auch als *Suburteile* verstanden werden (vgl. Koriat, 2012a), die sich als eigenes (fehlerbehaftetes) Urteil ausdrücken lassen und damit letztlich genauso modelliert werden können wie die Exemplare. So würde die Person aus Schleswig-Holstein allein auf Basis der Information „Celle hat ein Stadtschloss“ ein Urteil von „60.000“ treffen, welches als Suburteil mit anderen numerischen Informationen (bzw. anderen Suburteilen) zu einem Gesamturteil integriert werden kann. Auch ich nehme in meinem Ansatz später an, dass sehr unterschiedliche Formen von Informationen von Menschen zur Urteilsbildung herangezogen werden können. Im Sinne eines einfachen Verständnisses werde ich jedoch weiterhin Exemplare als Informationsbeispiele verwenden, die von einem breiteren Informationsbegriff natürlich miteingeschlossen werden. Der Vorteil der Exemplare liegt darin, dass diese nicht erst in Suburteile übersetzt werden müssen, was die Darstellung abstrakter Sampling-Prozesse vereinfacht. Nach dem GSM können jedoch, wie erwähnt, durchaus sehr unterschiedlichen Informationen und Überlegungen zur Urteilsbildung herangezogen werden (z.B. „Celle hat keinen Fußballverein in der ersten Bundesliga“:

Suburteil 40.000, „Celle hat einen eigenen Bahnhof“: Suburteil 110.000, aber auch „Celle könnte so groß sein wie Neumünster“: Suburteil 80.000). Für diese unterschiedlichen Formen von Informationen gelten die später getroffenen Annahmen des GSM und abgeleiteten Vorhersagen genauso wie für die Exemplare in den Beispielen.

Eine wesentliche Annahme der neueren Sampling-Ansätze ist, dass Menschen (wie Statistiker:innen) zwar ihre Informationsstichproben gut beschreiben oder zusammenfassen, aber nicht dafür kontrollieren können, dass die Auswahl dieser Informationen gegebenenfalls verzerrt ist. Diese Annahme wird auch als *Naivitätsannahme* bezeichnet, weshalb Menschen in Sampling-Ansätzen auch als *naive* intuitive Statistiker:innen verstanden werden (Fiedler & Juslin, 2006b; Juslin et al., 2007; siehe auch Fiedler, 2000). Naiv heißt in diesem Kontext, dass Menschen nicht die Ursprünge ihrer Informationen reflektieren und damit nicht für die damit einhergehenden Limitationen kontrollieren können. Sie können beispielsweise nicht dafür kontrollieren, dass bestimmte Formen oder Strategien der Informationssuche mit systematischen Fehlern einhergehen. So könnte beispielsweise eine andere Person aus Bayern häufig verschiedene Freund:innen in kleineren Städten in Oberfranken besuchen. Als bestmögliche Schätzung der Bevölkerungszahl von Celle würde diese Person die Bevölkerungszahlen von kleinen Städten aus dieser Region heranziehen, zum Beispiel Coburg (40.000), Forchheim (30.000) und Lichtenfels (20.000). Auch diese Person kann ihre Informationsstichprobe angemessen beschreiben und alle Informationen mit einem Mittelwert integrieren (30.000). Sie kann jedoch nach der Naivitätsannahme nicht für ihre Präferenz kontrollieren, kleine Städte zu besuchen. Ihr ist nicht bewusst, was diese Präferenz für einen Einfluss auf ihre Urteile hat und kann ihre Schätzung damit auch nicht nach oben korrigieren. Diese Person würde damit die Bevölkerung von Celle unterschätzen.

Die Effekte der Weisheit der Vielen lassen sich anhand von Sampling-Modellen damit erklären, dass unterschiedlichen Urteilen unterschiedliche Informationen zugrunde liegen und sich durch die Bildung eines Mittelwerts die unsystematischen Fehler der unterschiedlichen Informationen ausmitteln können (z.B. Hourihan & Benjamin, 2010). Mit Hilfe der Naivitätsannahme kann in einem erweiterten Verständnis zudem erklärt werden, warum Menschen dieses Potential häufig nicht spontan nutzen: Sie können nicht abstrahieren, dass den unterschiedlichen Urteilen unterschiedliche Informationen zugrunde

liegen. So könnte es beispielsweise sein, dass die Person aus Bayern lernt, dass die Person aus Schleswig-Holstein 70.000 geschätzt hat. Der naiven Person aus Bayern ist jedoch nicht bewusst, dass die Person aus Schleswig-Holstein andere Informationen für ihre Schätzung verwendet hat, weshalb sie die Abweichung als Fehler und nicht als Resultat anderer (neuer) Informationen sieht und somit keinen Mittelwert bildet (Minson et al., 2011; siehe auch, Hütter & Ache, 2016; Yaniv, 2004b; Yaniv & Kleinberger, 2000).

Mit meinem eigenen Ansatz (dem GSM) übernehme ich die wesentlichen Annahmen der bisherigen Sampling-Ansätze inklusive der Naivitätsannahme. Eine zentrale Annahme des GSM ist jedoch, dass Menschen Verbesserungspotentiale in ihren Urteilen teilweise selbst erkennen können, indem sie die Sampling-Prozesse bewusst *neu* auslösen und damit teilweise neue Informationen erhalten, die sie mit den bereits getroffenen Urteilen abgleichen können. Beispielsweise könnte die Person aus Bayern im nächsten Schritt Bamberg mit ca. 75.000 Einwohner:innen abrufen und würde auf Basis dieser Information erkennen, dass sie vermutlich den wahren Wert von Celle unterschätzt hat. Somit lässt sich das Weisheitsbewusstsein, das zentrale Phänomen dieser Arbeit, ableiten. Konkret ist mit Weisheitsbewusstsein gemeint, dass Menschen häufig in der Lage sind zu erkennen, dass der wahre Wert eher unter ihrem eigenen Urteil liegt, wenn sie ein zu hohes quantitatives Urteil abgegebenen haben, oder, dass der wahre Wert eher über ihrem Urteil liegt, wenn sie ein zu niedriges Urteil abgegeben haben. Mit anderen Worten: Sie können die Richtung ihres Fehlers überzufällig gut erkennen.

Im Folgenden werde ich in Abschnitt 2 das GSM ausführlicher vorstellen und hieraus das zentrale Phänomen dieser Arbeit, das Weisheitsbewusstsein, ableiten. Nach einer kurzen Beschreibung des allgemeinen empirischen Vorgehens zur Untersuchung meiner Fragestellungen (Abschnitt 3) wird in Abschnitt 4 empirisch getestet, inwiefern Menschen tatsächlich über eine solche Fähigkeit verfügen. Ziel von Abschnitt 5 und 6 ist es jeweils zu testen, ob Menschen mit Hilfe von einfachen Interventionen in der Lage sind, mit ihrem Weisheitsbewusstsein ihre eigenen Urteile zu verbessern und auch Ratschläge von anderen Personen besser zu nutzen. In Abschnitt 7 werde ich schließlich im Rahmen einer Metaanalyse meiner eigenen Studien einen umfassenden Test des GSM liefern, um auf Basis der Metaanalyse eine abschließende Bewertung des Rahmenmodells vorzunehmen. In Abschnitt 8 werde ich mit einer zusammenfassenden Diskussion und einem Fazit schließen.

## 2. Theoretisches Rahmenmodell

Für eine ausführliche Vorstellung meines Sampling-Ansatzes zur Beschreibung der Fehlererkennung und -korrektur ist es zunächst entscheidend den Urteilsbegriff eindeutig zu definieren, der diesem Ansatz zugrunde liegt. Eine solche Definition habe ich bereits in kurzer Form in der Einleitung vorgenommen, beziehungsweise von Pachur und Bröder (2013) übernommen. Ich werde im Folgenden diese Definition noch einmal präzisieren. Der Grund hierfür liegt darin, dass diese Arbeit sowohl die sozialpsychologische als auch die kognitionspsychologische Urteils- und Entscheidungsforschung adressiert, in denen der Urteilsbegriff teilweise unterschiedlich aufgefasst wird. Im Folgenden werde ich daher zunächst auf die unterschiedlichen Definitionen eingehen und eine Begriffsklärung für diese Arbeit vornehmen, bevor ich die Annahmen bisheriger Ansätze zur Urteilsbildung detaillierter beschreiben werde und schließlich meinen eigenen Ansatz im Detail vorstelle.

### 2.1. Klärung des Urteilsbegriffs

Der allgemeine Begriff *Urteil* (*Judgment*) wird, wie erwähnt, in der Literatur teilweise unterschiedlich definiert, was vermutlich darin begründet ist, dass es sich bei der Urteils- und Entscheidungsforschung um ein interdisziplinäres Forschungsfeld handelt. So ist beispielsweise in der sozialpsychologischen Forschung zu Gruppen ein wichtiges definitorisches Kriterium, dass Urteile auf einer kontinuierlichen Dimension erfolgen (Stasser & Dietz-Uhler, 2001). Urteilsfehler lassen sich hierbei relativ einfach über die (absolute) Abweichung vom entsprechenden wahren Wert bemessen, zumindest sofern dieser bekannt ist (bspw. um wie viele Jahre das Alter einer Person über- oder unterschätzt wurde). Der Begriff *Urteilen* kann nach dieser Definition mit *quantitativem Urteilen* gleichgesetzt werden und von *Entscheidungen* abgegrenzt werden, bei denen eine Wahl aus diskreten Antwortoptionen erfolgt. Eine solche enge Definition von Urteilen ist in der Gruppenforschung häufig insofern sinnvoll, da unterschiedliche Antwortformate oder Skalenniveaus unterschiedliche Lösungsstrategien von Gruppen erfordern. So kann beispielsweise ein Mittelwert nur über kontinuierliche Urteile der einzelnen Gruppenmitglieder gebildet werden, nicht jedoch über diskrete Antwortoptionen (z.B. „Welche Stadt hat mehr Brücken? Hamburg oder Venedig?“).

In der kognitionspsychologischen Urteilsforschung wird der Urteilsbegriff teilweise weiter gefasst (siehe Pachur & Bröder, 2013). Nach einer solchen Definition kann ein Urteil auch eine Wahl aus mehreren diskreten Optionen umfassen (A, B oder C), und es stellt sich eher die Frage, inwiefern sich die *kognitiven Prozesse* zwischen den unterschiedlichen Formen von Urteilen (z.B. quantitativ oder dichotom) unterscheiden oder ähneln. Zentral ist in diesen Ansätzen zudem, dass es eine objektiv erfassbare richtige Lösung gibt, nach der die Qualität von Urteilen bemessen werden kann. Teilweise werden hierfür auch normative Modelle rationaler Urteilsfindung herangezogen. Der Urteilsbegriff wird in dieser Hinsicht von *Präferenzen* abgegrenzt, die nicht objektiv bewertet werden können. Nach dieser Definition lässt sich der Urteilsbegriff jedoch weniger klar vom Begriff der *Entscheidungen* abgrenzen, die für diese Arbeit als eine spezielle Subgruppe von Urteilen mit diskreten Antwortoptionen betrachtet werden können.

In meiner Arbeit gehe ich einerseits auf konkrete Strategien der Urteilsbildung ein, die sich nur auf quantitative Urteile beziehen (z.B. den Mittelwert). Auf der anderen Seite setzte ich mich in der Ableitung meiner eigenen Intervention insbesondere mit den kognitiven Prozessen der Urteilsbildung auseinander und beziehe mich hierbei vornehmlich auf die kognitionspsychologische Literatur. Ich übernehme daher den inklusiveren Urteilsbegriff von Pachur und Bröder (2013), wobei ich, sofern es notwendig ist, auch immer die konkrete Form des Urteils mit ihrem jeweiligen Skalenniveau betone. Zentral sind in dieser Arbeit insbesondere die *quantitativen Urteile*, bei denen die Einschätzungen auf einer kontinuierlichen Dimension erfolgen, und später auch *Richtungsurteile*, mit denen ich das Weisheitsbewusstsein erfasse („Liegt der wahre Wert eher über oder unter meinem ersten Urteil? Oder weder über noch unter dem ersten Urteil?“).

### 2.2. Bisherige Ansätze

In der Literatur wurden verschiedene Ansätze postuliert, um Urteilsprozesse und die damit einhergehenden Fehler zu beschreiben (für einen Überblick siehe Pachur & Bröder, 2013). Trotz größerer Unterschiede in den konkreten Prozessannahmen ist allen diesen Ansätzen gemeinsam, dass der Urteilsprozess grundsätzlich aus (1) einer Suche/Abruf von Informationen besteht, und (2) eine Integration dieser Informationen zu einem Urteil umfasst. Fehler entstehen bei Urteilen zum einen, wenn grundsätzlich Informationen nicht

oder nur in unzureichendem Maße verfügbar sind und zum anderen, wenn bei der Urteilsbildung der Abruf und/oder die Integration der Informationen fehlerhaft durchgeführt wird (Stewart, 2001). Dies kann beispielswiese durch einen selektiven Abruf oder das selektive Verwenden bestimmter Informationen bei der Integration geschehen (Mussweiler & Strack, 1999).

Auch die Sampling-Modelle machen diese basalen Annahmen, um Urteilsprozesse zu beschreiben (Fiedler & Juslin, 2006b). Wie eingangs erwähnt, bleibt dabei häufig im Hintergrund, um was für Informationen es sich inhaltlich handelt, sondern es wird eher darauf fokussiert, welche Informationen gezogen werden beziehungsweise nach welchen Kriterien die Auswahl erfolgt (z.B. eine zufällige Auswahl oder eine Auswahl von Informationen mit kleinen Werten wie bei der Person aus Bayern). Sofern die Informationsverteilung, aus der gezogen wird, nicht systematisch verzerrt ist (es gibt nur Informationen, nach denen sich die Bevölkerungszahl von Celle unterschätzen lässt), sollten sich bei rationalen Entscheidungsträger:innen und zufälliger Ziehung im Schnitt keine systematische Verzerrungen der Urteile ergeben. Doch zumindest bei einem Teil der Sampling-Erklärungen wird angenommen, dass systematische Urteilsfehler entstehen, weil die Suche (*Selection Bias*) oder die Integration von Informationen (*Response Bias*) verzerrt ist (siehe z.B. Klayman et al., 2006). So wird beispielsweise zur Erklärung des Eingangs erwähnten Overconfidence-Effekts bei quantitativen Urteilen angenommen, dass Menschen zu einem Selection oder Response Bias neigen und primär nach Informationen suchen, die ihr quantitatives Urteil stützen oder solche Informationen besonders stark gewichten (Klayman et al., 2006; Soll & Klayman, 2004). Soll beispielsweise die Person aus Bayern aus dem vorherigen Beispiel einschätzen, wie sicher sie sich bei ihrem (zu niedrigen) Urteil ist, könnte sie entweder systematisch weitere kleine Städte in Bayern heranziehen anstatt größere (Selection Bias) und sich dadurch besonders sicher bei ihrem Urteil sein. Sie könnte jedoch auch sowohl größere als auch kleinere Städte heranziehen, aber die kleineren stärker berücksichtigen, weil sie ihrer anfänglichen Vorstellung von Celle eher entsprechen (Response Bias).

Im Folgenden werde ich mit dem GSM einen eigenen Ansatz vorstellen, der die basalen Annahmen der bisherigen Sampling-Ansätze übernimmt. Im Sinne der Sparsamkeit werde ich mit dem GSM weitestgehend auf zusätzliche Annahmen bezüglich Response und

Selection Biases verzichten und nehme an, dass alle Informationen zufällig gezogen werden und im Rahmen der möglichen Verarbeitungskapazitäten rational verarbeitet werden. Ich werde jedoch später in der Diskussion in Abschnitt 8 auf mögliche Einschränkungen dieses Idealmodells zur Erklärung anderer Urteilsprozesse eingehen. Ich werde auch später Interventionen ableiten, bei denen Menschen instruiert werden Informationen unverzerrt zu suchen (weder eher in die eine noch in die andere Richtung) und die damit möglichst robust gegenüber einer Verletzung dieser vereinfachenden Annahmen sind. Eine besondere Rolle nimmt zudem die Naivitätsannahme ein (Fiedler & Juslin, 2006b), die in gewisser Hinsicht auch als genereller Response Bias verstanden werden kann, obwohl Menschen zumindest ihre eigene Informationsstichprobe nach dieser Annahme angemessen zusammenfassen können (für eine Erklärung von Overconfidence mit der Naivitätsannahme siehe z.B. Juslin et al., 2007). Mit dem GSM übernehme ich die Naivitätsannahme und gehe davon aus, dass Menschen naiv gegenüber den Ursprüngen ihrer eigenen Informationen sind. Wie bereits erwähnt, gehe ich jedoch mit dem Weisheitsbewusstsein davon aus, dass Menschen durch das Sampling teilweise neuer Informationen grundsätzlich in der Lage sind, die Defizite ihrer vorherigen Urteile zumindest teilweise zu erkennen.

### 2.3. Das generalisierte Sampling-Modell (GSM)

Mit dem GSM übernehme ich weitestgehend die oben erwähnten Annahmen und betrachte die Menschen als naive intuitive Statistiker:innen (Fiedler & Juslin, 2006b), die Informationen zufällig ziehen und auf Basis statistischer Prinzipien verarbeiten. Hierbei kann es sich um sehr unterschiedliche Formen von Informationen handeln, wobei ich zur Formalisierung des Urteilsprozesses annehme, dass sich diese Informationen mit numerischen Werten beziehungsweise numerischen Suburteilen abbilden lassen. Wie eingangs erwähnt, könnte beispielsweise eine Person allein auf Basis der Information „Celle hat ein Stadtschloss“ ein (Sub-)Urteil von 60.000 abgeben. Neuerungen des GSM liegen darin, dass ich mehrere generalisierende Annahmen mache, die es mir erlauben, die Entstehung, sowie die Erkennung und Korrektur von Urteilsfehlern in einen Zusammenhang zu stellen. Die wesentliche neue Annahme des Modells ist, dass Menschen Teile dieser Sampling-Prozesse bewusst mehrfach auslösen können, und zwar unabhängig von der konkreten Form des Urteils. Das heißt, dass Menschen die Sampling-Prozesse auslösen können, um ein quantitatives Urteil zu treffen („Was ist die Bevölkerungszahl von Celle?“),

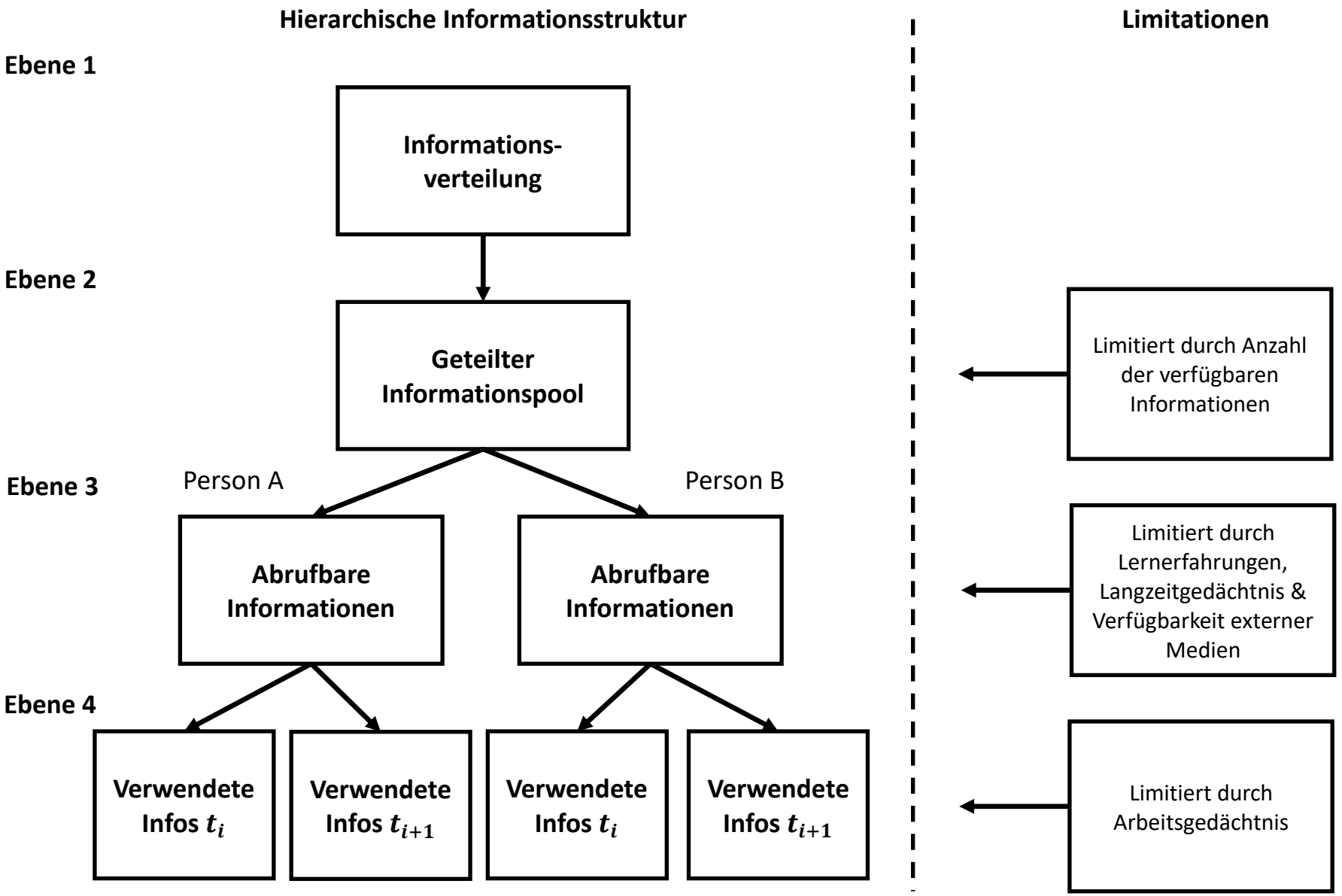
und auch, um diese später im Rahmen von Richtungsurteilen zu bewerten („Liegt der wahre Wert eher über oder eher unter dem ersten Urteil?“). Das Modell soll sich mit dieser Annahme später auch auf andere Urteilsprozesse anwenden lassen, weshalb ich das Modell als *generalisiertes* Sampling-Modell bezeichne. In dieser Arbeit fokussiere ich mich jedoch auf die Erkennung und Korrektur von Urteilsfehlern bei quantitativen Urteilen und werde auf andere Anwendungsformen des Modells erst in der Diskussion in Abschnitt 8 etwas detaillierter eingehen. In gewisser Hinsicht handelt es sich bei der Erkennung der Richtung von Fehlern um den ersten neuen Anwendungsfall des Modells, anhand dessen sich wesentliche Modellannahmen des GSM testen lassen. Zunächst werde ich jedoch in Abschnitt 2.3.1. auf allgemeine Annahmen des Modells eingehen, bevor ich in Abschnitt 2.3.2. auf konkrete Implikationen für die Fehlererkennung und -korrektur eingehen werde.

### 2.3.1. Hierarchische Struktur des GSM

Das GSM beschreibt Fälle, bei denen Menschen Sachverhalte mit einem wahren Wert auf Basis von Informationen einschätzen, die aus einer Informationsverteilung gezogen werden. Je mehr Informationen aus der Verteilung gezogen werden, desto besser lässt sich der wahre Wert über alle Informationen schätzen (zum Beispiel durch die Bildung eines Mittelwerts). Bei unbegrenzter abrufbarer Informationsmenge sollte dieser (theoretisch) eindeutig bestimmbar sein. Wären der Person aus Bayern aus dem Einführungsbeispiel alle relevanten Informationen über Celle bekannt, könnte sie die Bevölkerungszahl perfekt einschätzen. In der Realität sind jedoch zu jedem Zeitpunkt immer nur begrenzte Teilmengen dieser Informationen verfügbar, um Urteile zu treffen. Dies wird mit dem GSM über einen mehrschrittigen hierarchischen Sampling-Prozess formalisiert, bei dem immer nur eine Teilmenge an Informationen an die nächste Ebene weitergegeben wird (siehe Abbildung 1). Hierbei mache ich die vereinfachenden Annahmen, dass die Informationen auf allen Ebenen zufällig gezogen werden und am Ende durch einen gleichgewichteten Mittelwert integriert werden. Bei einer solchen hierarchischen Struktur handelt es sich wiederum um eine etablierte Konzeptualisierung, die verschiedenen Sampling-Ansätzen gemeinsam ist (z.B., Hourihan & Benjamin, 2010; Juslin et al., 2007; Koriat, 2012a). Bei anderen Ansätzen werden nicht alle Ebenen des GSM immer so explizit benannt und modelliert, aber meist zumindest implizit angenommen. So wird beispielsweise bei den metakognitiven Ansätzen betont, dass

**Abbildung 1**

*Das generalisierte Sampling-Modell (GSM)*



Menschen Informationen aus einem *geteilten* Informationspool ziehen (z.B. Koriat, 2012a), während insbesondere bei Ansätzen zur Weisheit der Vielen in einer Person betont wird, dass die *gleiche Person* ihre Informationen zufällig aus einem *privaten* Informationspool zieht (z.B. Hourihan & Benjamin, 2010). Im GSM werden, wie erwähnt, alle diese Ebenen explizit betont und modelliert. Nach der Naivitätsannahme sollten sich Menschen dieser Struktur jedoch nicht bewusst sein. Mit anderen Worten verstehen Menschen in der Regel nicht, welchen Einfluss diese Struktur auf ihre Urteilsprozesse hat. Sie haben nur Einblick in die Informationen, die ihnen nach dem hierarchischen Sampling-Prozess am Ende zur Verfügung stehen.

Die hierarchische Struktur beginnt mit der obersten Ebene, auf der sich die Informationsverteilung befindet, aus der letztlich alle Informationen gezogen werden. Auf der zweiten Ebene des GSM befindet sich der geteilte Pool von Informationen, auf die eine Population von Entscheidungsträger:innen bei der Bearbeitung einer Urteilsaufgabe auch grundsätzlich Zugriff hat (z.B. Koriat 2012a). Mit der zweiten Ebene wird modelliert, dass für verschiedene Urteilsgegenstände Wissen nur in limitierter Form in der Welt bewusst vorhanden ist, beziehungsweise nur eine begrenzte Menge von Informationen von Menschen entdeckt und geteilt wurde. Diese Ebene spiegelt damit die erste Begrenzung der Menge von Informationen wider, durch die der wahre Wert nicht immer eindeutig bestimmbar wird. Ein gutes Beispiel hierfür sind Prognoseaufgaben, wie die Vorhersage der Erderwärmung in den nächsten 10 Jahren. Eine allwissende Entität könnte die Erderwärmung perfekt vorhersagen, während wir uns in der Realität auf eine begrenzte Menge von Informationen von Wissenschaftler:innen verlassen müssen, die nur unter Annahme gewisser Szenarien Vorhersagen machen können (z.B. mit oder ohne Kohleausstieg).

Als nächste Begrenzung nehme ich mit der Ebene 3 an, dass eine einzelne Person nur auf einen bestimmten Anteil dieser Informationen zugreifen kann. Mit anderen Worten verfügt also jede Person nur über einen Teilausschnitt des gesamten verfügbaren Wissens. Genau wie bei anderen hierarchischen Ansätzen (z.B., Juslin et al., 2007) setze ich diese Ebene meist mit dem Langzeitgedächtnis gleich, das heißt, dass sich auf Ebene 3 nur die Informationen wiederfinden, die durch Lernerfahrungen auch Teil des Wissens einer Person geworden sind und zur Urteilsbildung auch von dieser Person grundsätzlich abgerufen

werden können. In der Realität kann es sich jedoch auch um externe Medien handeln (Bücher, Websites), zu denen nur manche Menschen unmittelbaren Zugang haben (siehe auch Fiedler, 2000). Die unterste Ebene (Ebene 4) spiegelt schließlich die Begrenzung des Arbeitsgedächtnisses wider, die auch in anderen Sampling-Ansätzen explizit betont wird (z.B., Hourihan & Benjamin, 2010; Juslin et al., 2007). Ich nehme mit dem GSM an, dass eine Person durch die Begrenzung ihres Arbeitsgedächtnisses zu einem bestimmten Zeitpunkt ($t_i$) nur eine begrenzte Menge von Informationen abrufen und verarbeiten kann (Cowan, 2001; Miller, 1956). Auf dieser Ebene findet schließlich über die Integration der noch vorhandenen Informationen die tatsächliche Urteilsbildung statt. Angewendet auf das Einführungsbeispiel könnte eine Person die Bevölkerungszahl von drei oder vier bekannten Städten aus dem Langezeitgedächtnis abrufen und mitteln.

Nach dem GSM können Urteilsfehler durch Informationsverluste auf drei unterschiedlichen Ebenen entstehen: weil die notwendigen Erkenntnisse noch nicht vorhanden sind (Ebene 2), nicht unmittelbar zugänglich sind (Ebene 3), oder nicht zeitgleich abgerufen und verarbeitet werden können (Ebene 4). Mit dem GSM lässt sich jedoch auch anhand der Weisheit der Vielen gut erklären, inwiefern Menschen ihre Urteilsfehler (theoretisch) korrigieren können. Geben zwei Personen Urteile zu dem gleichen Sachverhalt ab, würden diese zweimal den Sampling-Prozess auslösen und (teilweise) unterschiedliche Informationen aus dem geteilten Informationspool auf Ebene 2 ziehen. Durch die Auswahl der Informationen entstehen bei den beiden Urteilen teilweise unterschiedliche Fehler, die sie gegenseitig durch die Bildung eines Mittelwerts über beide Urteile korrigieren können. Ob und inwiefern hierdurch die Akkuratheit der Urteile gesteigert werden kann, lässt sich bei zwei Urteilen anhand der sogenannten *Bracketing Rate* (deutsch „einklammern“) erklären (Larrick & Soll, 2006). Diese quantifiziert, inwiefern zwei Urteile voneinander unabhängige Informationen über den wahren Wert liefern können und sich die Weisheit der Vielen durch einen Mittelwert entfalten kann (bzw. zu einem akkurateren Gesamturteil führt). Die Bracketing Rate gibt an, mit welcher Wahrscheinlichkeit zwei Urteile auf unterschiedlichen Seiten des wahren Werts landen. Wird beispielsweise das Alter einer 30-jährigen Person geschätzt, liegt ein Bracketing vor, wenn es mit einem Urteil unterschätzt (z.B. 29) und einem überschätzt wird (z.B. 33). In diesen Fällen ist der Mittelwert (31) mit einem absoluten Fehler von einem Jahr akkurater als der absolute Fehler der Einzelurteile im Durchschnitt (2

Jahre). Liegt kein Bracketing vor, so entspricht die Akkuratheit des Mittelwerts genau der Akkuratheit der beiden Einzelurteile im Durchschnitt, sodass der Mittelwert keinen Vorteil, aber auch keinen systematischen Nachteil gegenüber einer zufälligen Wahl eines der Einzelurteile hat. Wird die gleiche Person beispielsweise auf 32 und 34 geschätzt, so liegen der absolute Fehler des Mittelwerts (33) und der durchschnittliche absolute Fehler der Einzelurteile bei jeweils drei Jahren. Somit kann der Mittelwert im Schnitt nicht schlechter sein als eine zufällige Wahl der Urteile, jedoch häufig besser, sofern eine positive Bracketing Rate vorliegt. Folglich gibt die Bracketing Rate an, wie *häufig* zwei Menschen gegenüber einer zufälligen Wahl von der Weisheit der Vielen profitieren können. Sofern zwei Personen immer unterschiedliche Informationen aus der Informationsverteilung ziehen und die Verteilung selbst nicht systematisch verzerrt ist, sollte sich eine Bracketing Rate von 50% ergeben. Die unterschiedlichen Informationen würden mit 50% Wahrscheinlichkeit auf unterschiedlichen Seiten des wahren Werts landen und damit auch die Urteile, die hierauf basieren.

Mit dem GSM lässt sich auch erklären, warum die Weisheit der Vielen nicht immer funktioniert. Mit der Weisheit der Vielen lässt sich beispielsweise nicht der Informationsverlust kompensieren, der auf Ebene 2 entsteht, da über Ebene 2 hinaus von unterschiedlichen Personen keine zusätzlichen Informationen aus der Verteilung auf Ebene 1 gezogen werden können. Für das Städtebeispiel aus der Einleitung bedeutet dies, dass es nicht unendlich viele Städte in Deutschland gibt, die unterschiedliche Personen heranziehen können, um ihre Urteile zu treffen. Im schlimmsten Fall enthalten zwei Urteile dadurch genau die gleichen Informationen (die gleichen Städte), und durch die Bildung eines Mittelwerts kann kein Akkuratheitsgewinn erzielt werden. In der Realität ist es vermutlich meist der Fall, dass zwei Urteile von unterschiedlichen Personen anteilig gleiche und unterschiedliche Informationen enthalten, wodurch Bracketing Rates meist deutlich über 0% jedoch unter 50% liegen. So lagen die empirischen Bracketing Rates bei zwei Urteilen unterschiedlicher Personen bisher in entsprechenden Studien im Bereich von 26% und 41% (Soll & Larrick, 2009; Soll et al., 2021).

Mit dem GSM lässt sich auch veranschaulichen, inwiefern Menschen ihre Urteile (theoretisch) selbst korrigieren können. Dies lässt sich analog zur „klassischen“ Weisheit der Vielen anhand der Weisheit der Vielen in einer Person erklären, bei der die Bildung eines

Mittelwerts über unterschiedliche Urteile derselben Person erfolgt. Die gleiche Person kann mehrfach auf einen Pool von privaten Informationen zugreifen (zu $t_1$ und $t_2$), der Informationen aus dem geteilten Pool enthält (siehe Abbildung 1). Da der private Informationspool auf Ebene 3 tendenziell mehr Informationen enthält als auf Ebene 4 verarbeitet werden können, sollte bei zufälliger Ziehung den beiden Urteilen teilweise unterschiedliche Informationen zugrunde liegen. Die Person aus Bayern würde beispielsweise unterschiedliche Städte heranziehen, um die Bevölkerungszahl von Celle zu schätzen, wodurch sich tendenziell eine positive Bracketing Rate für die Urteile ergibt und Urteilsfehler (teilweise) durch die Bildung eines Mittelwerts korrigiert werden können (siehe auch Hourihan & Benjamin, 2010).

Für die Weisheit der Vielen in einer Person sollten sich nach dem GSM auf Ebene 2 die gleichen Limitationen ergeben wie bei der klassischen Weisheit der Vielen, da auch zwei Urteile der *gleichen* Person (zu $t_1$ und $t_2$) indirekt auf den gleichen Informationspool auf Ebene 2 zurückgehen und dieser begrenzt ist (siehe Abbildung 1). Je kleiner der geteilte Pool, desto eher liegen auch bei der gleichen Person anteilig mehr gleiche Informationen zugrunde. Für die Weisheit der Vielen in einer Person ergibt sich im Unterschied zur klassischen Weisheit der Vielen zusätzlich eine zweite Limitation, da die gleiche Person immer nur neue Informationen heranziehen kann, die vorher auch im Langzeitgedächtnis abgespeichert wurden (Ebene 3). Zumindest kurzfristig sollten sich die Restriktionen auf Ebene 3 dahingehend, welche Informationen einer bestimmten Person zugänglich oder nicht zugänglich sind, nicht ändern. Beispielsweise kann die Person aus Bayern keine neuen Städte aus Schleswig-Holstein heranziehen, die sie selbst nicht kennt. Somit sollten bei zwei Urteilen der gleichen Person anteilig mehr gleiche Informationen zugrunde liegen als bei zwei Urteilen unterschiedlicher Personen. So ergeben sich bei zwei Urteilen von der gleichen Person Bracketing Rates im Bereich von 14% und 22% (Herzog und Hertwig, 2014b), die zwar positiv sind, aber niedriger als bei zwei *unterschiedlichen* Personen (siehe auch Herzog & Hertwig, 2014a).

Anhand des GSM lässt sich also ableiten, dass für Menschen häufig ein großes Potential zur Reduzierung von Fehlern besteht. Wie eingangs erwähnt, zeigt sich jedoch empirisch, dass Menschen dieses Potential nur unzureichend nutzen (z.B., Müller-Trede, 2011; Soll & Larrick, 2009). Aus dem GSM lässt sich eine Lösung für dieses Problem ableiten,

indem Menschen dieses Potential selbst erkennen und dadurch (mit Hilfe von einfachen Interventionen) auch zur Verbesserung ihrer Urteile nutzen können. Im Folgenden werde ich zunächst für den individuellen Urteilskontext zeigen, inwiefern Menschen selbst ihr Korrekturpotential erkennen können, beziehungsweise inwiefern sie erkennen können, in welche Richtung sie ihre Urteile korrigieren sollten. Auf konkrete Interventionen zur Urteilsverbesserung in diesem individuellen Urteilskontext gehe ich in Abschnitt 5 ein. In Abschnitt 6 gehe ich detaillierter auf einen sozialen Urteilskontext ein, bei dem Menschen Ratschläge anderer Personen erhalten und bewerten sollen, inwiefern ihnen diese bei der Korrektur ihrer Urteile helfen können.

### 2.3.2. Das Weisheitsbewusstsein nach dem GSM

Ein wesentliches Merkmal des GSM besteht darin, dass Menschen nach dem GSM die gleichen Informationen für sehr unterschiedliche Urteile zum gleichen Sachverhalt (z.B. Schätzung der Bevölkerung der Stadt Celle) nutzen können. Bei solchen Urteilen kann es sich um quantitative Urteile handeln („Celle hat eine Bevölkerungszahl von 30.000“), aber auch metakognitive Urteile zur Urteilssicherheit („die Bevölkerungszahl von Celle liegt mit 80% Wahrscheinlichkeit/Sicherheit im Bereich 10.000 bis 60.000“) und auch Urteile zur Richtung von Fehlern („die Bevölkerungszahl von Celle liegt eher über als unter meiner ersten Schätzung von 30.000“). Die Informationen, die diesen Urteilen zugrunde liegen, folgen dabei nach dem GSM immer dem in Abbildung 1 dargestellten hierarchischen Sampling-Prozess. Dies ist insbesondere unabhängig davon der Fall, ob es sich um metakognitive Urteile (Urteile über andere Urteile) oder einfache quantitative Urteile handelt. So gibt es verschiedene Sampling-Ansätze, die sich entweder mit metakognitiven Urteilen beschäftigen, vornehmlich mit der subjektiven Urteilssicherheit (z.B. Juslin et al., 2007), oder mit quantitativen Urteilen und der Weisheit der Vielen (z.B., Hourihan & Benjamin, 2010; Vul & Pashler, 2008). Der vergleichsweise enge Fokus dieser Modelle auf bestimmte Formen von Urteilen oder bestimmte Phänomene führt jedoch dazu, dass die Weisheit der Vielen und metakognitive Urteile nicht in einen direkten Zusammenhang gestellt werden, auch wenn sich die zentralen Annahmen dieser Ansätze weitestgehend überschneiden (für eine Ausnahme bei dichotomen Urteilen siehe Koriat, 2012a, 2012b, 2019).

Nach dem GSM liegen Unterschiede in den unterschiedlichen Formen von Urteilen lediglich darin, dass sich gegebenenfalls der Bezugspunkt und gegebenenfalls das Antwortformat des Urteils ändert (z.B. Punktschätzung des wahren Werts vs. Richtungsurteil über die Richtung des eigenen Fehlers oder des wahren Werts). Dies ist zumindest so, solange der Sachverhalt der gleiche bleibt. Der Sampling-Prozess unterscheidet sich jedoch nicht zwischen den Urteilen. Es werden lediglich auf der untersten Ebene andere Antworten aus der Informationsstichprobe extrapoliert. Beispielsweise schätzte die Person aus Bayern in der Einleitung anhand von Coburg, Forchheim und Lichtenfels die Einwohnerzahl von Celle auf 30.000. Sie könnte auch gefragt werden, ob der wahre Wert über oder unter 20.000 liegt und könnte hierfür die gleichen Informationen heranziehen, mitteln und „drüber" sagen. Kritisch wird es, wenn sie gefragt wird, ob der wahre Wert eher über oder eher unter ihrem eigenen Urteil (30.000) liegt. Sie könnte keine explizite Richtung wählen, wenn genau die gleichen Informationen herangezogen und gemittelt würden. Nach dem GSM liegen in diesem Fall jedoch nicht zwangsläufig die *gleichen* Informationen zugrunde, da durch die Bewertung des Urteils das Sampling der Informationen erneut ausgelöst wird. Mit dem erneuten Sampling lässt sich das Weisheitsbewusstsein ableiten. So könnte die Person, wie eingangs erwähnt, im zweiten Schritt anstelle von Coburg, Bamberg als neue Information mit einer Bevölkerungszahl von ca. 75.000 aus ihrem Gedächtnis abrufen, und würde richtigerweise angeben, dass der wahre Wert eher über 30.000 liegt. In diesem Fall hätte die Person aus Bayern die Richtung ihres Fehlers erkannt und würde ein Weisheitsbewusstsein besitzen.

Noch einmal zusammengefasst lässt sich das Weisheitsbewusstsein wie folgt ableiten: Mit dem GSM nehme ich an, dass Menschen bei der Einschätzung der Richtung des Fehlers ihrer Urteile erneut auf ihren privaten Informationspool zugreifen können (Ebene 3) und zufällig eine begrenzte Menge von Informationen ziehen. Anhand dieser Informationen können sie bewerten, ob ihr ursprüngliches Urteil eher in die eine oder andere Richtung angepasst werden sollten, oder, als dritte Option, weder in die eine noch in die andere Richtung. Ferner nehme ich an, dass Menschen hierfür zu dem neuen Zeitpunkt, $t_{i+1}$, ein neues quantitatives Urteil simulieren (d.h. sie fragen sich, wie sie zu diesem Zeitpunkt urteilen würden). Dadurch lösen sie erneut den Sampling-Prozess aus, der der hierarchischen Struktur in Abbildung 1 folgt. Der Unterschied zu einem „normalen"

quantitativen Urteil liegt lediglich darin, dass im Anschluss eingeschätzt wird, ob das simulierte Urteil (1) über dem ersten Urteil liegt, (2) oder ob es darunterliegt, oder (3) ob es weder darüber noch darunter liegt (d.h. die Person kommt zu demselben Urteil wie zuvor). Nach dem GSM kommt es damit primär dann zu einem Richtungsurteil (Option 1 oder 2), wenn sich die Informationslage bei der Richtungseinschätzung – wie im oben genannten Beispiel – ändert. In diesen Fällen sollten die Richtungsurteile überzufällig häufig (d.h. in mehr als 50% der Fälle) in die Richtung des wahren Werts zeigen, da die neuen Informationen eine (zumindest teilweise) unabhängige Einschätzung der Richtung des wahren Werts erlauben. Zudem sollten Menschen in der Lage sein, ihre Urteile im Schnitt zu verbessern, sofern sie ihr erstes Urteil teilweise in Richtung des simulierten Urteils anpassen, also eine Anpassung relativ zum simulierten Urteil um mehr als 0% aber weniger als 100% vornehmen (siehe Herzog & Hertwig, 2014b).

Wesentlich ist bei der Ableitung des Weisheitsbewusstseins die Annahme, dass Menschen meist nur einen Teil der Informationen abrufen können, die sie auf Ebene 3 verfügbar haben, sonst wäre in dem Beispiel Bamberg zusammen mit Coburg schon in das erste Urteil eingeflossen. Nach dem GSM kommt dies im Wesentlichen durch die Beschränkung des Arbeitsgedächtnisses zustande (Cowan, 2001; Miller, 1956), da Menschen zu einem Zeitpunkt nur eine begrenzte Menge von Informationen abrufen und verarbeiten können (die Person in dem Beispiel war mit drei Informationen schon immer ausgelastet). Es ist jedoch auch möglich, dass bestimmte Informationen kurzzeitig nicht abrufbar sind oder aus Zeitdruck oder motivationalen Gründen nicht immer alle abgerufen oder verwendet werden (siehe z.B. Epley & Gilovich, 2006).

Mit dem GSM lässt sich jedoch nicht nur ableiten, dass Menschen ein Weisheitsbewusstsein besitzen, sondern auch, inwiefern das Phänomen beschränkt ist. Die Limitationen des Weisheitsbewusstseins sind nach dem GSM die gleichen wie bei der Weisheit der Vielen in einer Person und der klassischen Weisheit der Vielen, da auch die metakognitiven Urteile zum Weisheitsbewusstsein letztlich auf den gleichen Informationspool zurückgehen. Das heißt, dass Menschen Korrekturpotentiale nicht erkennen können, wenn ihr privater Informationspool auf Ebene 3 so beschränkt ist, sodass sie keine oder nur wenig neue Informationen abrufen können. Sie können auch ihre Fehler nicht erkennen, wenn der (begrenzte) Informationspool auf Ebene 2 zufällig oder auch

systematisch verzerrt ist. In diesem Fall sollte auch die Gesamtpopulation von Entscheidungsträger:innen den wahren Wert mit ihren Urteilen systematisch über- oder unterschätzen. Für die Richtungsurteile heißt dies, dass Menschen nach dem GSM vielmehr den Populationsmittelwert vieler Urteile (die klassische Weisheit der Vielen) detektieren als den wahren Wert, auch wenn sie nach der Richtung des wahren Werts explizit gefragt werden.

Dieser Umstand lässt sich an einem Beispiel erklären, bei dem der geteilte Informationspool auf Ebene 2 in Richtung niedrigerer Werte verzerrt ist. Bei der Schätzung der Bevölkerungszahl von Celle könnten die Entscheidungsträger:innen einer Population beispielsweise nur die Bevölkerungszahlen anderer Städte kennen, die tendenziell (d.h. im Mittel aber nicht in allen Fällen) kleiner sind als Celle. Werden aus diesem Pool zufällig Informationen gezogen und daraus quantitative Urteile gebildet, so sollte der Mittelwert aller Urteile (die Weisheit der Vielen) unter 69.399 liegen (der wahre Wert), zum Beispiel bei 60.000. Eine Person könnte jetzt zufällig (durch eine zufällige Auswahl der Informationen von Ebene 2) ein quantitatives Urteil zwischen wahrem Wert und Populationsmittelwert treffen und beispielsweise 65.000 schätzen. Sollte diese Person gefragt werden, ob der wahre Wert eher über oder eher unter 65.000 liegt, so sollten neue Informationen eher in Richtung 60.000 als in Richtung 69.399 tendieren, da auch alle „neuen" Informationen im Schnitt diesen Wert approximieren (sie kommen auch aus dem verzerrten Pool). Hierbei handelt es sich um einen Regression-zur-Mitte-Effekt (Fiedler & Krueger, 2012), wobei der Populationsmittelwert die Mitte bildet und nicht der wahre Wert. Die Bezeichnung *Weisheitsbewusstsein* bekommt damit eine zweite Bedeutung, weil im Schnitt die *Weisheit der Vielen* (der Populationsmittelwert) detektiert wird und nicht der wahre Wert. Da Populationsmittelwert und wahrer Wert jedoch bekanntlich häufig nah beieinanderliegen (die Weisheit der Vielen; Herzog et al., 2019; Larrick et al., 2012), sollten Menschen insgesamt in der Lage sein, hiermit die Richtung des wahren Werts zu erkennen, was der eigentliche Wortsinn des Weisheitsbewusstseins ist.

### 2.4. Bisherige Evidenz für das Weisheitsbewusstsein

Bisher gibt es nach meinem Kenntnisstand keine direkte Evidenz dafür, dass Menschen ein Weisheitsbewusstsein besitzen. Es gibt jedoch Evidenz dafür, dass

wesentliche Vorrausetzungen für dieses Phänomen gegeben sind. Zum einen zeigt die Literatur zur Weisheit der Vielen in einer Person, dass Menschen grundsätzlich in der Lage sind, zwei unabhängige Urteile abzugeben (z.B., Hertwig & Herzog, 2014a; Stroop, 1932; Vul & Pashler, 2008), also zwei Urteile, deren Fehler nur anteilig korreliert sind und die deshalb teilweise auf unterschiedlichen Seiten des wahren Werts landen. Zudem gibt es in der Literatur zur subjektiven Urteilssicherheit Indizien dafür, dass Menschen eine teilweise unabhängige Bewertung ihrer subjektiven Konfidenzintervalle vornehmen können. So zeigt sich bei Winman et al. (2004), dass sich das Phänomen der Overconfidence reduziert, wenn Menschen nach der Angabe eines Konfidenzintervalls später einschätzen sollen, mit welcher Wahrscheinlichkeit der wahre Wert in dem angegebenen Intervall liegt. Zum Beispiel könnten die Menschen gefragt werden, in welchem Bereich die Bevölkerungszahl von Celle mit 80% Wahrscheinlichkeit liegt und „zwischen 30.000 unter 65.000“ antworten. Werden sie später gefragt, mit welcher Wahrscheinlichkeit der wahre Wert zwischen 30.000 und 65.000 liegt, würden sie beispielsweise 70% und nicht 80% antworten[2]. Dieser Befund lässt sich damit erklären, dass für die spätere Bewertung teilweise neue Informationen herangezogen werden (z.B. Bamberg statt Coburg), die insgesamt seltener in dem Intervall landen als die alten Informationen zur Generierung des Intervalls, wobei auch das Antwortformat hierbei eine wesentliche Rolle spielt (für eine detailliertere Erklärung siehe auch Juslin et al., 2007). Dabei ist unklar, ob Menschen auch eine von dem ersten Urteil unabhängige Einschätzung vornehmen können, wenn sie die Richtung einschätzen, in der sie ihr eigenes Urteil korrigieren sollten. Ein solches Urteil stellt etwas höhere oder zumindest andere Anforderungen als ein Urteil über die Urteilssicherheit, da Menschen nicht nur annehmen müssen, dass Fehler entstanden sind, sondern auch, dass diese in einer bestimmten Richtung liegen (Wilson & Brekke, 1994). Es stellt sich damit die Frage, warum Menschen diese Fehler nicht schon mit dem ersten Urteil korrigieren, wenn sie die Richtung ihrer Fehler bereits kennen und zumindest bei einer monetären Incentivierung für akkurate Urteile hierzu auch motiviert sein sollten.

[2] Mit dem Beispiel kann zunächst nur gezeigt werden, dass sich die Sicherheit in das gleiche Intervall reduziert, da der wahre Wert nur innerhalb oder außerhalb des Intervalls liegen kann. Um zu testen, ob sich hiermit Overconfidence reduziert, müssten mehrere Stimuli getestet werden, an denen sich testen lässt, wie viel Prozent der wahren Werte insgesamt in den Intervallen liegen und inwiefern diese Zahl von der geforderten Wahrscheinlichkeit abweicht.

Darüber hinaus gibt es mehrere Arbeiten, die sich zumindest indirekt mit dem Weisheitsbewusstsein beschäftigen und versuchen, mit sehr ähnlichen Annahmen die Weisheit der Vielen in einer Person zu steigern. Diese Arbeiten liefern jedoch inkonsistente Ergebnisse. Ein prominentes Beispiel hierfür ist das sogenannte *Dialectical Bootstrapping* bei der Abgabe zweiter Urteile, bevor diese mit einem ersten (eigenen) Urteil gemittelt werden (Herzog & Hertwig, 2009, 2014b). Gemäß den Instruktionen zum Dialectical Bootstrapping sollen Menschen einschätzen, ob ihr erstes Urteil zu hoch oder zu niedrig war, und dann ein zweites Urteil auf Basis von gegenteiliger Evidenz generieren. Dazu sollten nur Menschen mit einem Weisheitsbewusstsein in der Lage sein. Herzog und Hertwig (2009, 2014b) berichten, dass das Dialectical Bootstrapping die Leistung der Weisheit der Vielen in einer Person steigert, was auch für ein generelles Weisheitsbewusstsein spricht. Auf der anderen Seite gibt es eine Reanalyse der früheren Arbeit von Herzog und Hertwig (2009) von White und Antonakis (2013), die für solche Strategien keinen Mehrwert zur Nutzung der Weisheit der Vielen in einer Person findet (siehe aber auch Herzog & Hertwig, 2013). Zudem gibt es eine unabhängige Arbeit von Gaertig und Simmons (2021) mit ähnlichen Instruktionen, die zeigt, dass sich durch die explizite Einschätzung der Richtung des wahren Werts vor einem zweiten Urteil die Weisheit der Vielen in einer Person sogar reduziert. Gaertig und Simmons argumentieren daher, dass Menschen bei solchen Richtungsurteilen primär raten, in welcher Richtung der wahre Wert eher liegt.

Nach der Argumentation von Gaertig und Simmons (2021) sollten Menschen kein Weisheitsbewusstsein besitzen. Ihre Studien unterstützen vermeintlich diese Vorhersage, auch wenn eine Testung eines Weisheitsbewusstseins nicht im Mittelpunkt ihrer Arbeit steht (sondern wie die Weisheit der Vielen in einer Person am besten genutzt wird). Bezüglich eines Weisheitsbewusstseins gibt es jedoch ein methodisches Problem, welches erklärt, warum Gaertig und Simmons ein Weisheitsbewusstsein nicht finden (konnten). So wurden die Versuchspersonen nach einem ersten Urteil explizit gefragt, ob der wahre Wert eher über oder eher unter ihrem ersten Urteil liegt und sollten explizit eine der beiden Richtungen wählen, bevor sie ein zweites quantitatives Urteil abgaben. Insgesamt zeigt sich, dass die Wahl der korrekten Richtung bei den meisten ihrer Studien nahe 50% lag (bzw. Anpassungen des zweiten quantitativen Urteils in die korrekte Richtung nahe 50% lagen), auch wenn Gaertig und Simmons teilweise Indikatoren finden, dass Menschen bei solchen

Fragen teilweise auch informierte Richtungsurteile treffen (z.B. bei sehr hohen oder sehr niedrigen wahren Werten). Ein entscheidender Unterschied zu dieser Arbeit ist jedoch, dass die Versuchspersonen nie die Option hatten, keine der beiden Richtungen zu wählen. Zum einen sollte dies die Wahrscheinlichkeit, die korrekte Richtung zu wählen, deutlich senken, da Menschen auch nach dem GSM teilweise keine neuen Informationen heranziehen können und in diesen Fällen tatsächlich gezwungen sind „nur zu raten", in welcher Richtung der wahre Wert eher liegt. Zum anderen lässt sich ein Weisheitsbewusstsein als etwas, was den Menschen tatsächlich *bewusst* ist, nur testen, wenn Menschen auch die Option haben keine Richtung zu wählen und damit auch angeben können, dass sie keine Intuition bezüglich der Richtung des wahren Werts haben.

Im ersten empirischen Abschnitt dieser Arbeit werde ich daher das Weisheitsbewusstsein als wesentliche Grundlage dieser Arbeit in einer eigenen Serie von Studien direkt testen. Hierfür wähle ich ein Vorgehen, bei dem Menschen erst mehrere quantitative Urteile abgeben und in einem zweiten Teil einschätzen sollen, ob der wahre Wert (1) eher über, (2) eher unter, oder (3) weder eher über noch eher unter dem ersten Urteil liegt. In Abschnitt 5 und 6 berichte ich zudem Studien, bei denen Menschen zuletzt ein zweites quantitatives Urteil abgeben und ihre ersten Urteile korrigieren können. In Abschnitt 6 erhalten sie hierfür teilweise zusätzlich den Ratschlag einer anderen Versuchsperson. Bevor ich jedoch detaillierter auf Methodik und Ergebnisse dieser Studien eingehe, werde ich zunächst im nächsten Abschnitt im Sinne einer transparenten Darstellung mein allgemeines Vorgehen bei der Datenerhebung und -auswertung beschreiben.

## 3. Vorgehen bei der Datenerhebung und -auswertung

Im Rahmen meiner Dissertation wurden acht Studien durchgeführt, um zu testen, inwiefern Menschen ein Weisheitsbewusstsein ausbilden können und inwiefern sie dieses mit Hilfe von einfachen Interventionen zur Verbesserung ihrer Urteile nutzen können. Alle Studien werden im Folgenden mitsamt aller Datenausschlüsse berichtet[3]. Sechs von acht Studien wurden im Open Science Framework präregistriert (siehe Tabelle 1). Die Präregistrierungen beschränkten sich dabei meist auf zentrale Tests bezüglich des

[3] Ausgenommen sind nur Datenausschlüsse bei Onlinestudien, bei denen die Versuchspersonen die Studie zwar gestartet haben, jedoch vor der Zuteilung zu den jeweiligen Versuchsbedingungen abgebrochen haben.

Weisheitsbewusstseins und der Wirksamkeit der Interventionen. Zur einfachen Lesbarkeit wird im Folgenden auf alle präregistrierten Hypothesen, Hypothesentests und andere präregistrierte Verfahren (z.B. Poweranalysen) mit Kürzeln verwiesen, die in Tabelle 1 mit Links zu den Präregistrierungen zu finden sind (z.B. P1 für die Präregistrierung von Studie 1). Zudem wurden für die vorliegende Arbeit teilweise neue Hypothesen aus dem GSM abgeleitet, die nicht präregistriert wurden. Diese Hypothesen erlauben jedoch insgesamt einen kritischeren Test des GSM und die Ergebnisse der entsprechenden Tests werden als konfirmatorische Evidenz für oder gegen das GSM beziehungsweise die entsprechenden Hypothesen betrachtet. Als Inferenzkriterium wurde in allen Fällen ein $p$-Wert < .05 angesetzt. Bei den präregistrierten Analysen wurde hierfür teilweise eine einseitige Testung explizit präregistriert, um die entsprechende Hypothese mit größerer statistischer Power zu testen. Alle anderen Vergleiche wurden zweiseitig getestet. Generell wurde bei allen präregistrierten Tests mit einer Power von mindestens $1–\beta = .80$ als angemessene Teststärke geplant. Sofern zusätzliche finanzielle Ressourcen zur Verfügung gestellt werden konnten, wurde jedoch an entsprechender Stelle mit einer höheren Power geplant, was meist der Fall war. Zudem sind in Tabelle 1 im Sinne eines transparenten Forschungsprozesses der Erhebungsstart aller Studien gelistet, da die Reihenfolge der Studien im Text zur strukturierten und einfach verständlichen Beantwortung der Forschungsfragen teilweise geändert wurde.

Von den Studien wurden fünf im Labor und drei als Onlinestudie durchgeführt (siehe letzte Spalte in Tabelle 1). Alle Laborstudien wurden an der Universität Göttingen erhoben und waren in deutscher Sprache verfasst. Die Rekrutierung der Versuchspersonen erfolgte für diese Studien über Onlinepostings, Aushänge, Campusakquise und über den Proband:innenpool der Abteilung für Wirtschafts- und Sozialpsychologie der Universität Göttingen mit Hilfe des Rekrutierungssystems ORSEE (Greiner, 2015). Bei den Stichproben handelt es sich damit überwiegend um Studierende. Alle Laborstudien wurden am Computer bearbeitet und mit dem Framework Alfred Version 0.2b5 (Treffenstädt & Wiemann, 2018) programmiert. Die Onlinestudien wurden über die digitale Infrastruktur der Duke University erhoben und konnten am Computer oder mobilen Endgeräten (Tablets, Smartphones) bearbeitet werden. Alle Onlinestudien wurden mit der Software Qualtrics (https://www.qualtrics.com/) programmiert und die Versuchspersonen über die Plattform

Amazon Mechanical Turk (MTurk, https://www.mturk.com/) rekrutiert. Hierbei handelte es sich um US-amerikanische Stichproben. Entsprechend waren alle Onlinestudien in englischer Sprache verfasst. Die Einladungen hierzu erfolgten über die Plattform *CloudResearch* (früher *TurkPrime*; https://www.cloudresearch.com/; Litman et al., 2017), die es erlaubt, gezielte Personengruppen bei MTurk zu rekrutieren und hierüber eine hohe Datenqualität zu sichern (beispielsweise indem nur Personen eingeladen wurden, deren Arbeit bei MTurk üblicherweise auch akzeptiert und nicht abgelehnt wird). Zudem erhielten alle Onlinestudien zur Durchführung ein positives Votum der Ethikkommission der Duke University vor der jeweiligen Datenerhebung. Die Materialen (PDFs mit Screenshots aller Seiten in den Experimenten) sind für alle Studien öffentlich zugänglich im OSF zu finden (https://osf.io/y23fs/)[4].

**Tabelle 1**

*Übersicht über Studien und Präregistrierungen*

| Erhebungsstart | Studie im Text | Präregistrierung/URL | Setting |
|---|---|---|---|
| 25.10.2018 | Studie 1 | P1: https://osf.io/vn64m | Labor |
| 29.3.2019 | Studie 2 | P2: https://osf.io/em3ga | Labor |
| 11.7.2020 | Studie 3 | - | Online |
| 17.6.2019 | Studie 4 | P4: https://osf.io/tychu | Online |
| 27.8.2019 | Studie 5 | P5: https://osf.io/hb4pu | Labor |
| 17.1.2020 | Studie 6 | P6: https://osf.io/7hr92 | Labor |
| 2.3.2020 | Studie 7 | - | Labor |
| 10.8.2020 | Studie 8 | P8: https://osf.io/khnz9 | Online |

[4] Teilweise wurde bei der Durchführung der Studien urheberrechtlich geschütztes Bildmaterial verwendet, für das im Text auf die entsprechenden Quellen verwiesen wird, das jedoch nicht im OSF geteilt wurden. Alle Fragen/Stimuli aus den anderen Studien befinden sich in Anhang A mit den entsprechenden Quellen und zusätzlich unter https://osf.io/xrqd2/.

Alle erhobenen Daten wurden mit R Version 3.6.2 (R Core Team, 2019) über die Bedienoberfläche RStudio Version 1.2.5033 (RStudio Team, 2019) ausgewertet. Für die Testplanung, Präregistrierung, Aufbereitung, Auswertung und Visualisierung der Daten wurden zudem die R Pakete *rmarkdown* (Allaire et al., 2021), *prereg* (Aust, 2019), *papaja* (Aust & Barth, 2018), *lme4* (Bates et al., 2015), *pwr* (Champely, 2018), *devEMF* (Johnson, 2020), *lmerTest* (Kuznetsova et al., 2017), *lsr* (Navarro, 2015), *yarrr* (Phillips, 2017), *mediation* (Tingley et al., 2014), *knitr* (Xie, 2021), und die R Pakete aus dem *tidyverse* (Wickham et al., 2019) verwendet. Die Daten und das Auswertungsskript sind im OSF unter https://osf.io/y23fs/ öffentlich zugänglich.

## 4. Studien zum Nachweis des Weisheitsbewusstseins

In dem ersten empirischen Abschnitt werde ich testen, inwiefern Menschen ein Weisheitsbewusstsein besitzen. Hierfür werde ich im Folgenden die Ergebnisse von Studien berichten, bei denen Menschen zunächst mehrere quantitative Urteilsaufgaben bearbeiteten (Phase 1) und im Anschluss ausgehend von diesen Urteilen die Richtung des wahren Werts einschätzten (Phase 2). Konkret sollten die Versuchspersonen in Phase 2 für ihre Urteile aus der ersten Phase einschätzen, ob der wahre Wert eher über, eher unter, oder weder eher über noch eher unter ihren Initialurteilen liegt. Es soll zum einen untersucht werden, wie häufig Menschen sich für eine der beiden Richtungen entscheiden, und, sofern sie sich für eine Richtung entscheiden, wie häufig sie hiermit richtigliegen. Nach dem GSM sollten Menschen immer dann eine Richtung wählen, wenn sie ein zweites Richtungsurteil auf Basis einer zumindest teilweise veränderten Informationslage mit teilweise neuen Informationen bilden. In den Studien wird deskriptiv untersucht, wie häufig dies der Fall war, da das GSM zumindest in diesen Studien hierzu keine genaue Vorhersage zulässt (abgesehen von Studie 2). Die zentrale Vorhersage bezüglich des Weisheitsbewusstseins ist, dass Menschen mit ihren Richtungsurteilen überzufällig häufig richtigliegen (in mehr als 50% der Fälle), sofern sie eine der beiden Richtungen gewählt haben (siehe Abschnitt 2.3.2.). Hierbei handelt es sich um das wesentliche Phänomen, auf dem später alle weiteren Interventionen aufbauen:

**Hypothese 1 (H1; P1-P8):** Menschen sind der Lage, die Richtung des wahren Wertes ausgehend von einem eigenen Urteil überzufällig häufig richtig einzuschätzen (d.h., in mehr als 50% der Fälle), sofern sie sich für eine der beiden Richtungen entscheiden.

Das Weisheitsbewusstsein unterliegt nach dem GSM den gleichen Limitationen wie die Weisheit der Vielen, da Menschen nur so gut in der Lage sein können, die Richtung des wahren Werts zu detektieren, wie die Gesamtmenge von Informationen dies zulässt (siehe Abschnitt 2.3.2.). Damit kann das Weisheitsbewusstsein auch nur die Fehler erkennen, die auch durch die Weisheit der Vielen korrigiert werden können. Für die Richtungsurteile lässt sich hieraus die Vorhersage ableiten, dass mit den expliziten Richtungsurteilen der Populationsmittelwert aller Initialurteile detektiert wird (die beste Schätzung, die auf Basis des geteilten Informationspools möglich ist). In Anbetracht der sehr positiven Evidenz zur Weisheit der Vielen (Herzog et al., 2019; Larrick et al., 2012), sollten wahrer Wert und Populationsmittelwert bei einer repräsentativen Auswahl von Stimuli häufig nah beieinanderliegen und in die gleiche Richtung zeigen, weshalb auch das Weisheitsbewusstsein im Schnitt die Richtung des wahren Werts detektieren sollte. Um zu untersuchen, ob Menschen hierbei jedoch insgesamt eher die Richtung des wahren Werts oder – wie es das GSM vorhersagen würde – die des Populationsmittelwerts wählen, sind insbesondere die Fälle diagnostisch, in denen Populationsmittelwert und wahrer Wert in unterschiedliche Richtungen zeigen. Für einen gründlichen Test des GSM wurden diesbezüglich zwei Hypothesen aufgestellt, da von vornherein nicht davon ausgegangen werden kann, dass die kritischen Fälle immer in ausreichender Zahl zur Verfügung stehen (da immer eine möglichst repräsentative Auswahl von Stimuli verwendet wurde und keine, bei der wahrer Wert und Populationsmittelwert häufiger besonders stark auseinanderliegen). Die erste Variante der Hypothese erlaubt einen generellen Test, ob die Daten insgesamt konsistent mit dem GSM sind. Die zweite Variante erlaubt hingegen den kritischen Test, ob Menschen auch eher zum Populationsmittelwert tendieren, wenn der wahre Wert in der anderen Richtung liegt. Bei niedrigen Fallzahlen kann diese zweite Variante gegebenenfalls erst später metaanalytisch über mehrere Studien teststark untersucht werden:

**Hypothese 2a (H2a; nicht präregistriert):** Die Richtungsurteile von Menschen liegen ausgehend von eigenen Urteilen überzufällig häufig (in mehr als 50% der Fälle) in der

Richtung des Populationsmittelwerts (der Weisheit der Vielen), sofern sich Menschen für ein Richtungsurteil entscheiden.

**Hypothese 2b (H2b; nicht präregistriert):** Die Richtungsurteile von Menschen liegen ausgehend von eigenen Urteilen überzufällig häufig (in mehr als 50% der Fälle) in der Richtung des Populationsmittelwerts, sofern sich Menschen für ein Richtungsurteil entscheiden und wahrer Wert und Populationsmittelwert in unterschiedlichen Richtungen liegen.

## 4.1. Studie 1

Hauptziel von Studie 1 war es, mit einer breiten Menge unterschiedlicher Aufgaben zu testen, inwiefern ein Weisheitsbewusstsein vorliegt (H1). Hierfür wurde eine Beobachtungsstudie mit zwei Phasen und ohne experimentelle Manipulation im Labor durchgeführt. In der ersten Phase gaben die Versuchspersonen am Computer quantitative Urteile zu insgesamt 30 unterschiedlichen Schätzaufgaben ab, die alle einen Bezug zum Thema Globalisierung hatten, bei denen es sich jedoch um unterschiedliche Typen von Aufgaben handelte (z.B. historische Schätzungen, Prozentschätzungen und Schätzungen von Entfernungen). In der zweiten Phase wurden diese Aufgaben in der gleichen Reihenfolge erneut präsentiert. Im Unterschied zur ersten Phase wurden den Versuchspersonen jedoch zusätzlich ihre ersten Urteile angezeigt und ihre Aufgabe war es einzuschätzen, ob der wahre Wert (1) eher über, (2) eher unter oder (3) weder eher über noch eher unter ihrem ersten Urteil lag.

### 4.1.1. Stichprobe

Maßgeblich für die Bestimmung der Stichprobengröße von Studie 1 war der Test des Weisheitsbewusstseins (H1) mit einem einseitigen Einstichproben-*t*-Test, mit dem später die relative Häufigkeit korrekter Richtungsurteile pro Person gegen 50% getestet wurde (unter Ausschluss der Urteile in keiner der beiden Richtungen). In einer präregistrierten Poweranalyse (P1) wurde für diesen Test eine statistische Power von $1–\beta = .90$ angestrebt und zur Berechnung der benötigten Stichprobengröße ohne weitere Vorinformationen ein kleiner bis mittelgroßer Effekt von $d = 0.3$ (oder größer) angenommen (Cohen, 1988). Für ein

Weisheitsbewusstsein mit einem kleineren Effekt wurde angenommen, dass sich dieses im Rahmen der darauf aufbauenden Interventionsstudien nicht mehr praktisch nutzen lässt oder solch ein Nutzen zumindest nicht im Rahmen von finanzierbaren Studien reliabel nachgewiesen werden kann. Die Poweranalyse hat eine Minimalstichprobe von 97 Versuchspersonen ergeben, die auf 100 Versuchspersonen aufgerundet wurde (siehe P1). Als zusätzliches Ausschlusskriterium wurde präregistriert (P1), dass alle Versuchspersonen ausgeschlossen und ersetzt werden würden, die sich in Phase 2 in 25 oder mehr der 30 Trials für keine der beiden Richtungen entschieden. Der Grund hierfür lag darin, dass sich in diesen Fällen das Weisheitsbewusstsein (H1) vergleichsweise schlecht schätzen lässt und gegebenenfalls hierdurch die notwendige Testpower für einen fairen Test der H1 fehlt. Dieses Kriterium traf jedoch in dieser ersten Studie in keinem Fall zu. Es wurde deshalb auch in keiner der weiteren Studien und Präregistrierungen formuliert, da im weiteren Verlauf immer von einer relativ hohen Anzahl expliziter Richtungsurteile ausgegangen wurde.[5] Während der Erhebung ergaben sich unerwartet in fünf Fällen technische Probleme, wodurch die Versuchspersonen die Studie nicht beenden konnten. Diese Versuchspersonen wurden ausgeschlossen und durch fünf zusätzliche Versuchspersonen ersetzt. Diese Entscheidung wurde während der Erhebung getroffen und damit vor der Analyse der Daten. Insgesamt wurden damit Daten von 100 Personen (73 Frauen und 27 Männer) mit einem Durchschnittsalter von $M$ = 23.10 ($Mdn$ = 22, $SD$ = 3.83) Jahren in der Analyse berücksichtigt, bei denen es sich überwiegend um Studierende der Universität Göttingen handelte. Alle Versuchspersonen erhielten entweder 3€ oder eine halbe Versuchspersonenstunde als Grundvergütung und unabhängig von der Wahl der Grundvergütung einen leistungsabhängigen Bonus von bis zu 3€ ($M$ = 0.72, $SD$ = 0.23).

#### 4.1.2. Aufgaben

Das Aufgabenset bestand in Studie 1 aus 30 verschiedenen Schätzaufgaben mit Bezug zum Thema Globalisierung, die mit den entsprechenden Quellen in Anhang A aufgelistet sind. Die Aufgaben lassen sich in drei Kategorien mit je 10 Fragen unterteilen (siehe Müller-Trede, 2011, für eine ähnliche Auswahl und Kategorisierung). Die erste Kategorie bestand

[5] Ein weiterer Grund für den Verzicht dieses Kriteriums in den späteren Studien war, dass hierbei teilweise mehreren Versuchsbedingungen erhoben wurden und es hierdurch zu artifiziellen Selektionseffekten kommen kann, sofern die Häufigkeit ein Richtungsurteil zu wählen systematisch mit der jeweiligen Manipulation zusammenhängt.

aus 10 Prozentschätzungen (z.B. „Wie viel Prozent der Weltbevölkerung lebte 2015 in einer Stadt?“), die am unteren und oberen Skalen Ende begrenzt sind (bei 0% und 100%). Die zweite Kategorie bestand aus 10 Jahresschätzungen („In welchem Jahr wurde die erste H&M-Filiale außerhalb Europas eröffnet?“), die am oberen Skalenende begrenzt sind, da die Ereignisse nicht nach dem Erhebungsjahr 2018 stattgefunden haben können. Bei der dritten Kategorie handelte es sich um eine gemischte Kategorie mit Aufgaben, deren Antwortformat durch einen natürlichen Nullpunkt nach unten begrenzt ist, aber nach oben zumindest theoretisch unbegrenzt ist (z.B. „Wie lange dauerte der erste Transatlantikflug der Lufthansa von Berlin nach New York?“ oder „Wie viele Filialen hatte IKEA 2016?“). Durch das Computerprogramm wurde der Bereich der möglichen Antworten bei Prozentschätzungen bei 0% und 100% begrenzt und bei der gemischten Kategorie bei null. Bei historischen Schätzungen hatten die Versuchspersonen grundsätzlich die Möglichkeit, Schätzungen außerhalb des möglichen Bereichs anzugeben, wobei keine Schätzung über dem Jahr 2014 lag.

### 4.1.3. Methode

Zur Durchführung von Studie 1 nahmen jeweils mehrere Personen zeitgleich in einem Computerlaborraum der Universität Göttingen teil. Jede Versuchsperson erhielt hierfür einen individuellen Computerarbeitsplatz und jede Session wurde von mindestens einer Versuchsleitung beaufsichtigt. Vor der Bearbeitung der Studie wurden alle Proband:innen mündlich instruiert, dass sie in der folgenden Studie mehrere Schätzaufgaben zum Thema Globalisierung bearbeiten würden. Alle Versuchspersonen wurden explizit darauf hingewiesen, dass Hilfsmittel, wie zum Beispiel das Verwenden von Smartphones oder Kommunikation mit anderen Teilnehmer:innen, während der Studie nicht erlaubt seien. Im Anschluss konnten die Versuchspersonen die Studie am Computer selbstständig bearbeiten. Auf der ersten Seite wurden alle Versuchspersonen aufgefordert, ihr Einverständnis zur Teilnahme an der Studie abzugeben. Die Einverständniserklärung umfasste allgemeine Informationen zu Ablauf, Länge und Vergütung der Studie, allgemeine Informationen zum Datenschutz, sowie die Einwilligungserklärung zur Datenerhebung und Datenverarbeitung.

Mit der Abgabe ihres Einverständnisses wurden die Versuchspersonen auf der nächsten Seite detaillierter über den Ablauf der Studie aufgeklärt, die in zwei Phasen

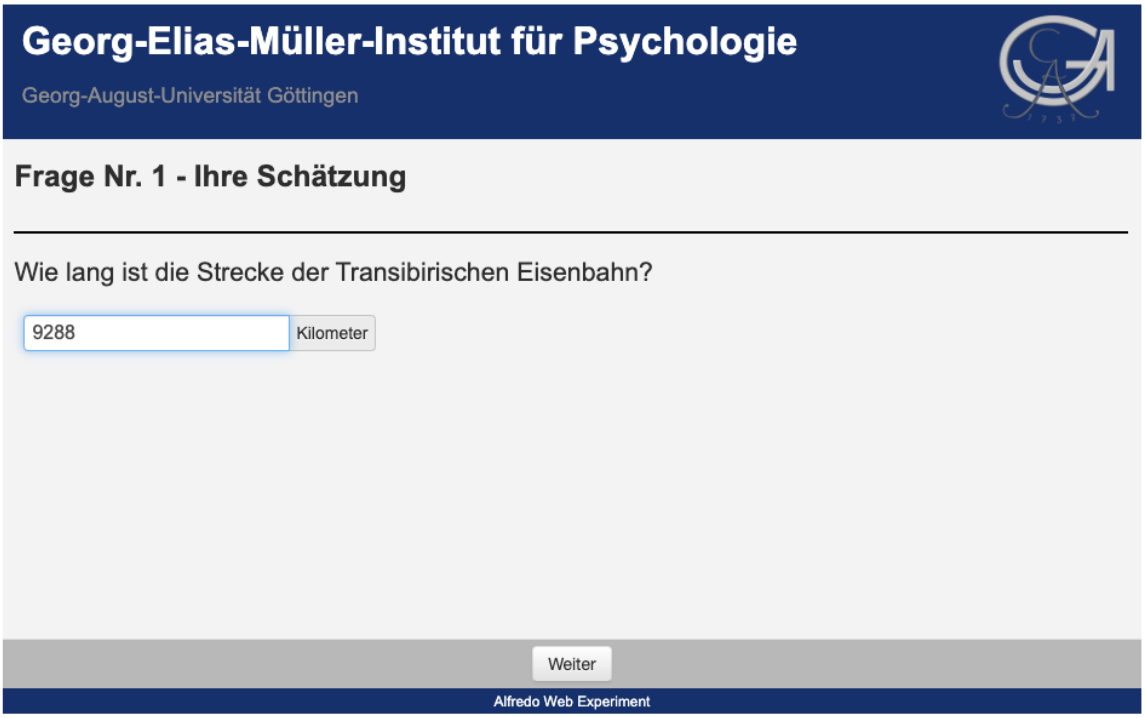
unterteilt war. Aufgabe der Versuchspersonen sei es in der ersten Phase, 30 Schätzaufgaben zum Thema Globalisierung möglichst genau zu beantworten. Für die zweite Phase erhielten die Versuchspersonen an dieser Stelle noch keine detaillierten Informationen und wurden lediglich darüber aufgeklärt, dass sie in der zweiten Phase wiederholende Fragen zu ihren Schätzungen beantworten würden, die ihr Wissen vertiefend abfragen würden. Zudem wurden die Versuchspersonen darüber informiert, dass sie einen leistungsabhängigen Bonus erhalten könnten, der von ihren Antworten in der ersten Phase abhing. Zuletzt wurden die Versuchspersonen noch einmal schriftlich daran erinnert, dass Hilfsmittel wie Smartphones in der Studie nicht erlaubt seien.

Auf der nächsten Seite erhielten die Versuchspersonen detaillierte Informationen zu der Vergütung und der Berechnung des erfolgsabhängigen Bonus, der von dem Aufgabentyp abhing. Für Prozentschätzungen („Wie viel Prozent der ...?“) erhielten die Versuchspersonen 10 Cent für jede Schätzung, die nicht weiter als 10 Prozentpunkte vom wahren Wert abwich; für historische Schätzungen („In welchem Jahr...?“) 10 Cent für jede Schätzung, die nicht weiter als 10 Jahre vom wahren Wert abwich. Für alle anderen Schätzungen wurde der prozentuale und nicht der absolute Fehler herangezogen und die Versuchspersonen erhielten 10 Cent für alle Schätzungen, die nicht weiter als 10% vom wahren Wert abwichen. Hierfür wurde der prozentuale Fehler an einem einfachen Beispiel erklärt: „Beispielsweise erhalten Sie 10 Cent, wenn Sie eine Distanz von 1000 km um nicht mehr als 100 km (10% von 1000) über- oder unterschätzen. Bei einer wahren Distanz von 100 km dürfen Sie diese jedoch nicht mehr als 10 km (10% von 100) über- oder unterschätzen.“ Auf der nächsten Seite wurden die Versuchspersonen darauf hingewiesen, dass sie eventuelle Boni unabhängig von ihrer Grundvergütung in Form von 3€ oder einer halben Versuchspersonenstunde erhielten. Auf der letzten Seite vor dem Start der ersten Phase wurden die Versuchspersonen darüber informiert, dass sie sich bei Fragen zu diesem Zeitpunkt mit einem Handzeichen bei der Versuchsleitung melden könnten, durch das Klicken auf „Zurück“ die Instruktionen ein weiteres Mal lesen könnten oder durch „Weiter“ mit der Studie beginnen könnten. Im weiteren Verlauf der Studie erhielten die Versuchspersonen keine weiteren Möglichkeiten, vorherige Seiten und eigene Angaben einzusehen oder vorherige Antworten zu ändern.

Auf der nächsten Seite startete Phase 1 der Studie und die Versuchspersonen bekamen auf insgesamt 30 Seiten je eine neue quantitative Urteilsaufgabe präsentiert (siehe Abbildung 2). Die Reihenfolge der Stimuli wurde hierbei für jede einzelne Versuchsperson randomisiert. Alle Fragen in den beiden Hauptphasen hatten ein sogenanntes *Forced-Input-Format* und mussten bearbeitet werden, um mit der Studie fortzufahren. Zudem wurde durch das Programm verhindert, dass nichtnumerische Antworten, Urteile außerhalb des vorab definierten Bereichs oder Urteile mit Nachkommastellen abgegeben wurden.

**Abbildung 2**

*Beispieldurchgang aus Phase 1 von Studie 1*




Nach Abgabe der 30 Urteile wurden die Versuchspersonen über den Ablauf von Phase 2 informiert. Hierfür erhielten sie präzise Instruktionen, die auf einer eigenen Seite angezeigt wurden. Die Instruktionen orientierten sich hierbei an vorherigen Interventionen zur Weisheit der Vielen in einer Person (dem Dialectical Bootstrapping, Herzog & Hertwig, 2009, 2014b), um Menschen nach einer möglichst unverzerrten Einschätzung der Richtung des wahren Werts beziehungsweise des eigenen Fehlers zu fragen (möglichst keine Response oder Selection Biases). Die zentrale Neuerung lag jedoch im letzten Schritt, bei

dem die Richtung des wahren Werts direkt abgefragt wurde und die Versuchspersonen zudem die Möglichkeit hatten keine der beiden Richtungen zu wählen:

> „In Phase 2 sollen Sie zu Ihren jeweiligen Schätzungen aus Phase 1 angeben, ob Sie den wahren Wert eher über, eher unter oder weder eher über noch eher unter der eigenen Schätzung vermuten.
>
> **Bitte folgen Sie bei jeder Aufgabe in Phase 2 immer den folgenden Schritten zur Beantwortung der Frage:**
>
> 1. Stellen Sie sich vor, Ihr Urteil weicht vom wahren Wert ab.
>
> 2. Überlegen Sie, was dafürspricht, dass der wahre Wert über Ihrer Schätzung liegt, und was dafürspricht, dass er unter Ihrer Schätzung liegt.
>
> 3. Entscheiden Sie sich für eine der beiden Richtungen auf Basis Ihrer Überlegungen oder geben Sie als dritte Option an, dass Sie den wahren Wert in keinen der beiden Richtungen eher vermuten."

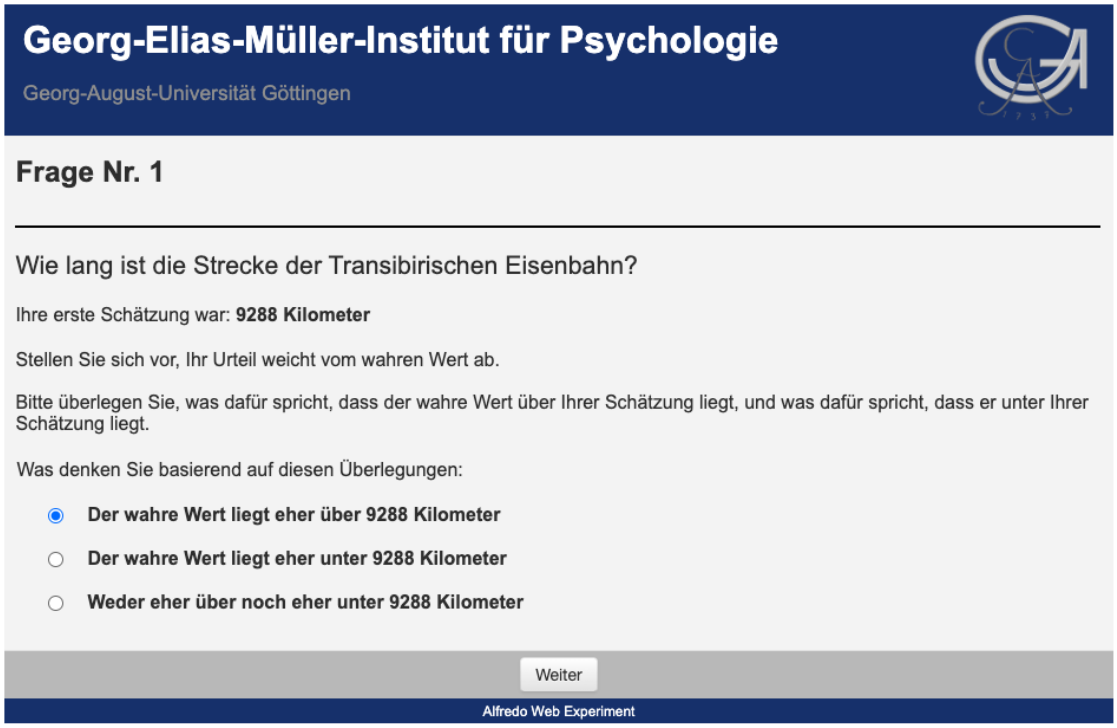
Um zu verhindern, dass diese Instruktionen ohne gründliches Lesen übersprungen wurden, wurde der „Weiter"-Knopf für 60 Sekunden gesperrt. Nach Ablauf der 60 Sekunden konnten die Versuchspersonen mit dem Klicken auf „Weiter" mit der zweiten Phase der Studie starten.

In Phase 2 wurden die Aufgaben und quantitativen Urteile aus Phase 1 in der gleichen Reihenfolge angezeigt. Aufgabe der Versuchspersonen war es, ihre Richtungsurteile abzugeben, wobei die detaillierten Instruktionen zur Abgabe dieser Urteile auf jeder Seite kurz zusammengefasst wurden (siehe Abbildung 3). Nach der Bearbeitung der letzten Aufgabe folgte ein kurzer Abschlussfragebogen, bei dem die Versuchspersonen optional Alter und Geschlecht angeben konnten, wobei sie auch explizit auf die Möglichkeit hingewiesen wurden, diese Fragen unbeantwortet zu lassen. Im Anschluss wurden sie auf einer eigenen Seite über Ziel und Zweck der Studie aufgeklärt, mit der Bitte, hierüber Stillschweigen zu bewahren. Schließlich wurden die Versuchspersonen darüber aufgeklärt, in wie vielen Fällen und in welcher Gesamthöhe sie einen Bonus erreicht haben. Zudem wurden sie zur Abrechnung und Dokumentierung gebeten, Namen und Geburtsdatum

anzugeben, wobei sie explizit darauf hingewiesen wurden, dass diese Daten mitsamt dem erfolgsabhängigen Bonus separat von den Experimentaldaten abgespeichert werden würden. Im letzten Schritt wurden die Versuchspersonen von der Versuchsleitung ausgezahlt.

**Abbildung 3**

*Beispieldurchgang aus Phase 2 von Studie 1*




### 4.1.4. Abhängige Maße

**Directional Judgment Rate (DJR).** Zur Untersuchung des Weisheitsbewusstseins wurden aus den Richtungsurteilen in Phase 2 verschiedene Kennwerte berechnet. Wichtig ist zunächst die Unterscheidung zwischen den Fällen, bei denen explizit eine Richtung gewählt wurde (hier auch *explizite* Richtungsurteile genannt), und den Fällen, bei denen keine der beiden Richtungen gewählt wurde („weder eher über noch eher unter dem Initialurteil“), da das Weisheitsbewusstsein nur für die Fälle definiert ist, in denen eine der beiden Richtungen gewählt wurde. Die erste abhängige Variable ist daher, wie häufig Menschen sich für ein Richtungsurteil anstelle der dritten Option entschieden, wobei es sich hierbei um ein Maß

handelt, welches zunächst nur deskriptiv ausgewertet wurde. Diese relative Häufigkeit wurde pro Versuchsperson berechnet, welche als *Directional Judgment Rate* (DJR) bezeichnet wird und nur Werte zwischen 0 (immer dritte Option) und 1 (immer erste oder zweite Option) annehmen kann. Bei der DJR beruht jeder Kennwert auf genau 30 Antworten. Bei allen folgenden Kennwerten wurden jedoch alle Beobachtungen ausgeschlossen, bei denen mit der dritten Option keine explizite Richtung gewählt wurde.

**Truth Detection Rate (TDR).** Die Operationalisierung des Weisheitsbewusstseins erfolgte über ein Maß, das als *Truth Detection Rate* (TDR) bezeichnet wird und zur Untersuchung des Weisheitsbewusstseins in alle Präregistrierungen aufgenommen wurde (siehe P1-P8). Für die Berechnung der TDR pro Person wurden zunächst alle Antworten ausgeschlossen, bei denen die dritte Option („weder noch“) gewählt wurde. Für alle restlichen Richtungsurteile wurde kodiert, ob diese ausgehend vom Initialurteil in die Richtung des wahren Werts zeigten, also bei einem wahren Wert über dem ersten Urteil die Option „über dem ersten Urteil“ und bei einem wahren Wert unter dem ersten Urteil die Option „unter dem ersten Urteil“ gewählt wurde. Alle diese Fälle wurden als richtiges Richtungsurteil mit 1 kodiert, und alle anderen Fälle mit 0 (wenn der wahre Wert in der anderen Richtung lag oder genau mit dem Initialurteil übereinstimmte). Schließlich wurden alle kodierten Werte für jede Versuchsperson gemittelt und es wurde pro Person eine TDR berechnet, die angibt, mit welcher relativen Häufigkeit die expliziten Richtungsurteile in der richtigen Richtung lagen.

**Crowd Detection Rate (CDR).** Als zweiten Kennwert zur Untersuchung des Weisheitsbewusstseins wurde berechnet, wie häufig die expliziten Richtungsurteile der Versuchspersonen in der Richtung des Populationsmittelwerts (der klassischen Weisheit der Vielen) lagen. Dieser Kennwert wird als *Crowd Detection Rate* (CDR) bezeichnet, wobei die Berechnung äquivalent zur TDR erfolgte. Der Unterschied zur TDR ist lediglich, dass alle Richtungsurteile nicht relativ zum wahren Wert, sondern relativ zum Populationsmittelwert mit 0 und 1 kodiert wurden. Zur Schätzung des Populationsmittelwertes wurden für jeden der 30 Stimuli alle zugehörigen 100 Initialurteile aus Phase 1 gemittelt und auf einen Wert ohne Nachkommastelle gerundet. Die Rundung erfolgte, um zu verhindern, dass die CDR im Vergleich zur TDR systematisch überschätzt wird. So konnten die Versuchspersonen in der Studie keine Urteile mit Nachkommastelle angeben und auch bei den wahren Werten

handelte es sich um Werte ohne Nachkommastelle. Ohne Rundung sollten die Populationsmittelwerte wesentlich seltener (praktisch nie) mit den ersten Urteilen übereinstimmen und damit sollten systematisch seltener Fälle auftreten, die bei expliziten Richtungsurteilen immer mit 0 (stimmt nicht mit der Richtung überein) kodiert werden. Alle Richtungsurteile in Richtung des zugehörigen Populationsmittelwerts wurden anschließend mit 1 und alle anderen mit 0 kodiert (unter Ausschluss aller „weder noch“-Urteile). Äquivalent zur TDR wurde mit dem Mittelwert aus dieser Kodierung pro Person die CDR berechnet, die die relative Häufigkeit bemisst, mit der die jeweilige Versuchsperson mit ihren Richtungsurteilen die Richtung des Populationsmittelwerts wählte.

#### 4.1.5. Ergebnisse & Diskussion

**Richtungsurteile.** Im ersten Schritt wurde explorativ untersucht, inwiefern sich Menschen überhaupt für ein Richtungsurteil entschieden, und hierfür die DJR für jede Versuchsperson berechnet. Mittelwerte und Standardabweichungen aller Kennwerte zum Weisheitsbewusstsein befinden sich für diese und alle weiteren Studien mit Richtungsurteilen in Tabelle 2 und sind dort zudem separat für alle Aufgabenkategorien gelistet. Über alle Aufgaben hinweg entschieden sich die Versuchspersonen in 80% der Urteile für eine der beiden Richtungen, was einer DJR von $M$ = .80 ($SD$ = .16) entspricht, wobei diese Zahl über die drei Aufgabenkategorien zwischen .76 und .84 schwankte (siehe Tabelle 2). Nur drei Personen (3%) entschieden sich insgesamt in weniger als 50% der Fälle für eine der beiden Richtungen und keine Person tat dies in weniger als 30% der Fälle, obwohl sie jederzeit die Option hatten, keine Richtung zu wählen. Die Wahl eines Richtungsurteils war somit eher die Regel als die Ausnahme. Auch wenn es sich hierbei um einen reinen deskriptiven Befund handelt, ist dieser konsistent mit den Annahmen des GSM, nach welchem Menschen zur Bewertung eines Urteiles im zweiten Schritt häufig neue Informationen heranziehen und dadurch bewusst eine der beiden Richtungen wählen. Die Grundvoraussetzung für das Weisheitsbewusstsein war somit in der deutlichen Mehrzahl der Fälle gegeben.

**Tabelle 2**

*Übersicht von DJR, TDR, & CDR für alle Experimentalbedingungen und Aufgabenkategorien*

| Stichprobe/Aufgabenkategorie | | DJR | | TDR | | CDR | |
|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | *M* (vs. .50) | *SD* | *M* (vs. .50) | *SD* |
| Studie 1 | | | | | | | |
| Historisch (Globalisierung) | 100 | .79 | .20 | .57** | .21 | .58*** | .21 |
| Unten begrenzt | 100 | .84 | .17 | .58*** | .20 | .66*** | .19 |
| Prozent | 100 | .76 | .19 | .49 | .19 | .54 | .23 |
| Gesamt[a] | 100 | .80 | .16 | .55††† | .11 | .59*** | .11 |
| Studie 2 (fremde Urteile) | | | | | | | |
| Historisch (Globalisierung) | 75 | .79 | .20 | .69*** | .14 | .66*** | .15 |
| Unten begrenzt | 75 | .84 | .14 | .69*** | .14 | .68*** | .17 |
| Prozent | 75 | .84 | .15 | .61*** | .17 | .67*** | .16 |
| Gesamt[a] | 75 | .82 | .14 | .66††† | .09 | .67*** | .09 |
| Studie 2 (eigene Urteile) | | | | | | | |
| Historisch (Globalisierung) | 73 | .74 | .22 | .61*** | .15 | .61*** | .19 |
| Unten begrenzt | 74 | .78 | .19 | .62*** | .16 | .71*** | .17 |
| Prozent | 74 | .70 | .23 | .55** | .15 | .58*** | .17 |
| Gesamt[a] | 74 | .74 | .19 | .60††† | .10 | .64*** | .10 |
| Studie 3 | | | | | | | |
| Kalorien | 296 | .67 | .21 | .57*** | .20 | .59*** | .20 |
| Studie 4 (EG + KI) | | | | | | | |
| Alter | 168 | .65 | .27 | .53 | .19 | .60*** | .20 |
| Studie 4 (EG) | | | | | | | |
| Alter[a] | 175 | .70 | .27 | .52 | .17 | .55*** | .17 |
| Studie 5 (EG) | | | | | | | |
| Historisch (Globalisierung) | 140 | .63 | .23 | .60*** | .20 | .59*** | .20 |
| Unten begrenzt | 139 | .68 | .22 | .59*** | .19 | .63*** | .18 |
| Gesamt[a] | 140 | .66 | .21 | .59††† | .14 | .61*** | .15 |
| Studie 6 | | | | | | | |
| Historisch (Globalisierung) | 138 | .68 | .24 | .61*** | .17 | .63*** | .18 |
| Unten begrenzt | 138 | .70 | .24 | .59*** | .16 | .65*** | .18 |
| Gesamt[a] | 138 | .69 | .22 | .60††† | .12 | .64*** | .12 |
| Studie 7 | | | | | | | |
| Historisch (Musik) | 73 | .59 | .27 | .60*** | .14 | .62*** | .14 |
| Studie 8 (R+) | | | | | | | |
| Kalorien[a] | 90 | .68 | .28 | .65††† | .23 | .67*** | .22 |
| Studie 8 (R-) | | | | | | | |
| Kalorien[a] | 90 | .55 | .27 | .56†† | .24 | .57** | .24 |

*Anmerkung*. DJR = Directional Judgment Rate. TDR = Truth Detection Rate. CDR = Crowd Detection Rate. EG = Experimentalgruppe. EG + KI = Experimentalgruppe + Abfrage Konfidenzintervall. R+ = Bedingung mit Ratschlägen. R- = Bedingung ohne Ratschläge.

[a] Einseitiger Test präregistriert (zweiseitige Tests in allen anderen Fällen).

† $p < .05$, einseitig. †† $p < .01$, einseitig. ††† $p < .001$, einseitig. * $p < .05$, zweiseitig. ** $p < .01$, zweiseitig. *** $p < .001$, zweiseitig.

**Weisheitsbewusstsein.** Im zweiten Schritt wurde für die Richtungsurteile, bei denen explizit eine der beiden Richtungen gewählt wurde, getestet, ob diese überzufällig häufig in der richtigen Richtung lagen (H1). Hierfür wurde für alle Versuchspersonen die TDR berechnet und in einem präregistrierten Einstichproben-*t*-Test einseitig getestet, ob diese im Schnitt über .50 lag (siehe P1). Bei dem Referenzwert von .50 handelt es sich um die TDR, die man erwarten würde, wenn Menschen sich zufällig für eine der Richtungen entscheiden (50%). Hierbei zeigte sich, dass die TDRs der Versuchspersonen mit $M = 0.55$ ($SD = 0.11$) im Schnitt über 50% lagen, $t(99) = 4.30$, $p < .001$ (einseitig; P1), $d = 0.43$. Hypothese 1 konnte damit angenommen werden. Die Entscheidungsträger:innen in Studie 1 waren überzufällig häufig in der Lage, mit ihren Richtungsurteilen die Richtung des wahren Werts relativ zum eigenen Urteil einzuschätzen; sie hatten also ein entsprechendes Weisheitsbewusstsein.[6] In Tabelle 2 sind die TDRs für alle drei Aufgabenkategorien separat gelistet. Hierbei zeigt sich, dass der entsprechende Effekt lediglich bei den Prozentschätzungen nicht auftrat und hier die TDRs mit $M = .49$ deskriptiv kaum von .50 abwichen. Bei den historischen Schätzungen ($M = .57$) und den nach unten begrenzten Schätzungen ($M = .58$) fielen sie entsprechend höher aus. Das Weisheitsbewusstsein ließ sich somit zumindest in dieser Studie nur eingeschränkt auf unterschiedliche Aufgabenkategorien generalisieren. Ein Problem dieser Subgruppenanalysen ist jedoch die – im Vergleich zur Analyse über alle Aufgaben hinweg – geringere statistische Power aufgrund der größeren Varianz der TDRs.

Um zu testen, ob die Richtungsurteile auch überzufällig häufig in der Richtung des Populationsmittelwerts lagen (H2a), wurde für alle Versuchspersonen die CDR berechnet und genau wie die TDR gegen .50 getestet, wobei ein zweiseitiger Test verwendet wurde, da diese Analyse nicht präregistriert war. Im Einklang mit dem GSM zeigte sich, dass die CDRs mit $M = .59$ ($SD = .11$) im Schnitt über 50% lagen, $t(99) = 8.16$, $p < .001$ (zweiseitig), $d = 0.82$. Somit zeigte sich in Studie 1 Evidenz für H2a. Die Versuchspersonen schätzen mit ihren

[6] Grundsätzlich besteht die Möglichkeit, dass dieser Effekt artifiziell dadurch zustande kommt, dass Menschen trotz der Bonuszahlungen bei ihren Initialurteilen extreme Fehler machen (bspw. durch Tippfehler), die sie im zweiten Schritt relativ einfach erkennen können. Um dies auszuschließen, wurden für diese und alle weiteren Studien die gleiche Analyse wiederholt und hierbei pro Stimulus jeweils die 10% der Urteile mit den extremsten Fehlern ausgeschlossen. Zur Berechnung der Fehler wurde bei allen Stimuli der absoluten Fehler verwendet und die absolute Differenz zwischen wahrem Wert und Urteil gebildet. In keiner der Analysen zeigten sich beim Ausschluss größere qualitative Unterschiede zwischen dieser und der im Haupttext berichteten Analysen. Zur Übersicht befindet sich in Anhang B eine Version von Tabelle 2, bei der für jede Studie dieser Ausschluss vorgenommen wurde.

expliziten Richtungsurteilen überzufällig häufig die Richtung der Weisheit der Vielen ein. Deskriptiv war dies für alle Aufgabenkategorien der Fall (siehe Tabelle 2), wobei sich erneut die CDRs für die Prozentschätzungen nicht statistisch signifikant von .50 unterschieden. Auch dieser Befund kann somit nur eingeschränkt auf die unterschiedlichen Aufgabenkategorien generalisiert werden.

Wie Eingangs erläutert ist für das Weisheitsbewusstsein entscheidend, dass wahrer Wert und Populationsmittelwert relativ häufig in der gleichen Richtung liegen. Dies war in dieser Studie der Fall. Bei 73% aller expliziten Richtungsurteile (ohne „weder noch“-Urteile) lagen beide Referenzwerte in der gleichen Richtung (konsistente Fälle), während bei 24% die beiden Referenzwerte in unterschiedliche Richtungen zeigten (inkonsistente Fälle). Um zu testen, ob Menschen mit ihren Richtungsurteilen tatsächlich eher die Weisheit der Vielen als den wahren Wert detektieren, wurden für beide Fälle separat die TDRs beziehungsweise CDRs berechnet. In weiteren 3% der expliziten Richtungsurteile stimmte das Initialurteil entweder genau mit dem wahren Wert oder genau mit dem Populationsmittelwert überein. Diese Fälle wurden in den folgenden Analysen ausgeschlossen, da sich nicht eindeutig definieren lässt, ob hier die Referenzwerte in der gleichen Richtung oder unterschiedlichen Richtungen lagen. Für die anderen Fälle sind TDR und CDR vollständig voneinander abhängig, da TDR = CDR, wenn beide Referenzwerte in die gleiche Richtung zeigen, und TDR = 1 – CDR, wenn beide in unterschiedliche Richtungen zeigen. Konsistent mit den vorherigen Analysen zeigte sich bei den konsistenten Fällen, dass die Richtungsurteile mit einer TDR/CDR von $M$ = .60 ($SD$ = .13) überzufällig häufig in der Richtung beider Referenzwerte lagen, $t(99) = 7.98$, $p < .001$ (zweiseitig), $d = 0.80$. Um zu testen, ob hierbei eher die Weisheit der Vielen detektiert wird, sind jedoch die inkonsistenten Fälle ausschlaggebend. Hier zeigten die Richtungsurteile eher in die Richtung der Populationsmittelwerte. Im Einstichproben-$t$-Test lagen die CDRs mit $M$ = .59 ($SD$ = .22) im Schnitt über .50, beziehungsweise die TDRs mit $M$ = .41 im Schnitt unter .50, $t(98) = 4.26$, $p < .001$ (zweiseitig), $d = 0.43$. Somit zeigte sich auch Evidenz für H2b. Die Versuchspersonen tendierten damit unabhängig von der Position des wahren Werts zum Populationsmittelwert und waren nicht mehr in der Lage die Richtung des wahren Werts einzuschätzen, wenn der Populationsmittelwert in die andere Richtung zeigte. Dieser Befund steht im Einklang mit dem GSM, nach dem Menschen zur Abgabe der Richtungsurteile auf

einen geteilten Informationspool zugreifen, dessen bestmögliches Urteil mit der Weisheit der Vielen approximiert werden kann.

Für die einzelnen Aufgabenkategorien fiel dieses Muster qualitativ ähnlich aus (siehe Tabelle 3). Ausnahmen sind zum einen wieder die Prozentschätzungen, bei denen Menschen bei gleicher Richtung der Referenzwerte nicht systematisch in diese Richtung tendierten, und die historischen Schätzungen, bei denen Menschen nur bei den konsistenten Fällen systematisch in die Richtung der Referenzwerte tendierten. Die statistischen Inferenzen für diese Subgruppenanalysen sind jedoch unter Vorbehalt zu betrachten, da gerade die Tests der inkonsistenten Fälle teilweise auf einer sehr geringen Anzahl von Beobachtungen beruhen. Grund dafür ist, dass Populationsmittelwert und wahrer Wert nur vergleichsweise selten in unterschiedliche Richtungen zeigten und die Standardabweichungen für TDR und CDR für die jeweilige Aufgabenkategorie entsprechend hoch ausfielen. Teilweise ließen sich für einzelne Versuchspersonen gar keine entsprechenden Kennwerte bestimmen, da es keine Fälle gab, um TDR und CDR zu bestimmen (siehe zweite Spalte in Tabelle 3). Insgesamt zeigen die Ergebnisse von Studie 1 jedoch Evidenz für das Weisheitsbewusstsein und die Annahmen des GSM. Es stellt sich jedoch die unmittelbare Frage, ob sich diese Ergebnisse replizieren lassen und ob sich bei einer größeren Menge von Stimuli auch Evidenz für das Weisheitsbewusstsein für alle Aufgabenkategorien zeigt.

**Tabelle 3**

*Übersicht von TDR & CDR nach Richtung von wahrem Wert und Populationsmittelwert für alle Experimentalgruppen und Aufgabenkategorien*

| Stichprobe/Aufgabenkategorie | *n* (T = C/T ≠ C) | % T = C | % T ≠ C | TDR/CDR (T = C) | | TDR/CDR (T ≠ C) | |
|---|---|---|---|---|---|---|---|
| | | | | *M* (vs. .50) | *SD* | *M* (vs. .50) | *SD* |
| Studie 1 | | | | | | | |
| Historisch (Globalisierung) | 100/81 | 72% | 23% | .62*** | .25 | .54/.46 | .40 |
| Unten begrenzt | 99/94 | 71% | 28% | .67*** | .23 | .36/.64*** | .36 |
| Prozent | 100/85 | 76% | 20% | .54 | .25 | .36/.64*** | .38 |
| Gesamt | 100/99 | 73% | 24% | .60*** | .13 | .41/.59*** | .22 |
| Studie 2 (fremde Urteile) | | | | | | | |
| Historisch (Globalisierung) | 75/73 | 66% | 30% | .77*** | .16 | .54/.46 | .29 |
| Unten begrenzt | 75/74 | 70% | 29% | .75*** | .18 | .51/.49 | .33 |
| Prozent | 75/65 | 78% | 18% | .69*** | .18 | .38/.62* | .39 |
| Gesamt | 75/75 | 71% | 26% | .74*** | .10 | .48/.52 | .18 |
| Studie 2 (eigene Urteile) | | | | | | | |
| Historisch (Globalisierung) | 73/72 | 66% | 30% | .67*** | .22 | .51/.49 | .30 |
| Unten begrenzt | 74/72 | 70% | 29% | .74*** | .19 | .38/.62** | .31 |
| Prozent | 74/61 | 78% | 18% | .59*** | .19 | .39/.61* | .37 |
| Gesamt | 74/74 | 71% | 26% | .67*** | .13 | .42/.58*** | .16 |
| Studie 3 | | | | | | | |
| Kalorien | 296/216 | 78% | 21% | .61*** | .23 | .46/.54 | .36 |
| Studie 4 (EG + KI) | | | | | | | |
| Alter | 168/149 | 68% | 21% | .62*** | .24 | .35/.65*** | .33 |
| Studie 4 (EG) | | | | | | | |
| Alter | 173/155 | 67% | 21% | .61*** | .20 | .43/.57** | .33 |
| Studie 5 (EG) | | | | | | | |
| Historisch (Globalisierung) | 140/122 | 73% | 24% | .65*** | .22 | .54/.46 | .37 |
| Unten begrenzt | 139/128 | 71% | 28% | .65*** | .21 | .45/.55 | .36 |
| Gesamt | 140/136 | 72% | 26% | .65*** | .17 | .48/.52 | .29 |
| Studie 6 | | | | | | | |
| Historisch (Globalisierung) | 137/113 | 77% | 21% | .66*** | .20 | .43/.57* | .35 |
| Unten begrenzt | 138/119 | 77% | 21% | .66*** | .19 | .38/.62*** | .36 |
| Gesamt | 138/134 | 77% | 21% | .66*** | .13 | .41/.59*** | .29 |
| Studie 7 | | | | | | | |
| Historisch (Musik) | 73/69 | 81% | 16% | .64*** | .16 | .44/.56 | .27 |
| Studie 8 (R+) | | | | | | | |
| Kalorien | 90/48 | 88% | 12% | .67*** | .24 | .48/.52 | .40 |
| Studie 8 (R-) | | | | | | | |
| Kalorien | 87/55 | 86% | 14% | .59** | .26 | .48/.52 | .43 |

*Anmerkung*. TDR = Truth Detection Rate. CDR = Crowd Detection Rate. EG = Experimentalgruppe. „T = C“ = Wahrer Wert und Populationsmittelwert liegen in der gleichen Richtung. „T ≠ C“ = Wahrer Wert und Populationsmittelwert liegen in unterschiedlichen Richtungen. EG + KI = Experimentalgruppe + Abfrage Konfidenzintervall. R+ = Bedingung mit Ratschlägen. R- = Bedingung ohne Ratschläge.

* p < .05, zweiseitig. ** p < .01, zweiseitig. *** p < .001, zweiseitig.

### 4.2. Studie 2

Ein Ziel von Studie 2 war, die Befunde der ersten Studie zu replizieren. Die Methodik von Studie 2 unterschied sich deshalb in nur wenigen Aspekten von Studie 1. Ein größerer Unterschied bestand darin, dass die Anzahl der Stimuli von 30 auf 45 erhöht wurde (5 neue Stimuli pro Aufgabenkategorie). Hierdurch lassen sich die Kennwerte zur Untersuchung des Weisheitsbewusstseins pro Person reliabler schätzen, was insbesondere für die Subgruppenanalysen relevant ist, die in der vorherigen Studie teilweise auf nur wenigen Beobachtungen beruhten.

Das zweite Ziel bestand darin, eine Vergleichsgruppe für das Weisheitsbewusstsein zu schaffen, um die Grenzen des Phänomens besser einzuschätzen und weitere Annahmen des GSM zu testen. Hierfür wurde eine zusätzliche Bedingung erhoben, bei der Menschen die Richtung des wahren Werts relativ zu dem Urteil einer anderen Person einschätzten. So zeigte sich mit Studie 1 erste Evidenz dafür, dass eine wesentliche Limitation des Weisheitsbewusstseins durch den geteilten Informationspool entsteht und das Weisheitsbewusstsein durch die relative Weisheit der Vielen limitiert ist (der Übereinstimmung der Richtung des Populationsmittelwerts und des wahren Werts). Nach dem GSM entstehen jedoch zusätzliche Limitationen durch den privaten Informationspool, auf den Menschen zur Bewertung ihrer eigenen Urteile erneut zugreifen können und der nur eine Teilmenge von Informationen des Gesamtpools enthält. Diese Teilmenge könnte vergleichsweise klein sein, wodurch Menschen bei der Bewertung zwar einige neue, jedoch auch primär alte Informationen erneut ziehen, um die Richtung ihrer Fehler zu bewerten. In meinem Beispiel in Abschnitt 2 konnte die Person aus Bayern beispielsweise nur eine von drei Informationen austauschen (Bamberg statt Coburg). Bei der Bewertung fremder Urteile lässt sich nach dem GSM keine solche Limitation erwarten, da für die Bewertung des Urteiles einer anderen Person auf einen neuen (den eigenen) privaten Informationspool zugegriffen werden kann und sich die Informationsstichproben *zwischen* Personen stärker unterscheiden sollten als *innerhalb* einer Person. Eine andere (fremde) Person könnte also eher völlig neue Informationen ziehen, um die gleiche Bewertung vorzunehmen und damit eine bessere Einschätzung der Richtung des wahren Werts vornehmen. Auch für die neue Bedingung lässt sich damit die Vorhersage ableiten, dass Menschen (ausgehend von fremden Urteilen) in der Lage sind, die Richtung des wahren Werts mit ihren expliziten

Richtungsurteilen überzufällig häufig richtig einzuschätzen. Darüber hinaus sollte dies häufiger der Fall sein als bei Einschätzungen relativ zu einem eigenen Urteil:

> **Hypothese 3 (H3; P2[7]):** Menschen sind in der Lage, die Richtung des wahren Wertes *ausgehend von einem fremden Urteil* überzufällig häufig richtig einzuschätzen (d.h., in mehr als 50% der Fälle), sofern sie sich für eine der beiden Richtungen entscheiden.

> **Hypothese 4 (H4; P2):** Menschen sollten mit ihren Richtungsurteilen häufiger richtig liegen, wenn sie die Richtung des wahren Werts relativ zu einem fremden (vs. eigenen) Urteil einschätzen (unter Ausschluss der Fälle, in denen keine der beiden Richtungen gewählt wird).

Zusätzlich lässt sich die Hypothese ableiten, dass Menschen bei der Bewertung fremder Urteile auch häufiger ein explizites Richtungsurteil wählen sollten als bei der Bewertung eigener Urteile. Der Grund hierfür liegt wieder darin, dass bei der Bewertung fremder Urteile seltener noch einmal genau die gleichen Informationen vorliegen sollten als bei der Bewertung eigener Urteile:

> **Hypothese 5 (H5; nicht präregistriert):** Menschen sollten sich häufiger für eine der beiden Richtungen anstelle der „weder noch"-Option entscheiden, wenn sie die Richtung des wahren Werts relativ zu einem fremden (vs. eigenen) Urteil einschätzen.

Für einen kritischen Test des GSM lassen sich zudem für die neue Bedingung äquivalente Hypothesen bezüglich der Richtungsurteile relativ zum Populationsmittelwert ableiten, da auch die Richtungsurteile bei der Bewertung fremder Urteile auf den gleichen geteilten Informationspool zurückgehen sollten und eher den Populationsmittelwert detektieren sollten als den wahren Wert (siehe Abbildung 1). Da jedoch anteilig mehr neue Informationen aus dem geteilten Pool gezogen werden, sollten Menschen bei fremden Urteilen besser in der Lage sein die zentrale Tendenz des geteilten Informationspools zu

[7] In P2 wurde ein einseitiger Test für H3 und ein zweiseitiger Test für H4 präregistriert, jedoch keine expliziten Hypothesen für die Effekte aufgestellt.

detektieren (den Populationsmittelwert), wodurch sie auch besser darin sind, den wahren Wert zu detektieren (siehe H4):

**Hypothese 6a (H6a; nicht präregistriert):** Die Richtungsurteile von Menschen liegen *ausgehend von fremden Urteilen* überzufällig häufig (in mehr als 50% der Fälle) in der Richtung des Populationsmittelwerts (der Weisheit der Vielen), sofern sich Menschen für ein Richtungsurteil entscheiden.

**Hypothese 6b (H6b; nicht präregistriert):** Die Richtungsurteile von Menschen liegen *ausgehend von fremden Urteilen* überzufällig häufig (in mehr als 50% der Fälle) in der Richtung des Populationsmittelwerts, sofern sich Menschen für ein Richtungsurteil entscheiden und wahrer Wert und Populationsmittelwert in unterschiedlichen Richtungen liegen.

**Hypothese 7a (H7a; nicht präregistriert):** Die Richtungsurteile von Menschen liegen ausgehend von fremden Urteilen häufiger in der Richtung des Populationsmittelwerts als ausgehend von eigenen Urteilen (unter Ausschluss der Fälle, in denen keine der beiden Richtungen gewählt wird).

**Hypothese 7b (H7b; nicht präregistriert):** Die Richtungsurteile von Menschen liegen ausgehend von fremden Urteilen häufiger in der Richtung des Populationsmittelwerts als ausgehend von eigenen Urteilen (unter Ausschluss der Fälle, in denen keine der beiden Richtungen gewählt wird und unter Ausschluss der Fälle, in denen wahrer Wert und Populationsmittelwert in gleicher Richtung liegen).

### 4.2.1. Design & Stichprobe

Bei Studie 2 handelte es sich um ein einfaktorielles Design mit einem Zwischensubjektfaktor mit zwei Ausprägung, mit dem manipuliert wurde, ob die Versuchspersonen die Richtung des wahren Werts relativ zu ihrem eigenen (wie in Studie 1) oder einem fremden Urteilen einschätzten. Hierfür wurden über ein sogenanntes *Yoking* Paare von zwei Versuchspersonen aus je einer der beiden Bedingungen gebildet, bei der eine Versuchsperson die Aufgaben wie in Studie 1 bearbeitete und erst quantitative Urteile abgab (Phase 1) und dann relativ zu diesen die Richtung des wahren Werts einschätzte

(Phase 2). Für die zweite Versuchsperson des jeweiligen Paares wurde Phase 1 der Studie übersprungen, und diese bekam direkt die Urteile der ersten Person präsentiert und sollte die Richtung des wahren Werts relativ zu diesen einschätzen.

Zur Bestimmung der Stichprobengröße wurden der Test von H1 und der Paarvergleich zur Testung von H4 zugrunde gelegt (siehe P2). Für H1 konnte in dieser Studie der beobachtete Effekt aus Studie 1 zugrunde gelegt werden ($d$ = 0.43). Für H4 wurde ein mittlerer bis starker Effekt angenommen ($d$ = 0.50; Cohen, 1988), da analog zum Weisheitsbewusstsein auch die Weisheit der Vielen bei zwei unterschiedlichen Personen deutlich stärker ausfällt als die Weisheit der Vielen in einer Person (z.B. Vul & Pashler, 2008). Ziel war es, beide Hypothesen mit einer statistischen Power von mindestens $1 - \beta = .85$ zu testen, weshalb eine Stichprobe von $N$ = 150 (75 pro Zelle) gewählt wurde, die bei den angenommenen Effektgrößen eine Power von $1 - \beta > .95$ für H1 bei einseitiger und $1 - \beta = .86$ für H4 bei zweiseitiger Testung ergab. Zudem wurde in die Präregistrierung für diese Studie explizit das Ausschlusskriterium aufgenommen, dass die Versuchspersonen die Studie komplett durchführen und beenden müssen, um in der Datenanalyse berücksichtigt zu werden (siehe P2). Sofern es zu einem Ausschluss kommt, würde die ausgeschlossene Versuchsperson durch eine neue ersetzt werden. Dieses Kriterium kam in einem Fall zur Anwendung, da eine Person die Studie aufgrund technischer Probleme nicht beenden konnte und auch entsprechend beim Yoking nicht berücksichtigt wurde. Entsprechend wurde eine Person mehr erhoben als ursprünglich geplant.

Im finalen Datensatz befanden sich die Antworten von $N$ = 150 Versuchspersonen (55 Männer, 92 Frauen und drei nichtbinär/divers) im Durchschnittsalter von $M$ = 24.79 ($Mdn$ = 23, $SD$ = 6.52) Jahren, bei denen es sich überwiegend um Studierende handelte. Aufgrund der höheren Anzahl von Stimuli und der damit verbundenen längeren Bearbeitungszeit erhielten die Versuchspersonen in dieser Studie eine Grundvergütung von 4€, wobei auch eine halbe Versuchspersonenstunde als alternative Kompensation abgezeichnet wurde. Zudem erhielten nur die Versuchspersonen in der Bedingung mit ihren eigenen Urteilen als Referenzurteile einen leistungsabhängigen Bonus von bis zu 4.50€ ($M$ = 1.09, $SD$ = 0.32), während die Versuchspersonen in der anderen Bedingung als Ausgleichszahlung einen zusätzlichen Euro erhielten, der sich nach dem durchschnittlichen erwarteten Bonus der anderen Bedingung richtete. Dieses Prozedere war notwendig, um sicherzustellen, dass die

Versuchspersonen in der neuen Bedingung eine möglichst äquivalente Bezahlung erhielten. Die Versuchspersonen wurden in dieser Bedingung darüber aufgeklärt, dass die Versuchspersonen in der anderen Bedingung Boni für ihre Schätzung erhielten. Durch die feste Zusatzzahlung sollte sichergestellt werden, dass es hierdurch nicht zu Frustration oder anderen unerwünschten Effekten kommt.

### 4.2.2. Methode

Methodik und Ablauf von Studie 2 entsprachen mit wenigen Ausnahmen denen von Studie 1. Der erste Unterschied bestand darin, dass die Versuchspersonen nach Abgabe ihres Einverständnisses im Hintergrund über einen automatisierten Prozess den zwei unterschiedlichen Versuchsbedingungen zugeordnet wurden. Hierfür wurde für jede Person durch die Experimentalsoftware geprüft, ob bereits eine Versuchsperson das Experiment in der Bedingung mit eigenen Urteilen beendet hatte und dieser Person zudem noch keine andere Versuchsperson als Partner:in zugewiesen worden war. War diese Suche erfolgreich, so wurde ein neues Paar gebildet, und die Versuchsperson erhielt die Urteile der ersten Person als Referenzurteile. War die Suche nicht erfolgreich, so wurde die Versuchsperson der Bedingung mit eigenen Urteilen zugeordnet und bekam zu einem späteren Zeitpunkt eine andere Versuchsperson in der anderen Bedingung als Partner:in zugeordnet. Die Zuteilung zu den Versuchsbedingungen war entsprechend durch den Zeitpunkt der Teilnahme determiniert, wobei im Laufe der Erhebung immer abwechselnd Versuchspersonen in der einen und der anderen Bedingung erhoben wurden und daher angenommen wird, dass dieser Prozess im Wesentlichen nicht von einer echten Randomisierung zu unterscheiden ist.

Ein weiterer Unterschied bestand darin, dass alle Versuchspersonen für einen reibungsloseren Ablauf nach Abgabe ihres Einverständnisses eine Seite mit dem Titel „Wichtige Informationen zur Bearbeitung des Experiments“ erhielten, bei der sie explizit aufgefordert wurden, nur die Schaltflächen des Experiments zu nutzen, um sich durch das Experiment zu navigieren.[8] Im weiteren Verlauf entsprach der Ablauf der Bedingung mit

[8] Ziel war es, hiermit die Risiken technischer Störungen zu minimieren, die vereinzelt aufgetreten waren, deren Ursprung jedoch bis zuletzt nicht vollständig geklärt werden konnte und die aufgrund einer kompletten Überarbeitung der Experimentalsoftware nicht weiterverfolgt werden mussten. Zumindest in Studie 2 lag die Zahl technischer Störungen unter dem Niveau von Studie 1. In den späteren Laborstudien (Studie 5-7) ergaben

eigenen Urteilen dem Ablauf von Studie 1 mit der bereits erwähnten Ausnahme, dass die Anzahl der Fragen zum Thema Globalisierung auf 45 erhöht wurde. Zudem wurden die Versuchspersonen in dieser Studie im Rahmen der Instruktionen über die Bonuszahlungen auch über den durchschnittlichen erwarteten Bonus informiert, der auf Basis von Studie 1 gerundet auf 1€ geschätzt wurde (hochgerechnet auf 45 Fragen) und der auch Basis der zusätzlichen Zahlungen in der zweiten Bedingung war.

In der Bedingung mit fremden Urteilen erhielten die Versuchspersonen angepasste Instruktionen zur Bearbeitung des Experiments und starteten zudem direkt mit der Phase 2 des Experiments. In dieser Bedingung erhielten die Versuchspersonen die Information, dass im Rahmen dieser Studie eine andere Person 45 Schätzaufgaben zum Thema Globalisierung beantwortet habe und es nunmehr ihre Aufgabe sei, vertiefende Fragen zu diesen Aufgaben zu beantworten. Zudem wurden sie darüber informiert, dass die andere Versuchsperson 10 Cent für besonders akkurate Schätzungen erhalten habe und sie, da sie keine solche Möglichkeit erhielten, 1€ als Ausgleichszahlung erhalten würden. Zudem erfolgte auch der Hinweis, dass das Verwenden von Hilfsmitteln nicht erlaubt sei. Im Unterschied zu der anderen Bedingung erfolgten keine vertiefenden Informationen zur Bonuszahlung, wobei die Versuchspersonen auch über die alternative Vergütung über eine halbe Versuchspersonenstunde informiert wurden, die sie unabhängig von zusätzlichen Zahlungen erhalten könnten. Auf der letzten Seite vor dem Start der Hautphase erhielten die Versuchspersonen den Hinweis, sich bei Fragen oder Problemen mit einem Handzeichen bei der Versuchsleitung melden zu können und mit „Zurück“ die Instruktionen noch ein weiteres Mal lesen zu können oder mit „Weiter“ mit der Studie zu starten.

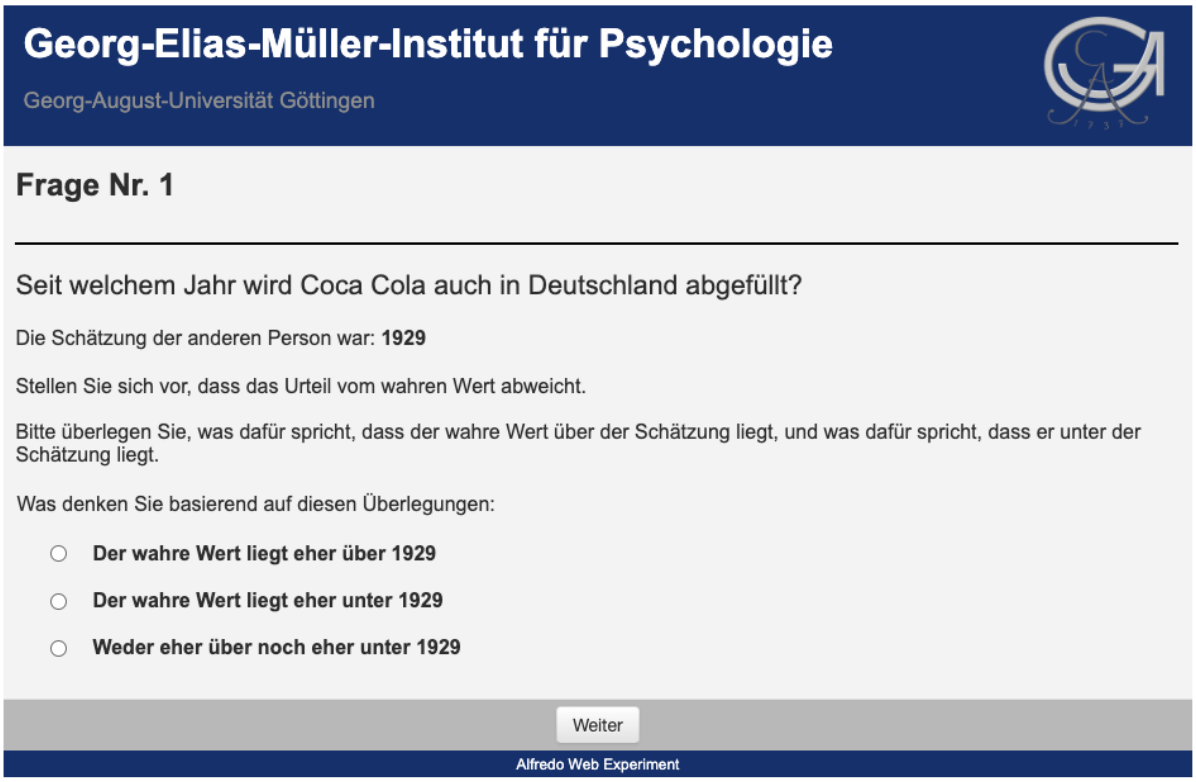
Mit dem Start der Hauptphase erhielten die Versuchspersonen in der neuen Bedingung die detaillierten Instruktionen zur Abgabe von Richtungsurteilen, die für mindestens 60 Sekunden angezeigt wurden. Der einzige Unterschied bestand darin, dass diese Urteile relativ zu dem Urteil der anderen Person abgegeben wurden (siehe Abbildung 4). Hierbei wurden die Aufgaben in genau der gleichen Reihenfolge präsentiert, in der die Versuchsperson aus der anderen Bedingung die Aufgaben bearbeitet hatte. Entsprechend wurde auch die Randomisierung der Aufgabenreihenfolge über das Yoking konstant

sich jedoch trotz der neuen Seite immer wieder technische Störungen, die jedoch nicht wesentlich häufiger auftraten als in Studie 1 und deren Ausschluss in den folgenden Präregistrierungen berücksichtigt wurde.

gehalten. Nach der Bearbeitung konnten die Versuchspersonen auch in dieser Bedingung Alter und Geschlecht angeben, erhielten ein kurzes schriftliches Debriefing und wurden ausgezahlt.

**Abbildung 4**

*Beispieldurchgang aus Studie 2 mit fremden Urteilen als Referenzurteile*



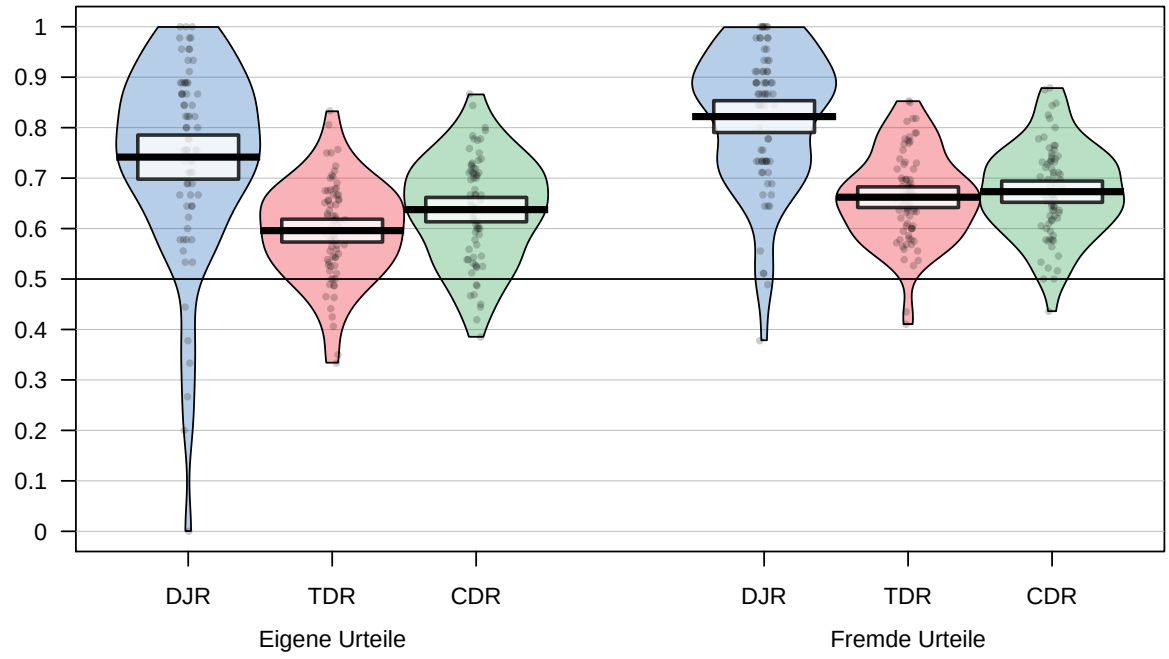

### 4.2.3. Ergebnisse & Diskussion

**Richtungsurteile**. Auch in Studie 2 wählten Menschen mit einer DJR von *M* = .74 (*SD* = .19) in der deutlichen Mehrzahl der Fälle ein Richtungsurteil, wenn sie ihre eigenen Urteile bewerteten. Dieser deskriptive Befund aus der ersten Studie konnte repliziert werden. Eine Versuchsperson in der Bedingung mit eigenen Urteilen entschied sich jedoch in keinem der 45 Durchgänge für ein Richtungsurteil und wurde entsprechend in den folgenden Analysen zum Weisheitsbewusstsein nicht berücksichtigt. Zum Test von H5 wurde zudem die DJR der neuen Versuchsbedingung berechnet und mit einem Welch's *t*-Test für unabhängige

Stichproben mit der DJR der ersten Bedingung verglichen[9]. Hierbei zeigte sich, dass die DJR mit *M* = .82 (*SD* = .14) bei fremden Urteilen höher ausfiel, *t*(134.62) = 2.97, *p* = .004 (zweiseitig), *d* = 0.49. H5 konnte somit angenommen werden. Die Versuchspersonen entschieden sich bei fremden Urteilen häufiger für eine der beiden Richtungen als bei eigenen Urteilen (siehe auch Abbildung 5).

**Abbildung 5**

*Vergleich von DJR, TDR, und CDR der Bedingungen von Studie 2*




*Anmerkung*. Abbildung der mittleren DJRs, TDRs und CDRs (schwarze Linien) mit 95% Konfidenzintervalle (weiße Rechtecke), Verteilungsdichte („Geigenplots“) und Darstellung einzelner Versuchspersonen (schwarze Punkte). DJR = Directional Judgment Rate. TDR = Truth Detection Rate. CDR = Crowd Detection Rate.

**Weisheitsbewusstsein.** Für die Analysen des Weisheitsbewusstseins wurden die gleichen Kennwerte wie in Studie 1 berechnet, wobei die Kennwerte in der neuen Versuchsbedingung immer ausgehend von den (fremden) Urteilen der anderen

[9] Der Welch’s *t*-Test korrigiert für ungleiche Varianzen, die insbesondere bei den TDRs und CDRs vorliegen können, wenn in einer Bedingung häufiger ein explizites Richtungsurteil vorliegt als in der anderen. Der Welch’s *t*-Test wurde für alle Paarvergleiche standardmäßig verwendet, da er bei gleichen Varianzen nahezu die gleichen Ergebnisse liefert wie der Student‘s *t*-Test, der gleiche Varianzen annimmt, und bei allen präregistrierten Zwischensubjektvergleichen immer explizit präregistriert wurde.

Versuchsperson kodiert wurden. In der Bedingung mit eigenen Urteilen waren die Versuchspersonen mit einer $TDR_{Selbst}$ von $M$ = .60 ($SD$ = .10) in der Lage, mit ihren Richtungsurteilen überzufällig häufig die Richtung des wahren Werts einzuschätzen, $t(73)$ = 8.46, $p$ < .001 (einseitig; P2), $d$ = 0.98. Somit konnte der zentrale Effekt aus Studie 1 repliziert werden, wobei der Effekt deskriptiv sogar größer ausfiel als in Studie 1 ($M$ = .55, $d$ = 0.43). H1 konnte auch für Studie 2 angenommen werden.

Für die Bedingung mit fremden Urteilen wurde der gleiche gerichtete Test präregistriert. Auch hier waren die Versuchspersonen in der Lage, die Richtung des wahren Werts mit einer $TDR_{Fremd}$ von $M$ = .66 ($SD$ = .09) überzufällig häufig richtig einzuschätzen, $t(74)$ = 15.79, $p$ < .001 (einseitig; P2), $d$ = 1.82. Somit zeigte sich auch Evidenz für H3. Nach H4 sollten sich zudem die TDRs zwischen den Bedingungen unterscheiden und bei fremden Urteilen größer ausfallen. Für diesen Vergleich wurde ein zweiseitiger Welch's $t$-Test für ungleiche Varianzen präregistriert (siehe P2). Hierbei zeigte sich, dass sich die TDRs der beiden Bedingungen statistisch signifikant unterschieden, $t(145.38)$ = 4.33, $p$ < .001 (zweiseitig; P2), $d$ = 0.71 (siehe auch Abbildung 5). Die Versuchspersonen waren bei fremden Urteilen häufiger als bei ihren eigenen Urteilen in der Lage, die Richtung des wahren Werts korrekt einzuschätzen. Auch H4 konnte somit angenommen werden. Deskriptiv fielen diese Effekte für alle Aufgabenkategorien ähnlich aus (siehe Tabelle 2). Zudem lagen die TDRs bei eigenen und fremden Urteilen, anders als in der ersten Studie, für alle Aufgabenkategorien statistisch signifikant über 50%. Diese Befunde stehen wieder im Einklang mit dem GSM, nach dem Menschen mit neuen Informationen die Richtung des wahren Werts systematisch einschätzen können, und zwar insbesondere, wenn sie die Urteile fremder Personen bewerten. Diese Befunde zeigten sich zudem für alle untersuchten Aufgabenkategorien.

Im nächsten Schritt wurde für die Bedingung mit eigenen Urteilen äquivalent zur ersten Studie getestet, inwiefern Menschen mit ihren Richtungsurteilen die Weisheit der Vielen detektierten. Hierbei lag in der Bedingung mit eigenen Urteilen die $CDR_{Selbst}$ mit $M$ = .64 ($SD$ = .10) über .50, $t(73)$ = 11.36, $p$ < .001 (zweiseitig), $d$ = 1.32. H2a konnte somit auch in Studie 2 angenommen werden. Auch bei der Unterscheidung der konsistenten Fälle (71%), bei denen Populationsmittelwert und wahrer Wert in der gleichen Richtung lagen, und der inkonsistenten Fälle (26%) zeigte sich ein ähnliches Muster zu Studie 1. Die $CDR_{Selbst}$ lag bei den konsistenten Fällen $M$ = .67 ($SD$ = .13) über .50, $t(73)$ = 11.45, $p$ < .001 (zweiseitig), $d$ =

1.33, und auch bei den inkonsistenten Fällen mit einer $CDR_{Selbst}$ von $M$ = .58 (SD = .16) über .50, $t(73) = 4.18$, $p < .001$ (zweiseitig), $d$ = 0.49. Somit konnte auch der Effekt zur H2b repliziert werden, das heißt die Versuchspersonen detektieren den Populationsmittelwert und nicht den wahren Wert. Deskriptiv fielen diese Effekte für alle Aufgabenkategorien ähnlich aus (siehe Tabelle 3). Eine Ausnahme waren lediglich die historischen Schätzungen, bei denen Menschen nur systematisch zum Populationsmittelwert tendierten, wenn wahrer Wert und Populationsmittelwert in der gleichen Richtung lagen, wobei auch diese Analyse wieder nur auf einer relativ geringen Fallzahl beruht. Insgesamt detektierten die Menschen mit ihren Richtungsurteilen die Weisheit der Vielen und zwar unabhängig davon, ob der wahre Wert mit dieser Richtung übereinstimmte.

Nach dem GSM sollten sich auch ähnliche Effekte in der Bedingung mit fremden Urteilen zeigen, da die Richtungsurteile auch in diesem Fall auf einen gemeinsamen Pool zurückgehen. Die CDRs sollten sogar noch größer ausfallen, da bei fremden Urteilen nach dem GSM anteilig mehr neue Informationen aus dem geteilten Pool gezogen werden. Es zeigte sich, dass die $CDR_{Fremd}$ mit $M$ = .67 ($SD$ = .09) größer ausfiel als .50, $t(74) = 16.39$, $p < .001$ (zweiseitig), $d$ = 1.89, und zudem gemäß Welch's $t$-Test größer als die $CDR_{Selbst}$, $t(144.08) = 2.21$, $p = .029$ (zweiseitig), $d$ = 0.36. Somit zeigte sich Evidenz für H6a und H7a. Bei einer Unterteilung der konsistenten (71%) und inkonsistenten Fälle (26%) zeigte sich jedoch, dass nur im ersteren Fall die $CDR_{Fremd}$/$TDR_{Fremd}$ mit $M$ = .74 ($SD$ = .10) über .50 lag, $t(74) = 21.53$, $p < .001$ (zweiseitig), $d$ = 2.49, und im letzteren Fall die $CDR_{Fremd}$ mit $M$ = .52 ($SD$ = .18) nicht statistisch signifikant von .50 abwich, $t(74) = 1.07$, $p = .287$ (zweiseitig), $d$ = 0.12. Für die fremden Urteile zeigte sich damit keine Evidenz dafür, dass Menschen eher den wahren Wert als die Weisheit der Vielen detektierten (H6b). Damit lag die $CDR_{Fremd}$ auch nur für die konsistenten Fälle über der $CDR_{Selbst}$, $t(135.43) = 3.69$, $p < .001$ (zweiseitig), $d$ = 0.60, und für die inkonsistenten Fälle zeigte sich sogar ein Trend in die entgegengesetzte Richtung, $t(145.50) = -1.99$, $p = .049$ (zweiseitig), $d$ = -0.33. Damit zeigte sich auch keine Evidenz für H7b, wobei für Tests von H6b und H7b die Kennwerte wieder auf einer wesentlich kleineren Fallzahl beruhten. Insgesamt wurde mit den expliziten Richtungsurteilen bei fremden Urteilen überzufällig häufig der Populationsmittelwert gewählt und auch häufiger als bei eigenen Urteilen. Diese Effekte traten jedoch nicht unabhängig von der Richtung des wahren Werts auf, sondern sind primär auf die Mehrheit der Fälle zurückzuführen, bei denen wahrer

Wert und Populationsmittelwert in gleicher Richtung lagen (für Statistiken pro Aufgabenkategorie siehe Tabelle 3).

### 4.3. Studie 3

Mit Studie 3 wird eine weitere Untersuchung berichtet, die es erlaubt, die Generalisierbarkeit des Weisheitsbewusstseins zu testen. Ziel ist es auch in dieser Studie, den zentralen Befund zum Weisheitsbewusstsein zu replizieren, wobei anders als in der zweiten Studie komplett neues Aufgabenmaterial verwendet wurde. In den bisherigen Studien wurde zur Untersuchung des Weisheitsbewusstseins primär Aufgabenmaterial verwendet, bei dem die Versuchspersonen zur Urteilsfindung Informationen komplett aus ihrem Gedächtnis abrufen mussten. Nach dem GSM bleibt das Weisheitsbewusstsein jedoch nicht auf solche reinen Wissensfragen beschränkt, sondern sollte sich auch dann zeigen, wenn zusätzliche Informationen durch externe Medien wie Bilder zur Verfügung gestellt werden. Auch bei solchen Urteilteilsaufgaben sollte sich ein Weisheitsbewusstsein ausbilden, sofern mehr Informationen zur Verfügung gestellt werden (und zusätzlich im Langzeitgedächtnis verfügbar sind) als im Rahmen eines Initialurteils verarbeitet werden können. In Studie 3 wird diese generalisierende Annahme mit einem Bilderset von 42 Gerichten aus Kochbüchern getestet, bei denen es die Aufgabe der Versuchspersonen war, allein auf Basis der Abbildung zu schätzen, wie viele Kilokalorien pro Portion in dem jeweiligen Kochbuch für die Mahlzeit angegeben wurden.[10] Zudem wurde mit Studie 3 das Weisheitsbewusstsein im Rahmen einer Onlinestudie anhand einer neuen Population von Entscheidungsträger:innen untersucht. Hierbei handelte es sich um eine US-amerikanische Stichprobe von Erwachsenen, die über MTurk rekrutiert wurden. Zur Anpassung der Instruktionen an den neuen Aufgabentyp und den Onlinekontext wurden teilweise Instruktionen von Soll et al. (2021) übernommen, die in ihrer Untersuchung den gleichen Aufgabentyp im Rahmen einer englischsprachigen Onlineuntersuchung zur Ratschlagsnutzung verwendet hatten und ihr Material zur Verfügung gestellt hatten.

[10] Mit der Erlaubnis des Verlags wurden hierfür Gerichte aus den folgenden Kochbüchern des Naumann & Göbel Verlags in der Studie abgebildet: Die Printausgabe „Das Ultimative Studentenkochbuch“ (2018, S. 23, 27, 60, 67, 74, 91, 117) und die E-Books „Soul Food - Veggie-Blitzrezepte“ (2015, S. 17, 21, 22, 37, 49), „1000 Vegetarische Gerichte“ (2014, S. 9, 21, 34, 65, 74, 84, 97, 307, 949, 760, 793), „1000 Rezepte für zwei“ (2014, S. 242, 279, 346, 350, 398, 480, 512, 715, 812), „Schnelle Gerichte“ (2015, S. 10, 11, 16, 17, 22, 23), und „30-Minuten-Rezepte“ (2017, S. 33, 35, 40, 47).

### 4.3.1. Design & Stichprobe

Auch bei Studie 3 handelt es sich wie bei Studie 1 um eine einfache Beobachtungsstudie mit zwei Phasen zur Untersuchung des Weisheitsbewusstseins. Anders als in den ersten beiden Studien wurde die angestrebte Stichprobengröße in Studie 3 jedoch anhand pragmatischer Überlegungen ermittelt. Zudem handelt es sich hierbei um eine der beiden Studien, die nicht präregistriert wurden. Konkret richtete sich die Planung der Stichprobengröße nach der Möglichkeit, die Antworten aus Phase 1 später als Ratschläge in einer anderen Studie zu verwenden (siehe Studie 8 in Abschnitt 6). Hierfür wurden mindestens 100 quantitative Urteile von unterschiedlichen Personen pro Gericht angestrebt, wobei jede Versuchsperson immer nur die Kalorien für eine Subgruppe von 14 von insgesamt 42 Gerichten schätzte (1/3 der Gerichte). Zudem wurden vor der Erhebung verschiedene Ausschlusskriterien definiert, um eine hohe Datenqualität sicherzustellen. Diese Ausschlusskriterien werden unten detaillierter beschrieben. Auf Basis eigener Erfahrungswerte wurde bei einem entsprechenden Ausschluss von einem Datenverlust von 10%-15% ausgegangen. Vor diesem Hintergrund wurden insgesamt 345 Proband:innen zur Teilnahme an der Studie bei MTurk angefragt, um im finalen Datensatz ein Stichprobengröße von $N = 300$ zu erreichen, und damit 100 Urteile pro Gericht. Aufgrund der einfacheren Möglichkeit, Proband:innen Online zu erheben, handelt es sich hierbei um ein Vielfaches der bisherigen Stichprobengröße, weshalb nach dem Stand der ersten beiden Studien von einer Power von $1 - \beta < .99$ zur Testung des Weisheitsbewusstseins (H1) auszugehen ist (unter Annahme der kleineren der beiden Effektgrößen aus Studie 1 und 2; $d = 0.43$). Wie in den vorherigen Studien wurden in den Analysen nur Versuchspersonen berücksichtigt, die die Studie komplett bearbeiteten. Zudem wurden zur Sicherung einer hohen Datenqualität direkt bei der Einladung über CloudResearch alle Personen ausgeschlossen, die bereits an einer meiner Studien zur quantitativen Urteilsbildung über die Plattform teilgenommen hatten oder deren Arbeit bei MTurk in mehr als 5% ihrer Aufträge abgelehnt worden war. Zudem wurden hierüber nur Proband:innen aus den USA eingeladen und es wurde die Funktion „Block Suspicious Geolocations“ genutzt, um Teilnahmen aus Regionen zu verhindern, die von der Plattform CloudResearch mit doppelten Teilnahmen und Teilnahmen von „Bots“ in Verbindung gebracht wurden.

Die Studie wurde von 354 Proband:innen bis zur Zuteilung zu den Bildersets bearbeitet. Hiervon haben sieben die Studie nicht beendet und weitere 41 wurden aufgrund wiederholter Falschantworten bei Verständnisfragen ausgeschlossen, die weiter unten detaillierter beschreiben werden. Zusätzlich wurden sieben Datensätze von Proband:innen ausgeschlossen, deren Nutzerkennung (MTurk ID) mehrfach im Datensatz auftauchte und die die Studie somit mehrfach bearbeitet hatten. Die finale Stichprobe enthält daher Antworten von *N* = 299 Proband:innen (171 Männer, 127 Frauen, eine nichtbinär/divers) im Alter von *M* = 39.70 Jahren (*Mdn* = 36, *SD* = 12.86). Die Versuchspersonen erhielten für die Bearbeitung der Studie eine Grundvergütung von 2$ und einen leistungsabhängigen Bonus von bis zu 2$ extra (*M* = 0.80, *SD* = 0.70).

### 4.3.1. Methode

Bei Studie 3 handelt es sich um eine weitere konzeptuelle Replikation von Studie 1. Sie entspricht dieser im Aufbau mit einer Phase 1 für quantitative Initialurteile und einer Phase 2 für Richtungsurteile. Da die Studie jedoch online von einer US-amerikanischen Stichprobe bearbeitet werden sollte, wurden neben der Übersetzung in die englische Sprache mehrere Änderungen vorgenommen, die im Folgenden detaillierter beschrieben werden. Es wurden insbesondere Änderungen vorgenommen, um eine Datenqualität zu gewährleisten, die möglichst der von Laboruntersuchungen entspricht. Als erstes Instrument wurde den Proband:innen auf der ersten Seite der Studie eine sogenannte reCAPTCHA-Aufgabe von Google präsentiert (https://www.google.com/recaptcha/about/), um zwischen Antworten von echten Menschen und automatisierten Antworten zu unterscheiden. Auf Basis des Algorithmus von Google wurden den Proband:innen teilweise kleine Aufgaben mit mehreren Bildern präsentiert und die Proband:innen mussten angeben, auf welchen Bildern bestimmte Dinge wie Ampeln, Busse oder Straßenschilder zu sehen sind. Sofern die reCAPTCHA-Aufgabe erfolgreich gelöst wurde, wurde auf der nächsten Seite ein kurzer Aufmerksamkeitscheck präsentiert. Hierbei handelte es sich um eine Mathematikaufgabe (7 x 9 = ?) mit unterschiedlichen Lösungen im Multiple-Choice-Format (84, 63, 30, 56, 34). In den Instruktionen wurden die Versuchspersonen darauf hingewiesen, dass es ihre Aufgabe sei, Kalorien zu schätzen und dass sie hierfür die folgenden Instruktionen bitte gründlich lesen sollten. Um zu zeigen, dass sie die Instruktionen gründlich gelesen hätten, sei es ihre Aufgabe, bei der unten dargestellten Matheaufgabe die Zahl 30 als Lösung zu wählen. Nur

bei erfolgreicher Bearbeitung des reCAPTCHA und des Aufmerksamkeitschecks konnten die Versuchspersonen mit der Studie starten und bekamen auf der ersten Seite die Einverständniserklärung mit den allgemeinen Informationen zur Studie und zur Datenerhebung präsentiert.

Nach Abgabe ihres Einverständnisses folgten detailliertere Informationen zu den Kalorienschätzungen, die größtenteils von Soll et al. (2021) übernommen wurden. Aufgabe sei es im Folgenden, mehrere Schätzungen zu Kalorien verschiedener Gerichte abzugeben. Bei dem gebräuchlichen Begriff „Kalorie" handele es sich um die Kurzform des Begriffs „Kilokalorie", der bemisst, wie viel Energie notwendig ist, um ein Kilogramm Wasser um ein Grad Celsius zu erhitzen. Als Referenz wurde darauf hingewiesen, dass die USDA (United States Department of Agriculture) für die meisten Erwachsenen einen täglichen Konsum zwischen 1.600 und 3.000 Kalorien empfiehlt. Es folgte ein kurzer einfacher Verständnischeck dieser Instruktionen mit zwei Multiple-Choice Fragen („Was ist die Aufgabe in dieser Studie?"/ Lösung: „Die Schätzung von Kalorien in verschiedenen Gerichten"; „Wie viele Kalorien empfiehlt die USDA pro Tag für die meisten Erwachsenen?"/ Lösung: „Zwischen 1.600 und 3.000"), wobei die Versuchspersonen die Lösungen in den Instruktionen nachlesen konnten. Bei Falschbeantwortung einer der beiden Fragen wurde hierauf hingewiesen, und Instruktionen und Verständnisfragen wurden dann erneut präsentiert. Proband:innen, die mindestens eine der beiden Verständnisfragen ein zweites Mal falsch beantworteten, durften die Studie zu Ende bearbeiten, ihre Daten wurden jedoch später aus dem Datensatz ausgeschlossen ($n$ = 41)[11]. Zudem wurden diese Proband:innen später bei der randomisierten Zuteilung zu den Bildersets separat berücksichtigt, um eine gleichmäßige Zellbesetzung *nach Ausschluss* zu gewährleisten. Dieses Prozedere wurde verwendet, um im Einklang mit den Ethikrichtlinien der Duke University eine Gleichbehandlung aller Proband:innen sicherzustellen, ohne hierfür Verluste in der Datenqualität in Kauf nehmen zu müssen.

[11] Hierbei handelt es sich um eine Zahl, die meinen eigenen Erfahrungswerten und denen meiner US-amerikanischen Kolleg:innen mit MTurk-Studien entspricht. Hierbei handelt es sich natürlich nur um anekdotische Evidenz, da sich die Kriterien zwischen den Studien teilweise deutlich unterscheiden. Detaillierte Zahlen finden sich beispielsweise auch bei Soll et al. (2021) in dem Anhang ihrer Arbeit, bei denen Versuchspersonen in ähnlichen Größenordnungen ausgeschlossen wurden.

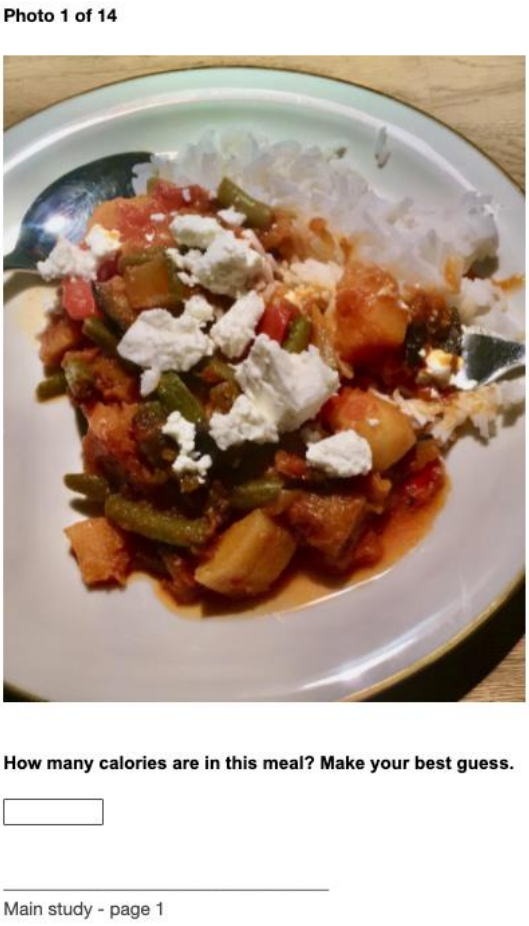
Auf den nächsten Seiten wurden äquivalent zu den ersten Studien Informationen zur Aufgabe in Phase 1 und zum leistungsabhängigen Bonus präsentiert. Aufgabe der Versuchspersonen sei es, ihre bestmöglichen Schätzungen der Kalorien von 14 Gerichten abzugeben. Bestmögliche Schätzungen seien die Schätzungen, bei denen sie sich gleichermaßen sicher seien, dass der wahre Wert über oder unter ihrer Schätzung liege. Für besonders akkurate Schätzungen würden sie einen leistungsabhängigen Bonus erhalten. Hierfür würde am Ende der Studie eines der präsentierten Gerichte zufällig ausgewählt werden und ein Bonus wie folgt berechnet: Der Bonus starte für das Bild bei 2$, wobei für je 10 Kalorien, um welche die Schätzung den wahren Wert verfehle, 10 Cent vom Bonus abgezogen werden würden. Je akkurater die Schätzung ausfalle, desto höher sei also der Bonus. Auf der letzten Instruktionsseite folgte eine kurze Zusammenfassung der Instruktionen mit dem Hinweis, dass die wahren Werte zur Berechnung der Boni aus den entsprechenden Kochbüchern entnommen wurden, und dass das komplette Gericht mit allen Beilagen pro Portion geschätzt werden solle.

Auf der nächsten Seite konnten die Versuchspersonen die Studie mit einem Code starten, der zu diesem Zeitpunkt generiert wurde, um die Versuchspersonen randomisiert einem der verschiedenen Bildersets zuzuordnen. Von dem Programm wurde (nach Ausschluss) eine gleichmäßige Zellbesetzung angestrebt. Die 42 Bilder wurden in sechs Sets aus je sieben Bildern unterteilt, und jede Versuchsperson sah immer zwei der sechs möglichen Sets (insgesamt 15 mögliche Kombinationen aus zwei Sets). Dieses Vorgehen wurde gewählt, um die Länge der Studie für das Onlinesetting zu reduzieren, ohne hierfür auf eine möglichst große Auswahl von Stimuli zu verzichten. Diese 14 Bilder wurden für jede Versuchsperson in randomisierter Reihenfolge präsentiert, wobei wieder die gleiche Reihenfolge für Phase 1 und 2 eingehalten wurde. Nach Eingabe des Codes folgte Phase 1 mit der Eingabe der Initialurteile nach dem Muster der ersten drei Studien, wobei die Antwortmöglichkeiten auf ein Minimum von 0 und ein Maximum von 5.000 Kalorien begrenzt wurden (siehe Abbildung 6[12]).

[12] In dem Beispiel wurde das urheberrechtlich geschützten Bildermaterial von Naumann & Göbel mit einem Beispiel aus meiner eigenen Küche ausgetauscht.

**Abbildung 6**

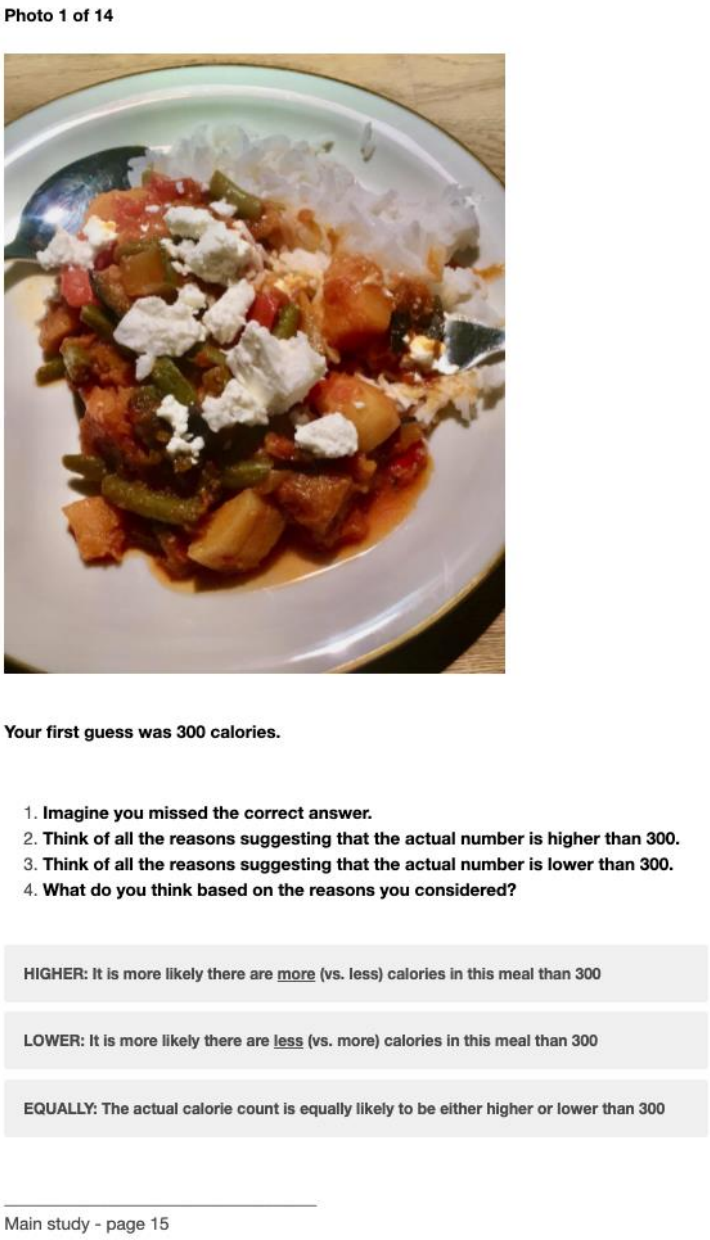
*Beispielaufgabe zu Phase 1 von Studie 3*




Nach Abgabe des letzten Urteils folgten die Instruktionen zu Phase 2, äquivalent zu den vorherigen Studien. Diese wurden jedoch schrittweise und ohne minimale Darbietungsdauer präsentiert. Zudem wurde eine Variante gewählt, bei der der dritte Schritt – das ist derjenige, in dem die Versuchspersonen sich erst Gründe für die eine und im nächsten Schritt Gründe für die andere Richtung überlegen sollten – in zwei Schritte aufgeteilt wurde. Die entsprechenden Instruktionen wurden im Anschluss auch bei der Abgabe jedes einzelnen Richtungsurteils in gekürzter Form gegeben. Für alle Richtungsurteile wurde erneut das Gericht und das quantitative Urteil aus der ersten Phase präsentiert (siehe Abbildung 7).

**Abbildung 7**

*Beispielaufgabe zu Phase 2 von Studie 3*




Nachdem das letzte Richtungsurteil abgegeben wurde, folgte auch in dieser Studie der optionale Abschlussfragebogen zu Geschlecht und Alter (mit der Möglichkeit diesen unbeantwortet zu lassen) und ein schriftliches Debriefing mit einem Hinweis auf den erreichten Bonus. Zudem gab es in dieser Studie ein optionales Kommentarfeld für Probleme und Hinweise jeglicher Art und es wurde auf dieser Seite mit Absprache des Verlags der Urheberverweis für das verwendete Bildmaterial platziert. Die Versuchspersonen wurden im Anschluss ausgezahlt, wobei die Bonuszahlungen erst kurz nach Abschluss der

Datenerhebung überwiesen wurde, worauf auf der letzten Seite der Studie explizit hingewiesen wurde.

### 4.3.2. Ergebnisse & Diskussion

Auch bei den Kalorienschätzungen in Studie 3 wählten die Proband:innen mit einer DJR von *M* = .67 (*SD* = .21) in der Mehrheit der Fälle ein Richtungsurteil. Drei Versuchspersonen wählten in keinem der Durchgänge ein explizites Richtungsurteil und wurden von den folgenden Analysen ausgeschlossen. Auch in Studie 3 lag die TDR mit *M* = .57 (*SD* = .20) über .50, $t(295) = 5.97$, $p < .001$ (zweiseitig), $d = 0.35$. H1 konnte somit auch für dieses Aufgabenmaterial angenommen werden und das Weisheitsbewusstsein ein weiteres Mal repliziert werden. Auch die CDR lag mit *M* = .59 (*SD* = .20) über .50, $t(295) = 7.88$, $p < .001$ (zweiseitig), $d = 0.46$. Werden nur die Fälle betrachtet, bei denen Populationsmittelwert und wahrer Wert in der gleichen (78%) vs. unterschiedlichen Richtungen lagen (21%), so zeigte sich, dass die CDR/TDR bei den konsistenten Fällen mit *M* = .61 (*SD* = .23) größer ausfiel als .50, $t(295) = 8.52$, $p < .001$ (zweiseitig), $d = 0.50$. Bei den inkonsistenten Fällen lag die CDR mit *M* = .54 (*SD* = .36) jedoch nur deskriptiv über .50, $t(215) = 1.81$, $p = .071$ (zweiseitig), $d = 0.12$. Somit zeigte sich in dieser Studie nur Evidenz für H2a, aber nicht für H2b. Das heißt, dass die Menschen mit ihren expliziten Richtungsurteilen zwar deskriptiv eher die Richtung der Weisheit der Vielen als die des wahren Werts wählten, wobei sich in dieser Studie inferenzstatistisch keine klare Evidenz für diese Vorhersage des GSM zeigte. Ein Problem könnte wieder die vergleichsweise geringe statistische Power für den Test von H2b sein, worauf in der folgenden Abschnittsdiskussion detaillierter eingegangen wird.

## 4.4. Diskussion zum Weisheitsbewusstsein

In dem ersten empirischen Abschnitt wurde mit dem Weisheitsbewusstsein eine wesentliche Vorhersage des GSM getestet, nach welcher Menschen selbst erkennen können, in welche Richtung sie ihre eigenen quantitativen Urteile korrigieren sollten. So zeigte sich zunächst, dass Menschen häufig nach einem ersten Urteil vermuten, dass der wahre Wert eher über oder eher unter ihrem ersten Urteil liegt. Dieser deskriptive Befund zeigte sich, obwohl die Menschen auch die Möglichkeit hatten, keine der beiden Richtungen zu wählen und zudem finanzielle Anreize bekommen haben, bei ihren ersten Urteilen möglichst

akkurate Schätzungen abzugeben. Wesentlich zentraler für mögliche Urteilskorrekturen ist jedoch, dass Menschen mit ihren expliziten Richtungsurteilen überzufällig häufig richtig lagen und damit häufig wussten, in welche Richtung sie diese ersten Urteile korrigieren sollten. Das Weisheitsbewusstsein zeigte sich in allen drei Studien und konnte zudem auf verschiedene Aufgabenkategorien und Populationen von Entscheidungsträger:innen generalisiert werden. Allein bei den Prozentschätzungen in Studie 1 zeigte sich kein systematisches Weisheitsbewusstsein und es fiel auch in Studie 2 deskriptiv kleiner aus als bei den anderen Aufgabenkategorien (auch wenn es in Studie 2 nachgewiesen werden konnte). Dies könnte auch erklären, warum die Daten von Gaertig und Simmons (2021) eher gegen ein Weisheitsbewusstsein sprechen (abgesehen von den in Abschnitt 2.4 diskutierten methodischen Problemen), da diese vornehmlich Prozentschätzungen als Aufgabentyp in ihrer Arbeit verwendeten. Eine einfache Erklärung könnte hierfür sein, dass bei Prozentschätzungen durch die Begrenzung der Skala bei 0% und 100% weniger starke Fehler auftreten, die schwerer zu erkennen sind. Insgesamt sprechen diese Ergebnisse dafür, dass es sich bei dem Weisheitsbewusstsein um eine robuste Fähigkeit handelt, die vermutlich aufgaben- und domänenübergreifend genutzt werden kann. Das Weisheitsbewusstsein sollte sich damit auch bei Aufgaben zeigen, bei denen tendenziell weniger starke Fehler auftreten, auch wenn der Effekt in diesen Fällen schwächer ausfallen könnte.

Aus dem GSM wurden zudem noch weitere Hypothesen abgeleitet, um das Modell zur Erklärung dieses Phänomens kritisch zu testen. So sagt das GSM vorher, dass Menschen bei der Bewertung ihrer eigenen Urteile meist eher den Populationsmittelwert als den wahren Wert detektieren. Die Befunde der ersten Studien stehen größtenteils im Einklang mit dieser Vorhersage. Dies zeigte sich insbesondere darin, dass die Versuchspersonen mit ihren expliziten Richtungsurteilen eher die Richtung des Populationsmittelwerts wählten als die des wahren Werts, wenn diese in unterschiedlichen Richtungen lagen (die inkonsistenten Fälle; H2b). Damit unterliegt das Weisheitsbewusstsein der gleichen Limitation wie die Weisheit der Vielen und kann damit nur eigene Fehler erkennen, die auch von der Weisheit der Vielen korrigiert werden können. Kritisch ist jedoch anzumerken, dass die Tests dieser Vorhersage allein auf Basis der inkonsistenten Fälle vermutlich eine relativ schwache statistische Power haben und sich der Effekt deswegen auch nur in zwei von drei Studien gezeigt haben könnte. Zwei Faktoren können zu einer geringen statistischen Power

beitragen haben: eine geringe Anzahl von Beobachtungen und eine potentiell geringere Effektstärke für den in H2b angenommenen Effekt im Vergleich zu den konsistenten Fällen. Zum einen hat sich bewahrheitet, dass diese Fälle bei einem repräsentativen Design (keine systematische Vorauswahl der Stimuli) vergleichsweise selten auftreten, wodurch naturgemäß auch die entsprechende Power schwächer ausgefallen sein muss. Zum anderen sollte der Effekt für die inkonsistenten Fälle kleiner ausfallen als für die konsistenten Fälle. Der Grund hierfür liegt darin, dass Menschen bei den inkonsistenten Fällen mit ihren ersten Urteilen vermutlich relativ nah am Populationsmittelwert liegen, wenn ihr Urteil zwischen dem häufig akkuraten Populationsmittelwert und dem wahren Wert liegt. Evidenz hierfür liefert die Weisheit der Vielen, die dem Populationsmittelwert eine starke Nähe zum wahren Wert attestiert (Herzog et al., 2019; Larrick et al, 2012). Je näher das erste Urteil am Populationsmittelwert liegt, desto schwerer sollte es sein, anhand neuer (zufällig gezogener) Informationen dessen Richtung zu detektieren. Wie in Abschnitt 2.3.2. kurz erwähnt, würde es sich auf Informationsebene um einen einfachen Regression-zur-Mitte Effekt handeln (Fiedler & Krueger, 2012), der umso schwächer ausfällt, je näher das erste Urteil an der Mitte der Verteilung liegt (dem Populationsmittelwert).

Für das Weisheitsbewusstsein sind diese Einschränkungen unter praktischen Gesichtspunkten grundsätzlich positive Nachrichten, da dies letztlich nur bedeutet, dass sich eine wesentliche Limitation bei einer repräsentativen Aufgabenauswahl zu selten zeigt und in diesen Fällen nicht besonders stark ausfällt. Nichtsdestotrotz erschweren diese Umstände einen kritischen und teststarken Test der Theorie. Dieses Problem ließe sich zum einen experimentell lösen, indem durch die Auswahl der Stimuli manipuliert wird, ob der Populationsmittelwert und der wahre Wert nah beieinander oder weit entfernt voneinander liegen. Eine zweite Möglichkeit ist, im Rahmen weiterer Studien, ohne Manipulation der Stimuli, erheblich mehr Daten zu erheben und H2b später anhand aller Daten metaanalytisch mit einer angemessenen Power zu testen. In dieser Arbeit wurde letzteres Vorgehen gewählt, da es erlaubt, parallel den praktischen Nutzen des Weisheitsbewusstseins im Rahmen mehrerer Interventionsstudien zu testen, womit als „positiver Nebeneffekt“ auch mehr Daten für einen späteren metaanalytischen Test von H2b erhoben wurden. Entsprechend wird in den folgenden Studien im Text nur berichtet, ob sich Menschen grundsätzlich konsistent mit dem GSM verhalten und über alle Fälle hinweg zum

Populationsmittelwert tendieren (H2a), während ein erneuter Test der H2b bezüglich der inkonsistenten Fälle metaanalytisch in Abschnitt 7 erfolgt (siehe jedoch Tabelle 3 für einen Test von H2b für die einzelnen Studien und Aufgabenkategorien).

Aus dem GSM lassen sich zudem weitere Vorhersagen bezüglich der Limitationen des Weisheitsbewusstseins im Vergleich zu einer Referenzgruppe ableiten, bei der Menschen die Richtung der wahren Werte relativ zu fremden Urteilen einschätzen (siehe Studie 2). Nach dem GSM sollten Menschen bei der Bewertung ihrer eigenen Urteile anteilig weniger neue Informationen heranziehen können als Menschen, die fremde Urteile bewerten, und daher seltener die Richtung des wahren Werts beziehungsweise des Populationsmittelwerts wählen. Im Einklang hiermit zeigte sich, dass fremde Personen häufiger ein explizites Richtungsurteil wählten als Menschen, die ihre eigenen Urteile bewerteten, und vor allem häufiger mit diesen richtig lagen. Es zeigte sich jedoch keine Evidenz dafür, dass Menschen bei fremden Urteilen eher zum Populationsmittelwert als zum wahren Wert tendieren und dass dieser Effekt stärker ausfällt als bei eigenen Urteilen. Ein Grund könnte hierfür wieder die gleiche Powerproblematik sein, die bei der Bewertung eigener Urteile diskutiert wurde. Das heißt, dass Menschen auch bei der Bewertung fremder Urteile bei einem repräsentativen Design zu selten quantitative Urteile zwischen wahrem Wert und Populationsmittelwert bewerten und der Effekt in diesen Fällen zudem durch eine größere Nähe der bewerteten Urteile zum Populationsmittelwert schwächer ausfällt.

Zusammenfassend lässt sich sagen, dass mit den ersten Studien die wesentliche Vorhersage des GSM, die Existenz des Weisheitsbewusstseins, bestätigt werden konnte, beziehungsweise diejenige Hypothese, die unmittelbar die größte praktische Relevanz hat. Im Folgenden werde ich konkrete praktische Implikationen des GSM ableiten und Interventionen entwickeln und testen, die es Menschen erlauben sollen, ihre Urteile mit dem Weisheitsbewusstsein auch selbst zu verbessern. Im Rahmen dieser Studien werde ich als wesentliche Voraussetzung dieser Interventionen jedoch auch immer das Weisheitsbewusstsein testen. Dies erlaubt es mir später auch die kritische Vorhersage des GSM (H2b) anhand einer großen Datenmenge metaanalytisch zu prüfen (siehe Abschnitt 7). Einzig die spezifischeren Vorhersagen bezüglich fremder Urteile (H6b und H7b) lassen sich mit diesem Vorgehen nicht abschließend bewerten, da für diese wesentlich mehr Daten mit Bewertung fremder Urteile erhoben werden müssten. Dies würde zu stark von der

wesentlichen Zielsetzung dieser Arbeit wegführen und müsste deshalb in zukünftigen Untersuchungen, metaanalytisch oder durch eine gezielte Auswahl von Stimuli, geklärt werden.

## 5. Interventionen auf Basis des Weisheitsbewusstseins im individuellen Urteilskontext

In diesem Abschnitt werde ich Studien zu Interventionen im individuellen Urteilskontext vorstellen, die Menschen auf Basis ihres Weisheitsbewusstseins in die Lage versetzen sollen, ihre eigenen Urteile selbständig zu korrigieren und zu verbessern. Interventionen zur Verbesserung von Urteilsprozessen werden in der Urteils- und Entscheidungsforschung auch als *Debiasing*-Strategien bezeichnet (Fischhoff, 1982; Soll et al., 2015)[13]. Insgesamt gibt es jedoch nur wenige erfolgreiche Debiasing-Strategien, sodass der Bedarf nach neuen Strategien zur Urteilskorrektur groß ist (Milkman et al., 2009), auch wenn für quantitative Urteile gerade in den letzten 10-15 Jahren Strategien entwickelt wurden, die die Urteilsprozesse zumindest teilweise verbessern können (z.B., Herzog & Hertwig, 2014a; Mellers et al., 2014; Welsh & Begg, 2018). Die Neuerung der hier vorgeschlagenen Interventionen ist, dass Menschen im Rahmen der Interventionen bereits entstandene Fehler mit ihrem Weisheitsbewusstsein selbstständig erkennen und korrigieren sollen.

Bisher gibt es nach meinem Kenntnisstand keine Evidenz dafür, dass Menschen ohne Intervention ihre eigenen quantitativen Urteile systematisch verbessern können. So zeigt beispielsweise die Literatur zur Weisheit der Vielen in einer Person, dass die zweiten quantitativen Urteile von Menschen nicht akkurater sind als die ersten, sondern tendenziell sogar einen größeren Fehler aufweisen (Fraundorf & Benjamin, 2014; Gaertig & Simmons, 2021; Herzog & Hertwig, 2014b; Steegen et al., 2014; Vul & Pashler, 2008). Diese Studien zielten jedoch nicht darauf ab, dass das zweite Urteil selbst akkurater war als das erste. Vielmehr sollte durch die Bildung eines Mittelwerts eine Verbesserung erzielt werden. Mit den hier entwickelten und untersuchen Interventionen ist es hingegen das Ziel, Menschen in

[13] Der Begriff des Debiasing ist etwas irreführend, da hiermit meist nicht nur Interventionen zur Vermeidung systematischer Fehler (*Bias*) gemeint sind, sondern auch Interventionen, die vornehmlich auf die Korrektur unsystematischer Fehler abzielen (*Noise*). So wird beispielsweise auch die Bildung eines Mittelwerts über mehrere Urteile von Soll et al. (2015) als Debiasing-Strategie aufgefasst.

die Lage zu versetzen ein zweites (korrigiertes) Urteil abzugeben, das akkurater ist als das erste.

Als wesentliche Voraussetzung für eine solche Intervention konnte im vorherigen Abschnitt gezeigt werden, dass Menschen überzufällig gut erkennen können, in welche Richtung sie ihre Urteile korrigieren sollten. Als nächstes stellt sich die Frage, ob Menschen (mit Hilfe von Interventionen) auch in der Lage sind, ihre Urteile in die entsprechende Richtung angemessen zu korrigieren. Eine Möglichkeit wäre, dass es allein ausreicht, dass Menschen bei der Bewertung ihres Urteils ein Weisheitsbewusstsein besitzen, damit sie spontan eine Korrektur in die entsprechende Richtung vornehmen und selbständig ihre Urteile verbessern. Ich nehme jedoch an, dass Menschen selbst bei einem vorliegenden Weisheitsbewusstsein keine solche spontanen Korrekturen vornehmen oder, selbst wenn sie diese vornehmen, dass diese zumindest eher schwach ausfallen. So zeigt sich, dass Menschen selbst dann häufig ihre Urteile nicht oder nur gering anpassen, wenn sie Ratschläge oder andere zusätzliche Informationen von externer Seite erhalten (Edwards, 1982; Rader et al., 2017; Soll & Larrick, 2009, Yaniv & Kleinberger, 2000). Zudem deuten die methodischen Entscheidungen bisheriger Arbeiten zur Weisheit der Vielen in einer Person darauf hin, dass Menschen nur mit sehr expliziten Instruktionen und starken Anreizen ihre Urteile ohne neue Informationen von externer Seite ändern. In diesen Arbeiten werden Versuchspersonen meist explizit dazu aufgefordert ein abweichendes Urteil abzugeben, um von dem Effekt zu profitieren (z.B., Gaertig & Simmons, 2021; Steegen et al., 2014; Vul & Pashler, 2008). Teilweise werden sie sogar mit finanziellen Anreizen dazu motiviert, ein abweichendes zweites Urteil abzugeben, indem sie immer für das bessere von den zwei Urteilen einen finanziellen Bonus für kleine Fehler erhalten (Gaertig & Simmons, 2021; Herzog & Hertwig, 2014b). Es lässt sich jedoch nur spekulieren, ob solche Maßnahmen notwendig sind, um ein abweichendes zweites Urteil zu erhalten, da diese Faktoren meist nicht experimentell manipuliert werden. Eine Ausnahme sind Arbeiten zum Dialectical Bootstrapping, die eine explizite Aufforderung enthalten, ein abweichendes Urteil abzugeben und die zeigen, dass Menschen mit diesen Instruktionen ihre Urteile zumindest häufiger anpassen, auch wenn sie damit nicht unbedingt akkuratere Urteile abgeben (White & Antonakis, 2013). Ich nehme aus diesen Gründen an, dass Menschen zwar grundsätzlich bereit sind, sich konsistent mit ihrem Weisheitsbewusstsein zu verhalten, es für die

Korrektur aber expliziter Instruktionen bedarf. In den folgenden Studien werde ich als eine Voraussetzung immer untersuchen, inwiefern Menschen sich im Rahmen der jeweiligen Intervention konsistent mit dem Weisheitsbewusstsein (bzw. ihren Richtungsurteilen) verhalten.

Die von mir vorgeschlagene Intervention zur Korrektur der Urteile baut grundsätzlich auf der Abfrage des Weisheitsbewusstseins aus den vorherigen Studien auf (in Anlehnung an den Instruktionen von Herzog & Hertwig, 2009, 2014b). Die zentrale Neuerung gegenüber den Studien des vorherigen Abschnitts ist, dass Menschen explizit aufgefordert werden ihre Urteile in die Richtung anzupassen, in der sie den wahren Wert vermuten. Anders als in der bisherigen Forschung (z.B., Gaertig & Simmon, 2021; Herzog & Hertwig, 2009, 2014b) wird diese Aufforderung jedoch selektiv nur dann gegeben, wenn Menschen ein Weisheitsbewusstsein entwickeln und den wahren Wert eher in einer der beiden Richtungen vermuten. Im Falle einer Anpassung soll diese so weit erfolgen, bis Menschen das Gefühl haben, dass der wahre Wert weder eher über noch eher unter ihrem finalen (korrigierten) Urteil liegt. Ich nehme an, dass Menschen hiermit, sofern sie sich für eine Richtung entscheiden, eine angemessene Korrektur einleiten können und somit insgesamt ihre Urteile verbessern. Unter der Annahme, dass Menschen solche Korrekturen nicht spontan vornehmen, sollte diese Verbesserung zudem größer ausfallen als in einer Kontrollgruppe ohne Intervention:

> **Hypothese 8a (H8a; H2a in P4-P5)**: Menschen sind in der Lage, die Akkuratheit ihrer Urteile mit der vorgeschlagenen Intervention zu verbessern. Diese Intervention instruiert sie dazu, ihre Urteile in die Richtung anzupassen, in der sie den wahren Wert eher vermuten. Die Korrektur soll so weit erfolgen, bis sie das Gefühl haben, dass der wahre Wert weder eher in der einen noch in der anderen Richtung liegt.
>
> **Hypothese 8b (H8b; H2b in P4-P5)**: Die Intervention zur Verbesserung der Urteile (H8a) führt zu einer stärkeren Verbesserung als in einer Kontrollgruppe ohne Intervention.

## 5.1. Studie 4

Der erste Test der neuen Intervention erfolgte mit einer über MTurk erhobenen Onlinestudie, in der zwei Experimentalgruppen und eine Kontrollgruppe erhoben wurden. Die beiden Experimentalgruppen unterschieden sich nur im Rahmen der Abfrage von Konfidenzintervallen, die in einer Bedingung zu explorativen Untersuchungszwecken erhoben wurde.[14] Für die konfirmatorische Testung von H8a und H8b wurden jedoch ausschließlich Tests der Experimentalgruppe ohne Konfidenzintervalle und deren Vergleich zur Kontrollgruppe präregistriert (siehe P4), weshalb die Ergebnisse zur anderen Experimentalbedingung nur ergänzend berichtet werden. In Studie 4 wurde ein weiteres Set neuer Aufgaben verwendet. Aufgabe der Versuchspersonen war es, das Alter von Menschen auf Bildern einzuschätzen.

### 5.1.1. Design und Stichprobe

Für den ersten Test der Intervention wurde ein einfaktorielles Zwischensubjektdesign mit drei Stufen im manipulierten Faktor gewählt: Intervention/Experimentalgruppe (EG) vs. Intervention + Konfidenzintervall (EG + KI) vs. Kontrollgruppe (KG). Die Testplanung richtete sich nach dem Test von H1, H8a, und H8b, die als zentrale Hypothesen für diese Studie präregistriert wurden (siehe P4). Hierbei wurde eine Testpower von mindestens $1 - \beta = .85$ für alle drei Hypothesen angestrebt und entsprechend eine Stichprobengröße von $N = 510$ ($n = 170$ pro Zelle) gewählt. Für die neuen Hypothesen (H8a und H8b) wurden kleine Effektstärken ($d = 0.30$) angenommen (Cohen, 1988). Für kleinere Effekte als $d = 0.30$ wurde angenommen, dass sich diese nicht im Rahmen der hier finanzierbaren Studien teststark untersuchen ließen (zumindest nicht im Zwischensubjektvergleich), weshalb kleinere Effektgrößen nicht von Interesse waren. Die Poweranalyse ergab für $n = 170$ eine statistische Power von $1 - \beta > .99$ für einen einseitigen Test der H1 (bei $d = 0.43$; vor der Erhebung die bis dahin kleinste beobachtete Effektgröße für H1), $1 - \beta = .99$ für einen einseitigen Test der H8a (bei $d = 0.30$), und $1 - \beta = 0.87$ für einen einseitigen Test der H8b (bei $d = 0.30$). Zur

[14] Ziel war es, den Zusammenhang der Schiefe der Konfidenzintervalle (wenn das erste Urteile nicht genau mittig zu den Endpunkten des Intervalls liegt) und der Richtungsurteile zu untersuchen (siehe P4). Für die in dieser Arbeit untersuchte Fragestellung liefern diese Analysen jedoch meines Erachtens keinen nennenswerten zusätzlichen Erkenntnisgewinn, weshalb im Sinne der Sparsamkeit keine Analysen zu den Konfidenzintervallen berichtet werden. Alle erhobenen Daten zu den Konfidenzintervallen sind jedoch zusammen mit allen anderen Daten im OSF öffentlich zugänglich (https://osf.io/y23fs/).

Studie wurden 600 Versuchspersonen über CloudResearch eingeladen, da wie in Studie 3 von einem Datenverlust von bis zu 15% nach Datenbereinigung ausgegangen wurde (siehe P4). Bei den Einladungen zur Studie wurden die gleichen Kriterien wie bei Studie 3 angesetzt (siehe P4), wobei zu dem früheren Erhebungszeitpunkt von Studie 4 (siehe Tabelle 1) die Funktion „Block Suspicious Geocode Locations“ noch nicht genutzt wurde.

Die Studie wurde von *N* = 600 Personen vollständig bearbeitet. Weitere 15 Personen haben die Studie mindestens bis zur Zuteilung zu den Versuchsbedingungen nach Phase 1 bearbeitet, jedoch vor der kompletten Bearbeitung abgebrochen. Von den vollständigen 600 Datensätzen wurden 62 Versuchspersonen wurden aufgrund von Falschantworten bei den Verständnisfragen ausgeschlossen, sieben weitere, weil sie die Studie zwei Mal bearbeitet hatten (siehe P4). Der finale Datensatz umfasste damit *N* = 531 Personen (302 Männer, 227 Frauen, zwei nichtbinär/divers) in einem Durchschnittsalter von *M* = 34.93 Jahren (*Mdn* = 32; *SD* = 10.64) mit *n* = 178 in der EG, *n* = 173 in der EG + KI, und *n* = 180 in der KG. Die Versuchspersonen wurden bei vollständiger Bearbeitung mit 2$ vergütet und erhielten zusätzlich in Abhängigkeit ihrer Leistungen einen Bonus von bis zu 2$ (*M* = 0.87, SD = 0.51).

### 5.1.2. Methode

Das Material von Studie 4 wurde genau wie das Material von Studie 3 in englischer Sprache verfasst. Auch der Ablauf der Bedingungen entsprach weitestgehend dem Ablauf von Studie 3, welche für den gleichen Proband:innenpool gestaltet worden war. Die Hauptunterschiede lagen in dem neuen Aufgabenmaterial, der Abgabe eines jeweils zweiten (und möglicherweise revidierten) Urteils zu jedem Stimulus, und dem Vergleich der drei Versuchsbedingungen. Hierdurch ergaben sich mehrere Anpassungen, weshalb im Sinne einer verständlichen Darstellung der gesamte Ablauf kurz zusammengefasst wird und nur punktuell auf ähnliche Methoden wie in Studie 3 verwiesen wird. Grundsätzlich bestand auch Studie 4 aus zwei Phasen. In der ersten Phase wurden mehrere quantitative Urteile abgegeben und in der zweiten Phase Richtungsurteile (in den EGs) und zweite (revidierte) quantitative Urteile.

Auch Studie 4 startete mit einer reCAPTCHA-Aufgabe. Anders als in Studie 3 erfolgte der zusätzliche Aufmerksamkeitscheck erst nach der Einverständniserklärung und nicht unmittelbar nach der reCAPTCHA-Aufgabe. Nach Abgabe des Einverständnisses wurde die

gleiche Mathematikaufgabe wie in Studie 3 präsentiert mit der Instruktion, dass es ihre Aufgabe sei das Alter von 20 Menschen auf Bildern zu schätzen und sie hierfür die Instruktionen genau lesen sollten. Zudem wurde ein weiterer Aufmerksamkeitscheck auf der nächsten Seite präsentiert („Was ist Ihre Aufgabe in dieser Studie?“/Lösung: „Meine Aufgabe ist es das Alter von Personen zu schätzen“), wobei die Personen nicht die vorherigen Instruktionen ein weiteres Mal lesen konnten und die Versuchspersonen nur einen Versuch zur Bearbeitung hatten. Im Einklang mit den Anforderungen der Ethikkommission wurde allen Versuchspersonen, die ihr Einverständnis zur Teilnahme gegeben hatten, erlaubt die Studie zu beenden und die Grundvergütung sowie den leistungsabhängigen Bonus zu erhalten. Die Daten derjenigen Personen, die den Aufmerksamkeitscheck nicht erfolgreich absolviert hatten, wurden aber nicht weiter berücksichtigt (siehe P4).

Nach den Screeningfragen erhielten die Versuchspersonen nochmals die Information, dass es ihre Aufgabe sei, das Alter von Menschen auf Bildern zu schätzen. Nachdem sie alle Altersschätzungen abgegeben hätten (Phase 1), sei es ihre Aufgabe, im Anschluss zusätzliche Fragen zu den Bildern zu beantworten (Phase 2). Als nächstes wurden die Versuchspersonen über mögliche Bonuszahlungen informiert. Hier erhielten die Versuchspersonen zunächst nur die Information, dass sie sowohl bis zu 1$ Bonus für Phase 1 und bis 1$ Bonus für Phase 2 erhalten könnten (insgesamt bis zu 2$). Der Bonus für Phase 1 würde auf der nächsten Seite erläutert werden, und der Bonus für Phase 2 nach der Bearbeitung von Phase 1. Zudem wurde darauf hingewiesen, dass die Antworten in Phase 1 nicht den Bonus in Phase 2 beeinflussen könnten und umgekehrt. Jedoch müssten beide Phasen komplett bearbeitet werden, um die Bonuszahlungen für beide Phasen ausgezahlt zu bekommen. Auf der nächsten Seite wurde der Bonus für Phase 1 erklärt, der der Berechnung des Bonus für die Kalorienschätzungen in Studie 3 ähnelte und ähnlich instruiert wurde. Auch hier wurde zufällig zur Berechnung eine Aufgabe und ein Urteil aus Phase 1 ausgewählt. Die Versuchspersonen erhielten dann 1$ minus 10 Cent pro Jahr, das ihre Schätzung vom wahren Wert entfernt lag.

Auf der nächsten Seite konnte Phase 1 des Experiments mit der Eingabe eines numerischen Codes gestartet werden. Diese Phase glich strukturell der Phase 1 in allen vorherigen Studien, wobei die quantitativen Urteile vom Computerprogramm immer auf ein

Alter von mindestens 10 Jahren und maximal 120 Jahren begrenzt waren. Gleiches galt für die quantitativen Urteile zu Phase 2 und die Ober- und Untergrenze der Konfidenzintervalle in der EG + KI. Eine Beispielaufgabe zu Phase 1 ist in Abbildung 8 dargestellt.[15] Genau wie in den vorherigen Studien wurden alle Stimuli auf eigenen Seiten in randomisierter Reihenfolge präsentiert. In der Erhebung wurde hierfür Bildmaterial von Minear und Park (2004) genutzt (schwarz/weiß Foto, Nahaufnahme des Gesichts, neutraler Gesichtsausdruck und neutraler Hintergrund). Auf den Bildern waren Personen unterschiedlichen Alters (von 18 bis 75) und Geschlechts mit unterschiedlichem ethnischem Hintergrund abgebildet.

**Abbildung 8**

*Beispielaufgabe zu Phase 1 von Studie 4*


Nach Abgabe der letzten Schätzung wurden die Versuchspersonen randomisiert den drei unterschiedlichen Versuchsbedingungen zugeteilt, und zwar so, dass alle

[15] Bei diesem Beispiel handelt es sich um ein Bild meines eigenen Gesichts, da ich für das verwendete Bildmaterial nicht die Erlaubnis hatte, dieses über die Durchführung der Studie hinaus zu teilen. Dieses Bild wurde nicht in der Hauptstudie verwendet.

Versuchsbedingungen gleichmäßig aufgefüllt wurden. Die Versuchspersonen, die eine der beiden Aufmerksamkeitschecks zu Beginn falsch beantworteten, wurden bei dieser gleichmäßigen Verteilung nicht berücksichtigt, konnten jedoch die Studie in einer der drei Bedingungen beenden, um nicht von den möglichen Bonuszahlungen ausgeschlossen zu werden. Die Instruktionen zu Phase 2 unterschieden sich im Anschluss je nach den drei Versuchsbedingungen.

In der EG erhielten die Versuchspersonen die Instruktion, dass sie im Folgenden die Bilder aus Phase 1 erneut mit ihren ersten Schätzungen präsentiert bekämen und es ihre Aufgabe sei einzuschätzen, ob der wahre Wert eher über, eher unter, oder weder eher in der einen noch in der anderen Richtung liege. Zusätzlich sollen sie eine zweite Schätzung abgeben, bei der sie das Gefühl hätten, dass der wahre Wert weder eher in der einen noch eher in der anderen Richtung liege. Im Anschluss wurden die detaillierten Instruktionen zu diesem Vorgehen präsentiert, wobei die Instruktionen wie in Studie 3 schrittweise präsentiert wurden. Die zentrale Neuerung war der zusätzliche Schritt, bei dem die Versuchspersonen aufgefordert wurden, ihre ersten Urteile anzupassen. Da es sich hierbei um die zentralen Interventionsinstruktionen handelt, wird im Folgenden der englische Originaltext einmal komplett abgedruckt (für die deutsche Laborvariante siehe Studie 5):

„Please follow the following steps for every photo:

1. Imagine you missed the correct answer.
2. Think of all the reasons suggesting that the person is actually older than your first guess.
3. Think of all the reasons suggesting that the person is actually younger than your first guess.
4. Decide on a direction. Is it more likely that the person is older or younger than your guess? You can also answer that they are equally likely.
5. Guess again. You should feel that the person is equally likely to be older or younger than this guess."

Im Anschluss wurden in der EG die Bonuszahlungen für Phase 2 erläutert, die genau wie in Phase 1 berechnet wurden. Die Versuchspersonen wurden darüber aufgeklärt, dass hierfür zusätzlich eine Schätzung aus Phase 2 zufällig gezogen werden würde und sie wieder 1$ abzüglich 10 Cent pro Jahr, das ihre Schätzung vom wahren Wert abweicht, erhalten würden. Im Anschluss konnte die zweite Phase bearbeitet werden. Die Bilder wurden in der gleichen Reihenfolge wie in Phase 1 präsentiert, wobei pro Aufgabe immer zwei Seiten bearbeitet wurden, bevor die nächste Aufgabe präsentiert wurde (siehe Abbildung 9). Auf der ersten Seite wurden die Richtungsurteile mit den entsprechenden Instruktionen abgefragt. Auf der zweiten Seite konnten die Versuchspersonen ihre zweite Schätzung abgeben, wobei ihnen sowohl ihr erstes Urteil als auch ihre Wahl beim Richtungsurteil angezeigt wurde. Die weiteren Instruktionen auf der zweiten Seite passten sich an die gewählte Richtungsoption an. Bei einer Wahl einer der beiden Richtungen wurden die Versuchspersonen aufgefordert ihr erstes Urteil anzupassen, bis sie das Gefühl hätten, dass der wahre Wert weder eher in der einen noch eher in der anderen Richtung liege. Im Fall der Wahl der dritten „weder noch“-Option wurden die Versuchspersonen lediglich aufgefordert ein zweites Urteil abzugeben.

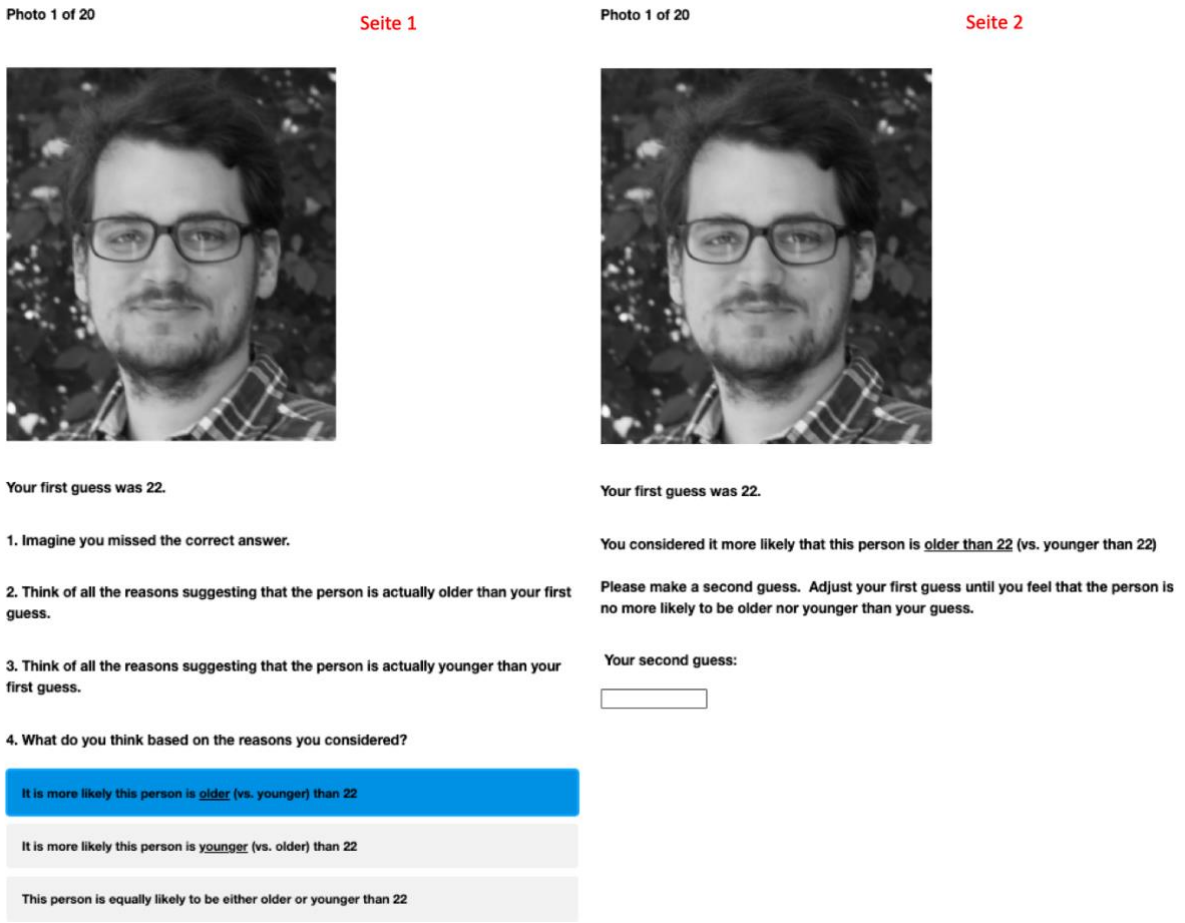
Der Ablauf in EG + KI glich der einfachen EG, wobei als Teil der Richtungsabfrage auch ein 80%-Konfidenzintervall abgefragt wurde. Hierfür unterschieden sich der zweite und dritte Schritt der detaillierten Instruktionen:

> „2. Think of all the reasons suggesting that the person is actually older than your first guess and make a high guess.
>
> *The high guess is the age for which you are 90% confident that the person is NOT older than this guess (only 1 chance in 10 of being older)*.
>
> 3. Think of all the reasons suggesting that the person is actually younger than your first guess and make a low guess.
>
> *The low guess is the age for which you are 90% confident that the person is NOT younger than this guess (only 1 chance in 10 of being younger).”*

Auch in dieser Bedingung wurden pro Aufgabe in Phase 2 zwei Seiten präsentiert. Der Unterschied zur EG war, dass auf der ersten Seite zu Schritt 2 und 3 jeweils nach der Ober- und Untergrenze des Konfidenzintervalls gefragt wurde. Die zweite Seite entsprach jedoch der EG.

**Abbildung 9**

*Beispieldurchgang zu Phase 2 der Experimentalgruppe (EG) von Studie 4*




In der KG erhielten die Versuchspersonen lediglich die Instruktion, dass sie in Phase 2 erneut die Bilder aus Phase 1 mit ihren ersten Urteilen präsentiert bekämen und es ihre Aufgabe sei ein zweites Urteil abzugeben. Dieses Urteil könne jeweils das gleiche wie in Phase 1 sein oder hiervon abweichen. Im Anschluss folgten keine detaillierten Instruktionen, sondern lediglich die Instruktionen zu den Bonuszahlungen. Auch hier wurden pro Aufgabe zwei Seiten präsentiert, wobei sie auf der ersten Seite nur ihre Schätzung sahen und keine weiteren Angaben machen konnten. Auf der zweiten Seite sahen sie erneut ihr erstes Urteil

und konnten zusätzlich ihr zweites Urteil abgeben mit dem Hinweis, dass sie ihr erstes Urteil korrigieren oder erneut eingeben könnten.

Nach Abgabe des letzten Urteils folgte in allen Bedingungen ein optionaler demographischer Fragebogen sowie eine kurze Aufklärung über den erreichten Bonus und ein kurzes schriftliches Debriefing. Genau wie in Studie 3 hatte die Versuchsperson zudem die Möglichkeit, in einem freien Textfeld Kommentare zu hinterlassen, sofern sie dies wünschten. Die Versuchspersonen wurden automatisiert für ihre Teilnahme ausgezahlt und erhielten nach der kompletten Erhebung ihren leistungsabhängigen Bonus als zusätzliche Zahlung.

### 5.1.3. Abhängige Maße

**Richtungsurteile.** Zur Untersuchung der Richtungsurteile in der EG und EG + KI wurden in diesem Abschnitt die gleichen Kennwerte wie in Abschnitt 4 berechnet (DJR, TDR, und CDR). Zur Kodierung der CDRs wurde der Populationsmittelwert anhand aller Urteile aus Phase 1 zu dem jeweiligen Bild berechnet, da sich die drei Versuchsbedingungen zu diesem Zeitpunkt noch nicht unterschieden. Diese geschätzten Populationsmittelwerte wurden vor der Berechnung der CDRs gerundet.

**Urteilsverbesserung**. Zur Operationalisierung der Urteilsverbesserung wurde ein Kennwert präregistriert (siehe P4), der auf den absoluten Fehlern (*Absolute Error,* AE) der quantitativen Urteile basiert:

$$AE = |Urteil - wahrer\ Wert| \quad (1)$$

Hierfür wurde für jeden Stimulus pro Person zunächst die Differenz der Fehler der Urteile aus Phase 1 und 2 gebildet:

$$\Delta_{AE} = AE_{Phase\ 1} - AE_{Phase\ 2} \quad (2)$$

Zuletzt wurden diese Differenzen über alle Stimuli pro Person gemittelt. Dieser Wert indiziert bei einem positiven Wert die mittlere Verbesserung pro Person, weshalb er auch als *Improvement* (IMP) bezeichnet wird. Der IMP Score kann auch negativ sein und damit Verschlechterungen indizieren.

**Urteilsänderungen.** Als Maße für Urteilsänderungen wurde berechnet, wie häufig Menschen ihre ersten Urteile ändern und wie stark diese Anpassung ausfällt, sofern sich Menschen entscheiden, eine Korrektur vorzunehmen. Diese Kennwerte werden jeweils als *Adjustment Rate* (AR) und *Opinion Shift* (OS) bezeichnet (vgl. Schultze et al., 2015). Die AR gibt lediglich an, wie häufig eine Person ihr erstes Urteil geändert hat und kann damit Werte zwischen 0 und 1 annehmen. Der OS wird hingegen nur auf Basis der Beobachtungen berechnet, bei denen sich die Versuchspersonen entschieden, ihre Urteile anzupassen. Hierfür wurde lediglich die mittlere absolute Differenz zwischen erstem und zweitem Urteil berechnet, wobei alle Beobachtungen ausgeschlossen wurden, bei denen diese gleich null ist.

### 5.1.4. Ergebnisse & Diskussion

**Vorabanalysen.** Vor dem Test des Weisheitsbewusstseins und der Intervention wurde eine einfache deskriptive Vorabanalyse durchgeführt, um zentrale Vorrausetzungen für den Erfolg der Intervention zu prüfen. Eine wichtige Voraussetzung für den Erfolg der Intervention war, dass Menschen sich häufig für ein Richtungsurteil entschieden und sich bei der Korrektur ihrer Urteile auch entsprechend ihrer Richtungsurteile verhielten. Ersteres war mit DJRs von $M$ = .70 ($SD$ = .27) in der EG und $M$ = .65 ($SD$ = .27) in der EG + KI der Fall. Drei Personen in der EG und fünf Personen in der EG + KI wählten in keinem der Durchgänge ein explizites Richtungsurteil und wurden bei der Analyse der TDRs und CDRs ausgeschlossen, jedoch nicht bei der Analyse der Urteilsverbesserungen. In der EG änderten die Versuchspersonen in 94% der Fälle, in denen sie ein explizites Richtungsurteil wählten, auch ihr zweites Urteil in die angegebene Richtung. In der EG + KI geschah dies in 90% der Fälle. Wurde hingegen keine explizite Richtung gewählt, passten die Versuchspersonen ihr zweites Urteil immer noch vergleichsweise häufig an. Das heißt konkret, dass in 39% der „weder noch“-Fälle der EG und in 46% der „weder noch“-Fälle der EG + KI ein zweites quantitatives Urteil gewählt wurde, welches nicht mit dem ersten übereinstimmte.

**Weisheitsbewusstsein.** Die zentrale Vorrausetzung für einen Erfolg der Intervention ist, dass Menschen beim gegebenen Aufgabenmaterial ein Weisheitsbewusstsein ausbilden und die Richtung des wahren Werts systematisch einschätzen können. Hierfür fand sich in dieser Studie jedoch keine hinreichende Evidenz. Die TDR lag sowohl beim einseitigen Test in

der EG (*M* = .52, *SD* = .17), *t*(174) = 1.63, *p* = .052 (einseitig; P4), *d* = 0.12, als auch beim zweiseitigen Test in der EG + KI (*M* = .53, *SD* = .19), *t*(167) = 1.97, *p* = .052 (zweiseitig; P4), *d* = 0.15, deskriptiv nur knapp über .50. Nach der Präregistrierung kann für diese Studie formal die H1 nicht angenommen werden, für deren Test allein der einseitige Test der TDR in der EG relevant war (siehe P4). In beiden Bedingungen zeigte sich deskriptiv ein Trend in die richtige Richtung, der jedoch deutlich unter dem Durchschnitt der bisherigen Effektstärken zum Weisheitsbewusstsein lag.

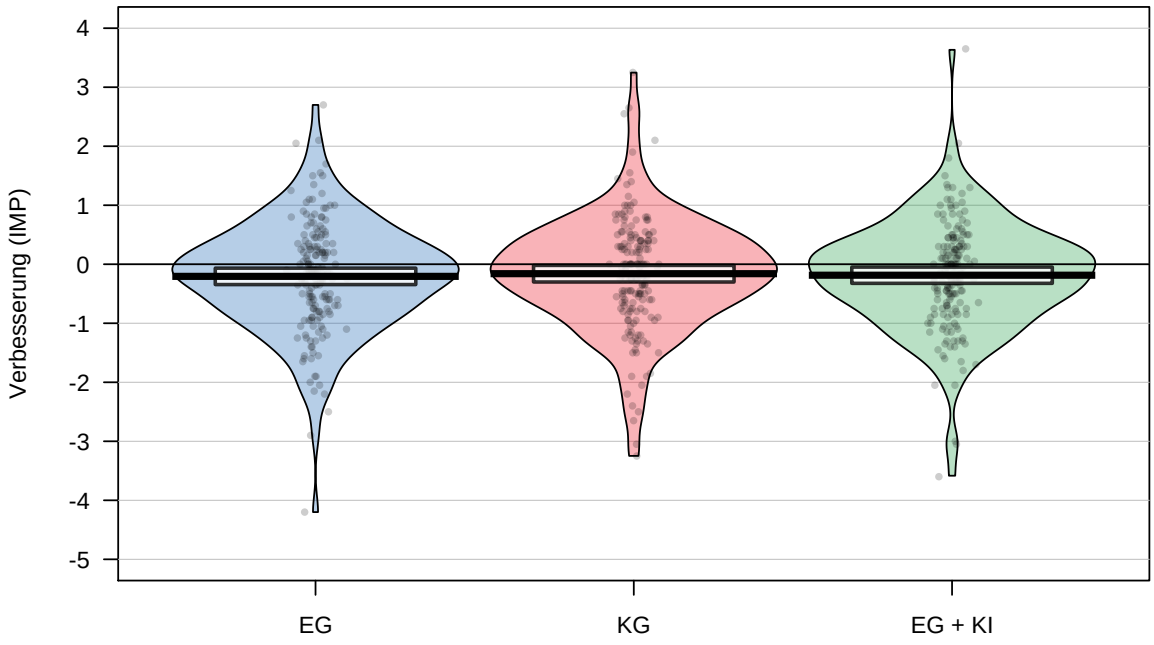
Nach dem GSM könnte ein Grund hierfür sein, dass bei diesen Aufgaben der Pool geteilter Informationen verzerrt ist, was sich auf alle Urteile auf der untersten Ebene des GSM auswirkt. In diesem Fall müsste sich jedoch zumindest zeigen, dass Menschen mit ihren expliziten Richtungsurteilen überzufällig häufig in Richtung des Populationsmittelwert tendieren (H2a). Dies war auch der Fall. Im Einklang mit H2a lagen die CDRs sowohl in der EG mit *M* = .55 (*SD* = .17) signifikant über .50, *t*(174) = 3.96, *p* < .001 (zweiseitig), *d* = 0.30, als auch in der EG + KI mit *M* = .60 (*SD* = .20), *t*(167) = 6.74, *p* < .001 (zweiseitig), *d* = 0.52. Somit zeigte sich in dieser Studie in beiden Experimentalbedingungen Evidenz für H2a. Obwohl sich kein deutliches Weisheitsbewusstsein zeigte, sind die Daten daher konsistent mit dem GSM. Doch auch die CDRs lagen zumindest in der EG unter dem Niveau der vorherigen Studien. Ein Grund könnte wie bei den Prozentschätzungen sein, dass die Versuchspersonen bei den Altersschätzungen vergleichsweise akkurate Urteile abgaben. So stimmten 5% der Urteile aus Phase 1 mit dem wahren Wert und 6% mit dem Populationsmittelwert genau überein, bei denen alle expliziten Richtungsurteile immer als in der falschen Richtung mit 0 kodiert werden. Insgesamt fehlt mit einem hinreichenden Weisheitsbewusstsein eine zentrale Voraussetzung für das Gelingen der Intervention. Im Sinne einer vollständigen Darstellung wird die Wirksamkeit der Intervention dennoch im nächsten Absatz anhand der Verbesserung beziehungsweise Verschlechterung der Urteile direkt getestet.

**Urteilsverbesserung.** Als erstes wurde die Hypothese getestet, dass Menschen mit der Intervention ihre Urteile von Phase 1 zu Phase 2 verbessern können (H8a). Für den Test dieser Hypothese wurde ein einseitiger Einstichproben-*t*-Test der IMP Scores in der EG gegen null präregistriert (siehe P4). Mit einem IMP Score von *M* = -0.21 (SD = 0.94) zeigte sich bei diesem Test keine Evidenz für eine Urteilsverbesserung, *t*(177) = -2.91, *p* = .998 (einseitig; P4), *d* = -0.22, und in einem explorativen zweiseitigen Test sogar ein deutlicher

Trend in die entgegengesetzte Richtung, $p$ = .004 (zweiseitig). Die Urteile in Phase 2 der EG fielen im Schnitt etwas schlechter aus als die in Phase 1. Auch in den anderen beiden Bedingungen lag eine Verschlechterung von Phase 1 zu Phase 2 vor (siehe auch Abbildung 10). Die IMP Scores in der EG + KI ($M$ = -0.19, $SD$ = 0.92) lagen im Schnitt unter 0, $t$(172) = -2.66, $p$ = .008 (zweiseitig; P4), $d$ = -0.20, und gleiches galt für die KG ($M$ = -0.16, $SD$ = 0.96), $t$(179) = -2.24, $p$ = .026 (zweiseitig; P4), $d$ = -0.17. Damit zeigte sich keine Evidenz für H8a. Somit kann auch H8b formal nicht getestet werden, weil sich keine Verbesserung in der EG zeigte. Im Sinne einer vollständigen Darstellung wurde jedoch der hierfür präregistrierte einseitige Welch‘s $t$-Test gerechnet (siehe P4). Es zeigten sich keine höheren IMP Scores in der EG als in der KG, $t$(355.99) = -0.45, $p$ = .674 (einseitig; P4), $p$ = .651 (zweiseitig), $d$ = -0.05, auch deskriptiv waren diese auf einem ähnlichen Niveau (siehe Abbildung 10). Auch bei den

**Abbildung 10**

*Urteilsverbesserung (IMP) der Bedingungen von Studie 4*




*Anmerkung*. Abbildung der mittleren Verbesserungen (IMP) pro Bedingung (schwarze Linien) mit 95% Konfidenzintervalle (weiße Rechtecke), Verteilungsdichte („Geigenplots“) und Darstellung einzelner Versuchspersonen (schwarze Punkte). EG = Intervention/Experimentalgruppe. KG = Kontrollgruppe. EG + KI = Intervention + Konfidenzintervall. IMP = Improvement Score auf Basis der absoluten Fehlerdifferenzen.

beiden anderen möglichen Bedingungsvergleichen zeigten sich keine Unterschiede in den IMP Scores, beide *p*s > .796 (zweiseitig; P4). Insgesamt war die Intervention damit weder im Innersubjektvergleich noch im Zwischensubjektvergleich wirksam; es kam insgesamt eher zu einer Verschlechterung, die jedoch nicht für die Interventionsbedingung spezifisch war.

**Urteilsänderungen.** Das Ausbleiben positiver Effekte der Intervention war angesichts des nicht hinreichend ausgeprägten Weisheitsbewusstseins erwartbar, da sich zumindest eine notwendige Voraussetzung für den Erfolg der Intervention nicht, beziehungsweise nicht ausreichend, zeigte. Unabhängig hiervon stellt sich jedoch die Frage, ob die Intervention zumindest Menschen anregt, ihre Urteile häufiger und zudem stärker zu korrigieren, als sie es ohne Intervention tun würden. Um dies zu testen, wurde im Folgenden zunächst die relative Häufigkeit angepasster Urteile verglichen. Hierbei zeigte sich, dass die Versuchspersonen in den beiden Experimentalgruppen ihre Urteile häufiger änderten als in der KG. Konkret passten die Versuchspersonen in der EG ihre Urteile mit einer Adjustment Rate (AR) von *M* = .81 (*SD* = .24) häufiger an als in der KG (*M* = .69, *SD* = .32), *t*(334.24) = 4.17, *p* < .001 (zweiseitig), *d* = 0.44. Zudem passten die Versuchspersonen ihre Urteile in der EG +KI (*M* = .78, *SD* = .26) häufiger an als in der KG, *t*(342.73) = 2.82, *p* = .005 (zweiseitig), *d* = 0.30, während die AR in der EG + KI ähnlich groß ausfiel wie in der EG, *t*(345.27) = 1.39, *p* = .166 (zweiseitig), *d* = 0.15. Beim Vergleich der Opinion Shifts zeigte sich ein anderes Muster. Hier lag das absolute Ausmaß der Korrekturen, sofern sie vorgenommen wurden, bei EG (*M* = 3.85, *SD* = 1.67) und KG (*M* = 3.84, *SD* = 1.56) auf einem ähnlichen Niveau, *t*(342.65) = 0.04, *p* = .970 (zweiseitig), *d* = 0.004. In der EG + KI (*M* = 3.20, *SD* = 1.16) fielen die Opinion Shifts jedoch geringer aus als in der EG, *t*(313.13) = -4.17, *p* < .001 (zweiseitig), *d* = -0.45, und der KG, *t*(311.13) = -4.27, *p* < .001 (zweiseitig), *d* = -0.47. Somit regte die Intervention die Versuchspersonen zumindest an, häufiger Änderungen an ihren Urteilen vorzunehmen, die jedoch nicht unbedingt größer ausfielen. Dieser Befund deutet zumindest an, dass die Intervention funktionieren könnte, sofern Menschen mit den Anpassungen ihrer Schätzungen (ganz im Sinne eines Weisheitsbewusstseins) auch überzufällig richtigliegen. In dieser Studie konnte hierdurch jedoch keine systematische Verbesserung erzielt werden, da das Weisheitsbewusstsein nur schwach ausgeprägt war. Entsprechend bedarf es eines erneuten Tests der Intervention mit Aufgaben, bei denen sich auch systematisch ein Weisheitsbewusstsein ausbilden kann.

## 5.2. Studie 5

Ziel von Studie 5 war es, die Intervention zur Verbesserung der Urteile unter verbesserten Voraussetzungen zu testen. Hierfür wurden in Studie 5 gezielt Aufgaben gewählt, bei denen sich das Weisheitsbewusstsein bereits in den vorherigen Studien deutlich zeigte. Konkret wurden 30 von den 45 Schätzaufgaben zum Thema Globalisierung aus Studie 2 verwendet. Im Unterschied zu Studie 4 wurde auf eine Experimentalgruppe mit Konfidenzintervallen verzichtet.

### 5.2.1. Design und Stichprobe

Für Studie 5 wurde ein einfaktorielles Design mit einem Zwischensubjektfaktor gewählt, welcher jedoch nur mit zwei Stufen manipuliert wurde: Intervention/Experimentalgruppe (EG) vs. Kontrollgruppe (KG). Die Testplanung richtete sich wieder nach dem Test von H1 und H8a und H8b (siehe P5), wobei mit einer statistischen Power von mindestens $1 - \beta = .80$ für alle Hypothesen geplant wurde, da für die teurere Laboruntersuchung keine zusätzlichen Ressourcen zur Verfügung standen (siehe Abschnitt 3). Es wurden die gleichen Effekte bei einseitiger Testung wie bei der Poweranalyse für Studie 4 angenommen ($d = 0.43$ für H1 und $d = 0.30$ für H8a und H8b; siehe P5). Dabei ergab sich bei einer Zellbesetzung von $n = 140$ eine statistische Power von $1 - \beta > .99$ für H1, $1 - \beta = .97$ für H8a und $1 - \beta = .80$ für H8b. Somit wurde eine Stichprobengröße von $N = 280$ (140 pro Zelle) angestrebt. Als Ausschlusskriterium wurde präregistriert, dass die Versuchspersonen die Studie vollständig (bis zur letzten Seite) bearbeiten müssen, um bei der Studie berücksichtigt zu werden, und dass zusätzliche Versuchspersonen erhoben werden, sobald sich aus den Ausschlüssen eine Zellbesetzung kleiner $n = 140$ ergibt.

Die Studie wurde genau wie die anderen Laborstudien an der Universität Göttingen durchgeführt. Bei der Erhebung der Studie kam es in sechs Fällen zu technischen Problemen, aufgrund derer die Studiendurchführung für diese Personen nicht beendet werden konnte. Diese Fälle wurden durch sechs zusätzliche Versuchspersonen ersetzt, wodurch sich eine finale Stichprobengröße von $N = 280$ (165 Frauen, 131 Männer, eine nicht-binär/divers, eine ohne Angabe) mit einem Durchschnittsalter von $M = 25.03$ Jahren ($Mdn = 24$, $SD = 5.96$) ergab. Für die Bearbeitung der Studie erhielten die Versuchspersonen eine Vergütung von 4€

oder eine halbe Versuchspersonenstunde und zusätzlich einen leistungsabhängigen Bonus von bis zu 6€ (*M* = 1.06, *SD* = 0.38).

### 5.2.2. Methode

Die Methodik von Studie 5 entsprach konzeptuell der Methodik von EG und KG von Studie 4. Der wesentliche Unterschied bestand darin, dass als Aufgabematerial die jeweils 15 historischen Schätzaufgaben und die 15 bei null begrenzten Aufgaben aus Studie 2 verwendet wurden, bei denen sich bereits ein deutliches Weisheitsbewusstsein gezeigt hatte. Da die Studie jedoch unter Laborbedingungen im deutschen Sprachraum durchgeführt wurde, wurden (im Vergleich zu Studie 4) mehrere zusätzliche Anpassungen vorgenommen, die sich an den früheren Studien im Labor orientierten (Studie 1 und 2) und daher nur kurz zusammengefasst werden. Ein Unterschied bestand darin, dass die Versuchspersonen bereits zu Beginn den unterschiedlichen Bedingungen zugeteilt wurden, was sich jedoch erst auf den Ablauf in Phase 2 auswirkte. Die Zuteilung erfolgte beim Start randomisiert durch das Experimentalprogramm, wobei durch das Programm eine gleichmäßige Zellbesetzung sichergestellt wurde. Datensätze, die aufgrund technischer Probleme nicht vollständig erhoben werden konnten, wurden bei späteren Zuteilungen nicht berücksichtigt. Auch in dieser Studie wurde den Versuchspersonen nach Abgabe ihres Einverständnisses eine Extraseite mit der Information präsentiert, bei der Bearbeitung nur die Navigationsfunktionen des Experiments zu benutzen, worauf die Erläuterungen zum Ablauf und den Bonuszahlungen folgten. Die Berechnung der Boni entsprach für Phase 1 den früheren Studien mit diesen Aufgabentypen, das heißt, dass die Versuchspersonen 10 Cent für historische Schätzungen erhielten, die nicht mehr als 10 Jahre vom wahren Wert abwichen, und 10 Cent für alle andere Schätzungen, deren prozentualer Fehler 10 Cent nicht überstieg. Bei der Abgabe der ersten Urteile zu Phase 1 waren alle historischen Schätzungen nach oben und unten unbegrenzt und alle anderen Aufgaben bei null begrenzt.

Der Ablauf von Phase 2 unterschied sich zwischen den beiden Bedingungen. In der EG erhielten die Versuchspersonen die kurze Information, dass sie im Folgenden erneut die Aufgaben aus Phase 1 mit ihren ersten Antworten präsentiert bekämen und es ihre Aufgabe sei, erneut möglichst akkurate Schätzungen abzugeben, wobei sie hierfür auf der nächsten Seite detaillierte Informationen erhalten würden. Diese Instruktionen waren weitestgehend

deckungsgleich mit den Instruktionen zur Abfrage des Weisheitsbewusstseins in Studie 1 und 2 und wurden für mindestens 60 Sekunden angezeigt. Genau wie in Studie 4 wurde jedoch ein weiterer Schritt präsentiert, bei dem die Versuchspersonen ihre erste Schätzung anpassen sollten, bis sie das Gefühl hätten, dass der wahre Wert weder eher in der einen noch eher in der anderen Richtung liege. Die vollständigen Instruktionen lauteten wie folgt:

> „In Phase 2 sollen Sie für Ihre ersten Schätzungen angeben, ob Sie den wahren Wert eher über, eher unter oder weder eher über noch eher unter der eigenen Schätzung vermuten. Anschließend ist es Ihre Aufgabe, eine zweite Schätzung abzugeben, bei der Sie das Gefühl haben, dass der wahre Wert weder über noch unter Ihrer ersten Schätzung liegt.
>
> **Bitte folgen Sie bei jeder Aufgabe in Phase 2 immer den folgenden Schritten zur Beantwortung der Frage:**
>
> 1. Stellen Sie sich vor, Ihr Urteil weicht vom wahren Wert ab.
>
> 2. Überlegen Sie, was dafürspricht, dass der wahre Wert über Ihrer Schätzung liegt, und was dafürspricht, dass er unter Ihrer Schätzung liegt.
>
> 3. Entscheiden Sie sich für eine der beiden Richtungen auf Basis Ihrer Überlegungen oder geben Sie als dritte Option an, dass Sie den wahren Wert in keiner der beiden Richtungen eher vermuten.
>
> 4. Geben Sie eine neue Schätzung ab, bei der Sie das Gefühl haben, dass der wahre Wert weder über noch unter Ihrer Schätzung liegt.“

Der Ablauf der KG und der weitere Verlauf der EG inklusive der Instruktionen zu den Bonuszahlungen und dem Ablauf von Phase 2 glichen konzeptuell dem Ablauf von Studie 4, wobei die Versuchspersonen in der KG direkt ihre zweiten Urteile abgeben konnten, ohne dass hierfür zunächst nur die Aufgabe und ihre erste Schätzung auf einer eigenen Seite präsentiert wurde. Nach Abschluss von Phase 2 endete das Experiment wie in allen Laborstudien mit einem optionalen Abschlussfragebogen, einer schriftlichen Aufklärung über Ziel und Bonus, sowie einer Dokumentation der Teilnahme auf einem separaten Server.

### 5.2.3. Abhängige Maße

Als abhängige Maße wurden die gleichen Maße wie in Studie 4 verwendet, wobei die Maße teilweise standardisiert wurden, da sich Varianz und Mittelwerte der absoluten Fehler und Korrekturmaße zwischen den Stimuli teilweise deutlich unterschieden. Für die Opinion Shifts wurde eine einfache *z*-Standardisierung pro Stimulus gewählt. Für die IMP Scores wurde zur Standardisierung ein ähnliches Vorgehen gewählt, wobei die jeweiligen IMP Scores anders als bei einer *z*-Standardisierung nur durch ihre Standardabweichung geteilt wurden und nicht von ihrem Mittelwert über alle Beobachtungen pro Stimulus subtrahiert wurden (siehe P5). Somit konnte pro Person ein Kennwert berechnet werden, bei dem positive und negative Vorzeichen weiterhin Verbesserungen und Verschlechterungen indizieren (anders als bei einer *z*-Standardisierung), wobei das Ausmaß dieser Verbesserungen und Verschlechterungen über die Stimuli hinweg standardisiert wurde.

### 5.2.4. Ergebnisse & Diskussion

**Vorabanalysen.** Genau wie in den vorherigen Studien entschieden sich die Versuchspersonen der EG mit einer DJR von $M$ = .66 ($SD$ = .21) häufig für eines der beiden expliziten Richtungsurteile und alle Versuchspersonen gaben mindestens ein explizites Richtungsurteil ab. Sofern sie sich für ein Richtungsurteil entschieden, korrigierten sie dies auch in 97% der Fälle in die angegebene Richtung. Im Gegenzug änderten die Proband:innen nur in 11% der Fälle ihr erstes Urteil, wenn sie vorher keine Richtung beziehungsweise die „weder noch“-Option wählten. Somit verhielten sich die Versuchspersonen in der EG größtenteils konsistent mit ihren Richtungsurteilen. Diese Voraussetzung für einen möglichen Erfolg der Intervention war somit gegeben.

**Weisheitsbewusstsein**. Im Einklang mit den bisherigen Studien, die die gleichen Aufgaben verwendeten, zeigte sich im Test der TDR von $M$ = .59 ($SD$ = .14) gegen .50, dass die Versuchspersonen mit ihren expliziten Richtungsurteilen überzufällig häufig in der Lage waren, die Richtung des wahren Werts korrekt einzuschätzen, $t(139) = 7.62$, $p < .001$ (einseitig; P5), $d$ = 0.64. Auch die CDR lag mit $M$ = .61 ($SD$ = .15) signifikant über .50, $t(139) = 8.14$, $p < .001$ (zweiseitig), $d$ = 0.69. Dieses Muster ließ sich auch für die einzelnen Aufgabenkategorien replizieren (siehe Tabelle 2). Hiermit zeigte sich konsistent Evidenz für

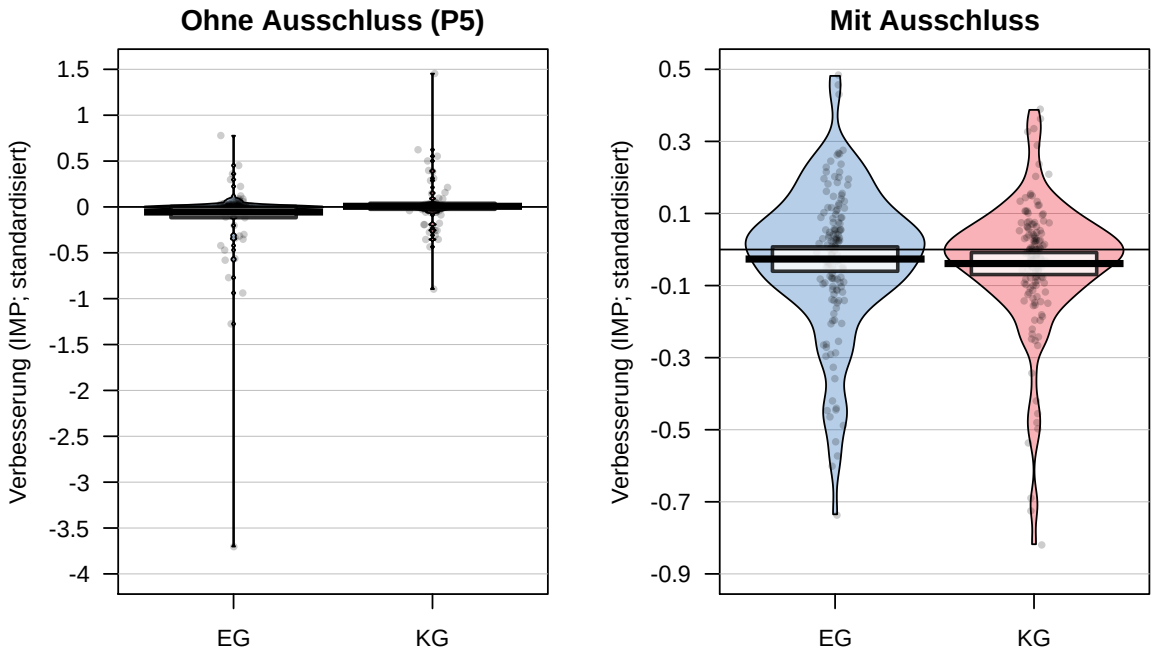
H1 und H2a, und die zentrale Voraussetzung für den Erfolg der Intervention war, anders als in Studie 4, klar gegeben.

**Urteilsverbesserung**. Für den Test der Intervention wurden zwei *t*-Tests für den neuen (standardisierten) IMP Score präregistriert, die den präregistrierten Tests von H8a und H8b in der vorherigen Studie entsprachen (siehe P5). Es zeigte sich in der EG erneut, dass die Versuchspersonen sich mit einem IMP Score von $M$ = -0.06 ($SD$ = 0.37) im einseitigen Einstichproben-*t*-Test gegen null nicht verbesserten, $t(139) = -1.76$, $p = .960$ (einseitig; P5), $d$ = -0.15. Deskriptiv fiel der IMP Score auch in der EG von Studie 5 negativ aus, $p = .080$ (zweiseitig). H8a musste entsprechend auch in dieser Studie abgelehnt werden. Es zeigte sich in der explorativen Folgeuntersuchung auch kein statistisch signifikanter Unterschied des standardisierten IMP Scores in der KG von null ($M$ = 0.01, $SD$ = 0.20), $t(139) = 0.41$, $p = .684$ (zweiseitig), $d$ = 0.03. Aufgrund der fehlenden Verbesserung in der EG lässt sich H8b auch hier nicht sinnvoll testen. Doch der entsprechende Test erlaubt zumindest Aussagen darüber, ob die die Verschlechterung in der EG kleiner ausfällt als in der KG. Dies war aber nicht der Fall, $t(211.74) = -1.75$, $p = .959$ (einseitig; P5), $p = .082$ (zweiseitig), $d$ = -0.21. Bei einer deskriptiven Analyse der IMP Scores zeigte sich jedoch, dass es vergleichsweise häufig statistische Ausreißer gab, die einen starken Einfluss auf die Analysen gehabt haben könnten und somit dafür gesorgt haben könnten, dass sich keine Evidenz für H8a und H8b zeigte (siehe Abbildung 11). Um diese Möglichkeit auszuschließen, wurden weitere Tests von H8a und H8b durchgeführt, bei denen zunächst die standardisierten Fehlerdifferenzen für alle Beobachtungen herangezogen wurden. Die 1% extremsten positiven und 1% extremsten negativen Durchgänge pro Bedingung wurden ausgeschlossen und die IMP Scores danach neu berechnet (siehe rechte Bildhälfte in Abbildung 11). Auch hier zeigte sich keinerlei Evidenz für H8a und H8b, beide $p$s > .125 (zweiseitig), sondern deskriptiv eine Verschlechterung der Urteile in der EG (siehe Abbildung 11). Auch eine separate Analyse der beiden Aufgabenkategorien führte mit und ohne Ausschluss qualitativ zu den gleichen Ergebnissen. Es zeigte sich keine Verbesserung in der EG oder KG in diesen Folgeuntersuchen pro Aufgabenkategorie mit und ohne Ausschluss, alle $d$s < 0.15. Insgesamt liefern die präregistrierten Analysen und die Folgeuntersuchung ein eindeutiges Bild bezüglich der Wirksamkeit der Intervention: Genau wie in Studie 4 zeigte sich weder im Innersubjektvergleich noch im Zwischensubjektvergleich ein Akkuratheitsgewinn durch die

Intervention, und zwar obwohl die Versuchspersonen im Gegensatz zu denen in Studie 4 über ein deutliches Weisheitsbewusstsein verfügten.

**Abbildung 11**

*Urteilsverbesserung (IMP) in Studie 5 mit und ohne Ausschluss extremer Werte*




*Anmerkung*. Mittlere Verbesserung (IMP) der Urteile in Studie 5 mit und ohne Ausschluss extremer Verbesserungen und Verschlechterungen. Beim Ausschluss wurden pro Bedingung je 1% der extremsten positiven und 1% der extremsten negativen standardisierten Fehlerdifferenzen ausgeschlossen (auf Basis der absoluten Fehler) und die standardisierten IMP Scores neu berechnet. Abgebildet sind die Mittelwerte der Verbesserungen (IMP) (schwarze Linien), 95% Konfidenzintervalle (weiße Rechtecke), die Verteilungsdichte („Geigenplots“) und Datenpunkte von einzelnen Versuchspersonen (schwarze Punkte). EG = Intervention/Experimentalgruppe. KG = Kontrollgruppe. IMP = Improvement Score auf Basis der absoluten Fehlerdifferenzen und standardisiert anhand der Standardabweichung der einzelnen Stimuli (mit und ohne Ausschluss).

**Urteilsänderungen.** Um zu verstehen, warum sich trotz Weisheitsbewusstsein kein positiver Effekt der Intervention zeigte, wurden die Urteilsänderungen vertiefend untersucht. Zunächst wurden die Bedingungsunterschiede exploriert. Auch in dieser Studie zeigte sich, dass die Versuchspersonen in der EG mit einer AR von *M* = .69 (*SD* = .22) ihre

Urteile häufiger änderten als in der KG ($M$ = .56, $SD$ = .28), $t$(260.47) = 4.24, $p$ < .001 (zweiseitig), $d$ = 0.51. Genau wie in der vorherigen Studie fielen die Korrekturen, sofern sie vorgenommen wurden, jedoch nicht systematisch größer aus. Die standardisierten Opinion Shifts unterschieden sich auch in dieser Studie nicht zwischen EG ($M$ = -0.003, $SD$ = 0.34) und KG ($M$ = 0.05, $SD$ = 0.59), $t$(218.31) = -0.87, $p$ = .383 (zweiseitig), $d$ = -0.11.[16] Damit wurden die Versuchspersonen wie in Studie 4 durch die Intervention zu Korrekturen angeregt, ohne deren Ausmaß zu beeinflussen, sofern sie überhaupt eine Korrektur vornahmen.

Als nächstes stellt sich die Frage, ob die Versuchspersonen der EG ihre Schätzungen auch häufiger in Richtung des wahren Wertes anpassten. Hierfür wurde für alle Fälle, in denen sich das zweite Urteil vom ersten unterschied, dichotom kodiert, ob das zweite Urteil in die gleiche Richtung zeigte wie der wahre Wert, oder nicht. Zudem wurden Korrekturen als entgegen des wahren Wertes kodiert, wenn wahrer Wert und erstes Urteil genau übereinstimmten (2% der Fälle). Für die EG lag der Anteil richtiger Anpassungen mit $M$ = .59 ($SD$ = .15) auf einem ähnlichen Niveau wie die TDRs der Richtungsurteile und damit auch über .50, $t$(139) = 7.10, $p$ < .001 (zweiseitig), $d$ = 0.60. Dies war zu erwarten, da sich die Versuchspersonen der EG fast immer konsistent mit ihren Richtungsurteilen verhielten. Neue Erkenntnisse liefert der Anteil richtiger Anpassungen in der KG, der mit $M$ = .62 ($SD$ = .17) ähnlich hoch ausfiel wie in der EG, $t$(271.69) = 1.43, $p$ = .154 (zweiseitig), $d$ = 0.17, und zudem auch über .50 lag, $t$(137) = 8.22, $p$ < .001 (zweiseitig), $d$ = 0.70. Damit lagen die Korrekturen in der EG nicht häufiger in der richtigen Richtung als in der KG, auch wenn in der EG insgesamt häufiger Korrekturen vorgenommen wurden. Die Intervention hat in dieser Hinsicht jedoch auch nicht geschadet, da die Korrekturen auch nicht systematisch seltener in die richtige Richtung getätigt wurden.

Eine Erklärung für dieses Muster ist, dass die Menschen auch in der KG ein Weisheitsbewusstsein besitzen. Die geringere Häufigkeit der Anpassungen könnte dann sowohl daran liegen, dass sie das Weisheitsbewusstsein ohne Instruktion seltener ausbilden oder dass sie ohne explizite Aufforderung eine solche Korrektur schlichtweg seltener vornehmen. In jedem Falle stehen die Daten im Einklang mit der Idee, dass Menschen auch

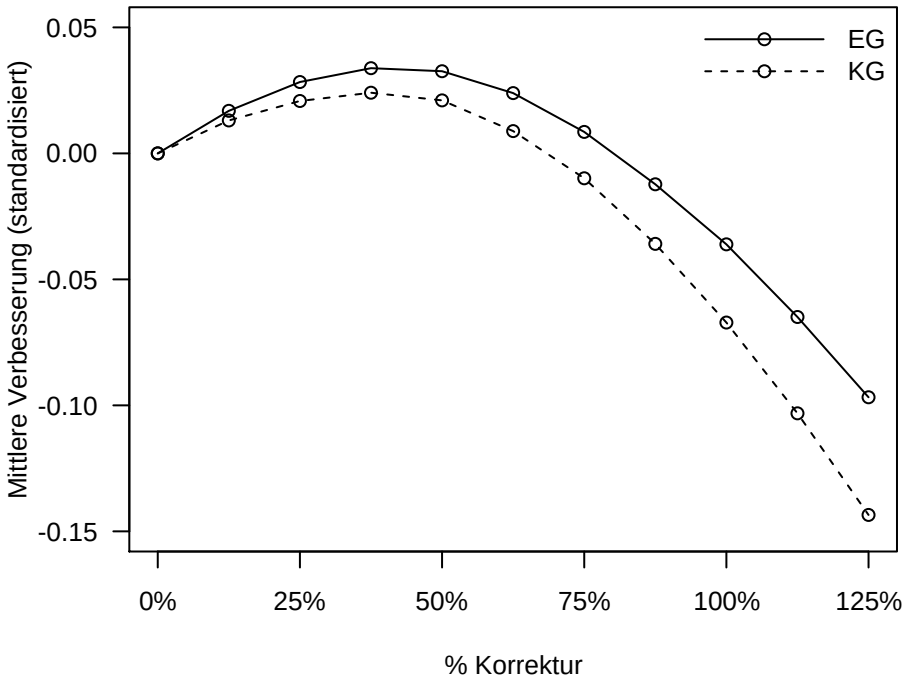
[16] Es zeigten sich in den standardisierten Opinion Shifts auch keine Unterschiede, wenn pro Bedingung vorher jeweils die 1% extremsten standardisierten Verbesserungen und Verschlechterungen pro Bedingung ausgeschlossen wurden, $p$ = .670 (zweiseitig).

spontan (d.h. ohne Abfrage eines Richtungsurteil) ein Weisheitsbewusstsein besitzen. Dass es nun weder in der EG noch in der KG zu Verbesserungen der Urteile kam, legt nahe, dass die metakognitive Bewertung der eigenen Urteile nicht das zentrale Problem ist. Entsprechend könnten also die Korrekturen selbst, das heißt ihr Ausmaß, dysfunktional sein. Zum einen könnten die Bewegungen zu klein sein oder zu selten auftreten, um bedeutsame Verbesserungen herbeizuführen. Zum anderen besteht die Möglichkeit, dass die Korrekturen insgesamt so groß ausfallen, dass Menschen bei Korrekturen in die richtige Richtung über das Ziel hinausschießen, was im extremen Fall sogar zu einer systematischen Verschlechterung führen kann.

Um direkt zu untersuchen, inwiefern die vorgenommenen Korrekturen eher zu klein oder zu groß ausfielen, wurde eine deskriptive Analyse durchgeführt. Mit dieser Analyse wurde näherungsweise untersucht, wie groß die Korrekturen der Urteile hätten sein müssen, damit die Versuchspersonen ihre Urteile am stärksten verbessert hätten. Die Ergebnisse der deskriptiven Analyse sind in Abbildung 12 dargestellt. Hierfür wurden die tatsächlichen Korrekturen herangezogen und schrittweise um einen bestimmten prozentualen Anteil reduziert oder verstärkt. Es wurde untersucht, wie stark sich die Versuchspersonen mit Korrekturen zwischen 0% und 100% oder sogar über 100% hinaus verbessert oder verschlechtert hätten. Referenzpunkte dieser Analyse sind die ersten Urteile (keine Korrektur/0% Korrektur) und die zweiten Urteile der Versuchspersonen (100% Korrektur). Eine Korrektur von 50% entspricht dem hypothetischen Fall, bei dem die tatsächlichen Korrekturen nur halb so stark ausgefallen wären. Die Berechnung dieser künstlichen Korrekturen entspricht einem gewichteten Mittelwert aus erstem Urteil und zweitem Urteil. Der prozentuale Anteil der Korrektur entspricht dem Gewicht auf dem zweiten Urteil im gewichteten Mittel. Fälle, bei denen keine Korrektur vorgenommen wurde (das erste Urteil entspricht dem zweiten), wurden von der Analyse ausgeschlossen (31% in der EG und 44% in der KG; siehe ARs). Zudem wurden bei dieser Analyse 1% der Beobachtungen mit den extremsten standardisierten Verbesserungen und 1% mit den extremsten standardisierten Verschlechterungen pro Bedingung ausgeschlossen, die andernfalls einen zu starken Einfluss auf die deskriptive Analyse hätten. Zuletzt wurden für jede Stufe standardisierte Verbesserungen beziehungsweise Verschlechterungen pro Trial berechnet und gemittelt.

**Abbildung 12**

*Urteilsverbesserung in Studie 5 für theoretische Urteilskorrekturen*




*Anmerkung*. Mittlere Verbesserung auf Basis der Fehlerdifferenzen zwischen den ersten Urteilen (Phase 1) und den stufenweise korrigierten Urteilen. Zur Berechnung des korrigierten Urteils (% Korrektur) wurde ein gewichteter Mittelwert aus dem jeweils ersten empirischen Urteil (Phase 1) und dem zweiten (Phase 2) berechnet. Für die Fehlerdifferenz wurde der absolute Fehler des korrigierten Urteils von dem absoluten Fehler des ersten Urteils subtrahiert. Diese Fehlerdifferenz wurde anhand der Standardabweichung der Fehlerdifferenzen von dem ersten und zweiten empirischen Urteil pro Stimulus standardisiert und über alle Trials gemittelt. Darstellung unter Ausschluss aller Trials ohne Korrektur der Urteile und unter Ausschluss der 1% extremsten positiven und 1% extremsten negativen standardisierten empirischen Fehlerdifferenzen pro Bedingung. EG = Intervention/Experimentalgruppe. KG = Kontrollgruppe.

Diese deskriptive Analyse liefert drei Erkenntnisse. Erstens zeigt sich, dass die zweiten Urteile (100% der Korrektur) tendenziell schlechter ausfielen als die ersten Urteile (0% der Korrektur). Zweitens hätten größere Korrekturen als die tatsächlichen Korrekturen zu noch stärkeren Verschlechterungen geführt (>100% der Korrektur). Drittens hätten zumindest

deskriptiv reduzierte Korrekturen bei 50% oder knapp unter 50% in beiden Bedingungen zu Verbesserungen geführt. Damit wurden tendenziell eher zu starke als zu schwache Korrekturen vorgenommen, sofern sie denn überhaupt vorgenommen wurden.

Zusammengefasst lässt sich sagen, dass Menschen in der Interventionsbedingung häufig ein Weisheitsbewusstsein ausbildeten und sich konsistent hiermit verhielten. Zudem regte die Intervention Menschen zu häufigeren Korrekturen an, die jedoch tendenziell (genau wie in der KG) zu stark ausfielen, um eine systematische Verbesserung zu erzielen. Ziel ist es im Folgenden eine Intervention zu testen, die die Menschen weiterhin zu häufigen Korrekturen in Richtung des Weisheitsbewusstseins anregt. Die Intervention soll jedoch dahingehen angepasst werden, dass das Ausmaß der Korrekturen, sofern sie denn vorgenommen werden, im Vergleich zu spontanen Korrekturen reduziert wird. Einfach ausgedrückt: Menschen sollen mit der angepassten Intervention nur kleine Schritte in die (häufig) richtige Richtung vornehmen und damit ihre Urteile im Schnitt verbessern.

### 5.3. Studie 6

Mit Studie 6 wurde auf Basis der Erkenntnisse aus Studie 5 eine neue, angepasste Intervention entwickelt, die grundsätzlich den Schritten der ursprünglichen Intervention entsprach. Der zentrale Unterschied dieser angepassten Intervention lag im letzten Schritt, bei dem die Menschen in der neuen Intervention aufgefordert wurden, ihre Urteile nur *einen kleinen Schritt* in die von ihnen eingeschätzte Richtung anzupassen. Die Annahme war, dass dadurch überschießende Korrekturen in die (häufig) richtige Richtung vermieden werden und auch extreme Korrekturen in die (teilweise) falsche Richtung reduziert werden. Die Korrekturen sollten daher im Schnitt zu einer Verbesserung der Urteile führen und zudem im Schnitt größer ausfallen als in einer Kontrollgruppe ohne Intervention. Hieraus ergaben sich neue Hypothesen, die strukturell den Hypothesen H8a und H8b entsprechen, sich jedoch auf eine neue (angepasste) Intervention beziehen:

> **Hypothese 9a (H9a; H2a in P6)**: Menschen sind in der Lage, mit einer Intervention die Akkuratheit ihrer Urteile zu verbessern, die Menschen dazu instruiert ihre Urteile *einen kleinen Schritt* in die Richtung anzupassen, in der sie den wahren Wert eher vermuten.

**Hypothese 9b (H9b; H2b in P6)**: Die (angepasste) Intervention zur Verbesserung der Urteile (H9a) führt zu einer stärkeren Verbesserung als in einer Kontrollgruppe ohne Intervention.

Zum Test dieser Hypothesen wurde eine weitere Studie durchgeführt, die (abgesehen von der veränderten Intervention) der vorherigen Studie entsprach, wobei in dieser Studie weitere Maßnahmen ergriffen wurden, um extreme Urteile direkt bei der Eingabe zu verhindern.

### 5.3.1. Design und Stichprobe

Das Untersuchungsdesign und die Stichprobenplanung entsprachen bei Studie 6 weitestgehend der vorherigen Studie. Der zentrale Unterschied war, dass in dem einfaktoriellen Zwischensubjektdesign die Experimentalbedingung gegen eine neue Experimentalbedingung ausgetauscht wurde, bei der die Versuchspersonen aufgefordert wurden, ihre Urteile lediglich einen kleinen Schritt in die angegebene Richtung zu korrigieren. Bei der Stichprobenplanung wurde die gleiche Poweranalyse wie in Studie 5 für die neuen Hypothesen angesetzt (mit H9a und H9b anstelle von H8a und H8b), wodurch sich eine Stichprobengröße von $N$ = 280 (140 pro Zelle) ergab (siehe P6). Auch die Ausschlusskriterien waren die gleichen wie in Studie 5 (siehe P6). Bei der Studie kam es in fünf Fällen zu technischen Problemen, wodurch die Datenerhebung für diese Personen nicht beendet werden konnte. Diese Datensätze wurden ausgeschlossen und durch Daten von fünf zusätzlichen Versuchspersonen ersetzt. Die finale Stichprobengröße betrug damit $N$ = 280 (163 Frauen, 116 Männer, eine ohne Angabe) mit je 140 Personen in den beiden Versuchsbedingungen. Hierbei handelte es sich überwiegend um Studierende im Alter von $M$ = 23.42 Jahren ($Mdn$ = 22, $SD$ = 5.01). Bei gleichem Vergütungssystem wie in Studie 5 erhielten die Versuchspersonen einen leistungsabhängigen Bonus von $M$ = 1.06 Euro ($SD$ = 0.41).

### 5.3.2. Methode

Methode und Ablauf von Studie 6 entsprach weitestgehend der Methodik von Studie 5 mit zwei zentralen Unterschieden. Der erste Unterschied lag darin, dass die Eingaben der quantitativen Urteile zu Phase 1 und 2 in beiden Bedingungen bei allen Aufgaben durch das

Experimentalprogramm nach oben und unten begrenzt waren, um anders als bei Studie 5 extreme Urteile und Fehler direkt bei der Eingabe zu verhindern. Hierfür wurden die quantitativen Urteile zu Phase 1 aus den vorherigen Studien mit dem (teilweise) gleichen Aufgabenmaterial (Studie 1, 2, & 5) analysiert (für eine detaillierte Übersicht siehe Anhang A). Ziel war es Grenzen zu finden, die extreme Fehleingaben verhindern, ohne dabei plausible Eingaben der Versuchspersonen zu verhindern. Bei den historischen Schätzungen waren hierfür die Urteile auf einen Bereich zwischen 1700 und 2019 begrenzt, in dem alle wahren Werte (1835-2001) und mindestens 95% aller Urteile aus den früheren Studien lagen. Für die bei null begrenzten Aufgaben wurden wieder alle Urteile nach unten bei null begrenzt. Nach oben wurde jedoch ein variabler Wert auf Basis vergangener Urteile gesetzt. Als Obergrenze wurde das 0.975-Quantil gewählt, also der Wert, bei dem nur 2.5% der Urteile aus den vorherigen Studien größer als dieser Grenzwert waren.[17]

Der zweite zentrale Unterschied zwischen Studie 5 und 6 lag in der Instruktion der neuen Intervention zu Phase 2. Nach Phase 1 erhielten die Versuchspersonen auch in der neuen EG die Instruktion, im Folgenden neue und möglichst akkurate Urteile abzugeben, worauf die detaillierteren Interventionsinstruktionen folgten. Diese entsprachen weitestgehend den detaillierten Instruktionen in der EG von Studie 5, wobei der letzte Schritt durch folgenden ersetzt wurde:

> „4. Geben Sie eine neue Schätzung ab. Bewegen Sie sich hierfür bitte ausgehend von Ihrer ersten Schätzung **einen kleinen Schritt** in die von Ihnen angegebene Richtung. Diese Bewegung sollte nur erfolgen, falls Sie sich für eine der beiden Richtungen entschieden haben."

Anders als in den ersten Laborstudien wurden diese Instruktionen Schritt für Schritt präsentiert und die minimale Darbietungszeit wurde nicht auf mindestens 60 Sekunden gesetzt. Wurde in Phase 2 keine Richtung gewählt, so erfolgte lediglich die Aufforderung

[17] Für fast alle Fragen lag dieser Grenzwert deutlich über dem wahren Wert. Die einzige Ausnahme war eine Frage („Wie viele Handelshäfen befinden sich in den Niederlanden?"), bei der der wahre Wert von nahezu allen Versuchspersonen unterschätzt wurde und somit 97.5% der Urteile unter dem wahren Wert lagen. Für eine konsistente Begrenzung wurde diese Frage mit einer neuen ausgetauscht („Was ist die Luftlinienentfernung zwischen London und Amsterdam?"), von der das 0.975-Quantil aus einer anderen Erhebung bekannt war (Schultze et al., 2018), welches deutlich über dem wahren Wert lag.

eine neue Schätzung abzugeben mit dem Hinweis, dass auch die erste Schätzung erneut eingegeben werden könne (kein Hinweis darauf, dass sie diese auch korrigieren könnten). Der restliche Ablauf in der EG und der komplette Ablauf in der KG entsprachen jeweils dem Ablauf von EG und KG in Studie 5. Auch die abhängigen Maße entsprachen alle denen von Studie 5.

### 5.3.3. Ergebnisse & Diskussion

**Vorabanalysen.** Mit einer DJR von *M* = .69 (*SD* = .22) entschieden sich die Versuchspersonen auch in der neuen EG vergleichsweise häufig für ein explizites Richtungsurteil. In 98% dieser expliziten Richtungsurteile korrigierten sie zudem ihr erstes Urteil in die angegebene Richtung. Wurde kein explizites Richtungsurteil gewählt („weder noch"), so änderten die Versuchspersonen nur in 7% dieser Fälle ihre ersten Urteile. Insgesamt verhielten die Versuchspersonen sich damit bei ihren Korrekturen wieder konsistent mit ihren Richtungsurteilen.

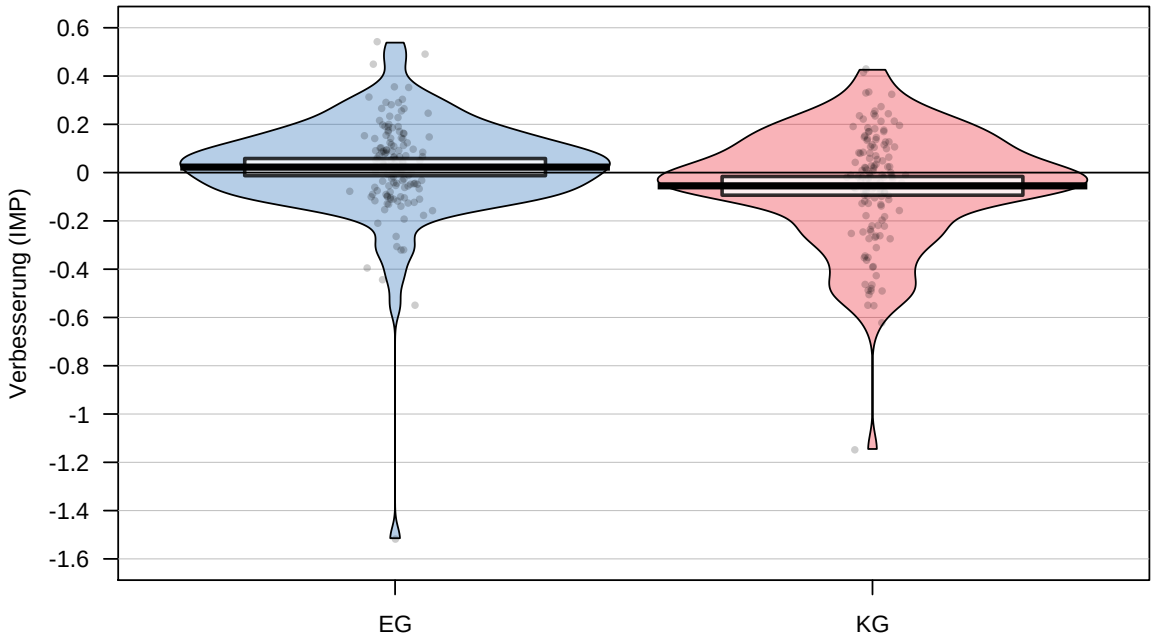
**Weisheitsbewusstsein.** Auch in Studie 6 waren die Versuchspersonen in der Lage, mit ihren expliziten Richtungsurteilen die Richtung der wahren Werte beziehungsweise der Populationsmittelwerte zu detektieren. Sowohl die TDR (*M* = .60, *SD* = .12) lag über .50, $t(137) = 9.74$, $p < .001$ (einseitig; P6), $d = 0.83$, als auch die CDR (*M* = .64, *SD* = .12), $t(137) = 13.50$, $p < .001$ (zweiseitig), $d = 1.15$. Somit konnten auch in dieser Studie H1 und H2a angenommen werden. Die zentrale Voraussetzung für die Wirksamkeit der (angepassten) Intervention war hiermit gegeben.

**Urteilsverbesserung.** Für den Test der neuen Intervention (H9a und H9b) wurden strukturell die gleichen Tests und abhängigen Variablen präregistriert wie in Studie 5 (siehe P6). Beim Test der standardisierten IMP Scores in der EG zeigte sich deskriptiv ein positiver Effekt (*M* = 0.02, *SD* = 0.21), der jedoch nicht signifikant von null verschieden war, $t(139) = 1.27$, $p = .103$ (einseitig; P6), $d = 0.11$. Somit zeigte sich keine hinreichende Evidenz für H9a. In der KG zeigte sich eine leichte Verschlechterung im IMP Score (*M* = -0.05, *SD* = 0.23), $t(139) = -2.80$, $p = .006$ (zweiseitig), $d = -0.24$. Formal lässt sich damit auch die H9b (eine stärkere *Verbesserung* in der EG) nicht testen, da sich keine Verbesserung in der EG zeigte, beziehungsweise kann H9b auch ohne weiteren Test abgelehnt werden. Mit dem präregistrierten Test von H9b lässt sich jedoch testen, ob die Verschlechterung in der KG

größer ausfiel als in der EG. Dies war der Fall, $t(276.03) = 2.92$, $p = .002$ (einseitig; P6), $d = 0.35$. Die Versuchspersonen verschlechterten sich in der KG stärker als in der EG (siehe Abbildung 13). So konnten mit der Intervention zumindest Verschlechterungen der Urteile reduziert, beziehungsweise verhindert werden. Um zu verstehen, warum sich insgesamt keine systematische Verbesserung in der EG einstellte, werden in dem nächsten Abschnitt die Korrekturen der Urteile genauer untersucht und darauffolgend potentielle Unterschiede nach Aufgabenkategorien exploriert.

**Abbildung 13**

*Urteilsverbesserung (IMP) der Bedingungen von Studie 6*




*Anmerkung*. Mittlere Verbesserung (IMP) der Urteile in Studie 6 (schwarze Linien) mit 95% Konfidenzintervallen (weiße Rechtecke), Verteilungsdichte („Geigenplots“) und Darstellung einzelner Versuchspersonen (schwarze Punkte). EG = Intervention/Experimentalgruppe. KG = Kontrollgruppe. IMP = Improvement Score auf Basis der absoluten Fehlerdifferenzen und standardisiert anhand der Standardabweichung der einzelnen Stimuli.

**Urteilsänderungen.** Auch die angepasste Intervention zielte darauf ab, dass Menschen weiterhin möglichst häufig ein Weisheitsbewusstsein ausbilden und ihre Urteile entsprechend korrigieren. Anders als bei den vorherigen Studien sollte aber zugleich

sichergestellt werden, dass diese Korrekturen nicht zu stark ausfallen. Um zu testen, ob dies der Fall war, wurden zunächst die ARs der Bedingungen verglichen, wobei sich wieder zeigte, dass die Urteile in der EG ($M$ = .71, $SD$ = .23) häufiger angepasst wurden als in der KG ($M$ = .62, $SD$ = .27), $t(269.67) = 3.00$, $p = .003$ (zweiseitig), $d = 0.36$. Im Anschluss wurden wieder alle Beobachtungen ausgeschlossen, bei denen das erste Urteil überhaupt nicht angepasst wurde. Bei den verbleibenden Beobachtungen zeigte sich, dass die z-standardisierten Opinion Shifts in der EG ($M$ = -0.10, $SD$ = 0.33) im Schnitt kleiner ausfielen als in der KG ($M$ = 0.14, $SD$ = 0.53), $t(232.56) = -4.50$, $p < .001$ (zweiseitig), $d = -0.54$. Die Intervention war somit zumindest insofern wirksam, dass Menschen häufiger ihre Urteile (im Sinne des Weisheitsbewusstseins) korrigierten und (im Vergleich zur KG) kleinere Korrekturen vornahmen. Auch hier stellt sich die Frage, ob die Korrekturen damit in der EG häufiger in der richtigen Richtung lagen als in der KG. Äquivalent zu Studie 5 wurden hierfür Kennwerte berechnet, die angeben, wie häufig eine Person ihre Urteile in die richtige Richtung korrigierte, sofern sie überhaupt eine Korrektur vornahm. In beiden Bedingungen erfolgten die Korrekturen überzufällig häufig in die richtige Richtung, das heißt sowohl in mehr als 50% der Fälle in der EG ($M$ = .60, $SD$ = .11), $t(137) = 10.21$, $p < .001$ (zweiseitig), $d = 0.87$, als auch in der KG ($M$ = .57, $SD$ = .14), $t(139) = 5.62$, $p < .001$ (zweiseitig), $d = 0.48$. Deskriptiv fiel dieser Wert in der EG etwas größer aus, wobei diese Differenz knapp nicht statistisch signifikant war, $t(263.55) = 1.96$, $p = .051$ (zweiseitig), $d = 0.23$. Insgesamt zeigt sich, dass die Intervention damit insofern wirkte, dass häufig kleinere Korrekturen in die richtige Richtung erfolgten, wobei sich, wie bereits erwähnt, insgesamt hierdurch keine positiven Effekte bezüglich der Verbesserung einstellten. Im Folgenden wurde explorativ untersucht, ob sich zumindest für eine der beiden Aufgabenkategorien positive Effekte der Intervention zeigten, was gegebenenfalls weitere Aufschlüsse darüber geben kann, unter welchen Voraussetzungen das Weisheitsbewusstsein zur Verbesserung von quantitativen Urteilen genutzt werden kann.

**Explorative Analyse der Aufgabenkategorien.** Für die separate Analyse der Aufgabenkategorien wurden äquivalente Tests zu den präregistrierten Analysen von H9a und H9b gerechnet, wobei für die explorativen Analysen nur zweiseitige Tests berichtet werden. In diesen explorativen Folgeuntersuchungen zeigte sich bei Aufgaben mit natürlichem Nullpunkt mit einem standardisierten IMP Score von $M$ = -0.01 ($SD$ = 0.29) keine

Verbesserung in der EG, $t(139) = -0.56$, $p = .577$ (zweiseitig), $d = -0.05$, auch wenn sich bei diesen Aufgaben eine geringere Verschlechterung zeigte als in der KG ($M = -0.09$, $SD = 0.35$), $t(267.93) = 1.98$, $p = .049$ (zweiseitig), $d = 0.24$. Diese Ergebnisse stehen im Einklang mit den generellen Ergebnissen über alle Aufgaben hinweg. Bei den historischen Schätzungen lagen jedoch die standardisierten IMP Scores in der EG statistisch signifikant über null ($M = 0.06$, $SD = 0.25$), $t(139) = 2.76$, $p = .007$ (zweiseitig), $d = 0.23$. Bei diesen Aufgaben zeigte sich zum ersten Mal in einer Bedingung eine Verbesserung der Urteile im Innersubjektvergleich. Die Versuchspersonen gaben in der EG zu Phase 2 akkuratere historische Schätzungen ab als zu Phase 1. Zudem fielen die Verbesserungen statistisch signifikant größer aus als in der KG ($M = -0.02$ $SD = 0.30$), $t(271.37) = 2.41$, $p = .017$ (zweiseitig), $d = 0.29$. In der KG zeigte sich keine Verbesserung der Urteile im Test gegen null, $t(139) = -0.82$, $p = .416$ (zweiseitig), $d = -0.07$.

Im Falle dieser ersten positiven Evidenz wurden allein die historischen Aufgaben weiter analysiert. Für diese Analysen kann das untersuchte Versuchsdesign auch als 2x2-Design betrachtet werden mit der Interventionsmanipulation als Zwischensubjektfaktor (EG vs. KG) und der Phase (1 vs. 2) als Messwiederholungsfaktor und der Größe der Urteilsfehlern zu Phase 1 und 2 als abhängige Variable. Mit H9a und H9b wird jeweils der kritische bedingte Haupteffekt und die kritische Interaktion dieses Designs getestet. So besagt H9a, dass die Fehler in der EG zur Phase 2 kleiner sind als zu Phase 1, und H9b, dass diese Reduktion der Fehler in der EG größer ist als in der KG. Diese Tests beruhen auf der Annahme, dass sich die Fehler zu Phase 1 (vor der Manipulation) durch die Randomisierung zwischen den Bedingungen nicht unterscheiden. Im Falle positiver Evidenz für H9a und H9b sollten somit auch die Fehler in der EG zu Phase 2 kleiner sein als die Fehler der KG zu Phase 1 und 2. Diese Annahmen wurde anhand der mittleren absoluten Fehler (MAE) der separaten Phasen getestet, die hierfür *z*-standardisiert wurden. Positive Werte indizieren damit eine schwächere Leistung. Zu Phase 1 zeigten sich anhand der standardisierten MAEs kein statistisch signifikanter Unterschied zwischen EG ($M = 0.04$, $SD = 0.39$) und KG ($M = -0.03$, $SD = 0.40$), $t(277.65) = 1.55$, $p = .122$ (zweiseitig), $d = 0.19$. Die Randomisierung war somit erfolgreich, wobei die Urteile in der KG deskriptiv etwas akkurater waren als in der EG. Die Analysen zu den IMP Scores haben bereits gezeigt, dass sich die Versuchspersonen in der EG verbesserten und die Versuchspersonen in der KG nicht verbesserten. Zu Phase 2 produzierten die Personen in der EG ($M = 0.01$, $SD = 0.37$) mit ihrer Verbesserung jedoch

keine kleineren Fehler als in der KG zu Phase 2 ($M$ = -0.02, $SD$ = 0.40), $t$(275.60) = 0.77, $p$ = .442 (zweiseitig), $d$ = 0.09, sondern waren vielmehr auf einem ähnlichen Niveau. Die Fehler der EG zu Phase 2 waren damit auch nicht kleiner als die Fehler der KG zu Phase 1, $t$(275.51) = 0.94, $p$ = .348 (zweiseitig), $d$ = 0.11. Es zeigte sich also nur bei einem der drei relevanten Vergleiche Evidenz für die Intervention, nämlich beim Vergleich EG Phase 2 vs. EG Phase 1 (Test IMP gegen null) und nicht bei EG Phase 2 vs. KG Phase 2 und EG Phase 2 vs. KG Phase 1. Der statistisch signifikante Test sollte jedoch eine wesentliche größere statistische Power haben, da es bei diesem Test um einen Innersubjektvergleich handelt und bei den anderen beiden um Zwischensubjektvergleiche. Auch der statistisch signifikante Zwischensubjektvergleich der IMP Scores sollte eine etwas höhere statistische Power haben als die Zwischensubjektvergleiche der MAEs, da die IMP Scores für die idiosynkratischen Unterschiede in den Leistungen der einzelnen Versuchspersonen kontrollieren und somit unsystematische Fehlervarianz reduzieren.

Zusammengefasst zeigte sich insgesamt (über beide Aufgabenkategorien hinweg) kein Nutzen der Intervention, wobei zumindest potentielle Verschlechterungen reduziert werden konnten. In einer explorativen Untersuchung zeigten sich jedoch zumindest bei einer Aufgabenkategorie, dass mit der überarbeiteten Intervention eine Verbesserung der Urteile erzielt werden konnte, auch wenn die Versuchspersonen der EG hiermit insgesamt nicht besser abschnitten als die Versuchspersonen in der KG. Somit liefert diese Studie erste Indizien dafür, unter welchen Voraussetzungen Menschen tatsächlich ihre Urteile selbst verbessern können. In Anbetracht der explorativen Natur dieser Analysen und der fehlenden Evidenz bezüglich aller relevanter Vergleiche (insbesondere der mit vermutlich geringer Power) sollten diese Ergebnisse jedoch nur als Ansatzpunkte für zukünftige Untersuchungen betrachtet werden. Bevor jedoch spekuliert wird, warum Menschen bei historischen Aufgaben ihre Urteile mit der Intervention verbessern konnten (und bei den anderen Aufgaben nicht), stellt sich die Frage, ob sich dieser explorative Befund überhaupt replizieren lässt. Im Zuge der Vorarbeiten zum letzten empirischen Abschnitt zum Umgang mit Ratschlägen (Abschnitt 6) wurde ein Datensatz erhoben, der zwar ein anderes Untersuchungsziel verfolgte (die Erhebung von quantitativen Ratschlägen), dessen Methode jedoch so erweitert wurde, dass dieser zumindest einen weiteren Test von H9a bei neuen historischen Aufgaben erlaubt.

## 5.4. Studie 7

In der letzten Studie dieses Abschnitts lässt sich anhand von einer neuen unabhängigen Stichprobe von Versuchspersonen und neuen Aufgaben testen, inwiefern die angepasste Intervention mit kleinen Korrekturen zu einer Verbesserung der Urteile bei historischen Schätzungen führt. In dieser Studie wurden von den Versuchspersonen insgesamt 50 neue historischen Schätzaufgaben zum Thema Musikgeschichte bearbeitet (für eine detaillierte Auflistung der Schätzaufgaben siehe Anhang A). Hierbei handelt es sich um eine Laborstudie, bei der es nur das sekundäre Ziel war, einen erneuten Test von H9b zu liefern, und die nicht präregistriert wurde. Primäres Ziel dieser Studie war, Urteile zu erheben, die in einer späteren Studie als Ratschläge verwendet werden können. Aufgrund der Corona-Pandemie musste diese Studie jedoch abgebrochen werden, und für die während der Pandemie zwingenden Onlinestudien wurde auf andere Aufgaben als die historischen Schätzaufgaben zurückgegriffen. Der Grund hierfür lag darin, dass die historischen Schätzaufgaben relativ einfach über Internetrecherchen gelöst werden können und sich daher nur für kontrollierte Laboruntersuchungen eignen. Neben der etwas kleineren Stichprobe ergeben sich aufgrund der ursprünglich anderen Zielsetzung gewisse Limitationen. Zum einen wurde in dieser Studie keine Kontrollgruppe ohne Intervention erhoben, zum anderen erhielten die Versuchspersonen nur für ihre ersten und nicht für ihre zweiten Urteile finanzielle Bonuszahlungen. Außerdem wurde diese Studie nicht präregistriert. Bisher gab es jedoch keinerlei Hinweise, dass Menschen ohne Intervention ihre Urteile verbessern können, und das Fehlen der Bonuszahlungen bei den zweiten Urteilen sollte einen positiven Interventionseffekt eher erschweren. Im Falle positiver Evidenz sollte diese Studie damit das Vertrauen in den explorativen Befund zu historischen Schätzungen steigern.

### 5.4.1. Design & Stichprobe

Bei Studie 7 handelt es sich um eine einfache Beobachtungsstudie zum Weisheitsbewusstsein und zur Intervention. Die Intervention mit kleinen Korrekturschritten wird in dieser Studie nur im Innersubjektvergleich getestet. Zur Bestimmung der Stichprobengröße wurde in dieser Studie keine Poweranalyse durchgeführt. Die Stichprobengröße basierte ausschließlich auf pragmatischen Überlegungen. Ziel war es,

mindestens 90 Versuchspersonen am gleichen Standort wie die anderen Laborstudien zu erheben, um für eine spätere Studie eine hinreichend große Menge an Ratschlägen zur Verfügung zu haben. Die Laborstudie musste jedoch aufgrund des Ausbruchs der Corona-Pandemie Mitte März 2020 abgebrochen werden, weshalb nur Daten von insgesamt 76 Versuchspersonen erhoben wurden und auch keine Folgestudie mit diesem Material zum Thema Ratschlagsnutzung folgen konnte. Zudem musste zwei Versuchspersonen aufgrund technischer Probleme ausgeschlossen werden. Der finale Datensatz bestand damit aus 74 Versuchspersonen (42 Frauen und 32 Männer) in einem Durchschnittsalter von *M* = 25.01 Jahren (*Mdn* = 24, *SD* = 4.99). Für ihre Teilnahme erhielten die Versuchspersonen einen Festbetrag von 4€ oder auf Wunsch eine halbe Versuchspersonenstunde und eine Bonuszahl für ihre Initialurteile aus Phase 1 von bis zu 5€ (*M* = 1.19, *SD* = 0.40).

### 5.4.2. Methode

Die Methodik von Studie 7 entsprach weitestgehend der Methodik der EG von Studie 6 mit zwei zentralen Unterschieden. Zum einen wurden keine Bonuszahlungen für Phase 2 des Experiments gezahlt, weswegen alle Verweise auf diese Zahlungen aus den Instruktionen gestrichen wurden. Zudem wurden die Aufgaben für diese Studie durch neue ersetzt. Hierbei handelte es sich um 50 historische Schätzaufgaben zum Thema Musik- und Musikinstrumentgeschichte (z.B. „In welchem Jahr wurde die Stimmgabel erfunden“; siehe Anhang A für eine detaillierte Auflistung). Die Berechnung der finanziellen Bonuszahlungen entsprach für Phase 1 der Berechnung von Studie 6, wobei nur historische Urteile präsentiert wurden und die Instruktionen zu den Bonuszahlungen entsprechend gekürzt wurden. Da es sich um die erste Studie mit diesem Material handelte, wurden die Versuchspersonen nicht wie in der vorherigen Studie über den durchschnittlichen Bonus informiert. Und schließlich wurden die quantitativen Urteile zu Phase 1 und 2 auf einen Bereich von 1500 und 2020 begrenzt, wobei die wahren Werte im Bereich von 1711 („In welchem Jahr wurde die Stimmgabel erfunden?“) bis 1995 („In welchem Jahr wurde das Museum der "Rock and Roll Hall of Fame" eröffnet?“) lagen.

### 5.4.3. Ergebnisse & Diskussion

**Vorabanalysen.** In dieser Studie wählten die Versuchspersonen mit einer DJR von *M* = .59 (*SD* = .27) in mehr als der Hälfte der Fälle ein Richtungsurteil, wobei eine Person in

keinem Fall ein explizites Richtungsurteil wählte und entsprechend von der Analyse der TDRs und CDRs ausgeschlossen wurde. In 99% der Fälle, in denen ein explizites Richtungsurteil gewählt wurde, wurde im nächsten Schritt auch das quantitative Urteil in die angegebene Richtung geändert. Wurde kein explizites Richtungsurteil gewählt („weder noch“), so wurde auch das quantitative Urteil in nur 7% der Fälle geändert. Somit verhielten sich die Versuchspersonen auch in dieser Studie weitestgehend konsistent mit ihren Richtungsurteilen.

**Weisheitsbewusstsein.** Entschieden sich die Versuchspersonen für ein Richtungsurteil, so lag dieses überzufällig häufig in der richtigen Richtung und auch in der Richtung des Populationsmittelwerts. So lag die TDR insgesamt mit $M$ = .60 ($SD$ = .14) über .50, $t(72) = 6.25$, $p < .001$ (zweiseitig), $d$ = 0.73, und auch die CDR mit $M$ = .62 (SD = .14) über .50, $t(72) = 7.31$, $p < .001$ (zweiseitig), $d$ = 0.86. Somit zeigte sich auch in Studie 7 deutliche Evidenz für H1 und H2a. Diese Ergebnisse sprechen noch einmal für die Robustheit des Weisheitsbewusstseins, da bei dieser Studie wieder komplett neue Stimuli verwendet wurden.

**Urteilsverbesserung.** Zum Test der Urteilsverbesserung (H9a) wurde wie in Studie 5 und 6 der standardisierte Improvement Score berechnet und mit einem Einstichproben-$t$-Test gegen null getestet. Dieser fiel mit $M$ = 0.07 ($SD$ = 0.18) signifikant größer als null aus, $t(73) = 3.21$, $p$ = .002 (zweiseitig), $d$ = 0.37. Damit zeigte sich auch in dieser explorativen Studie Evidenz für H9a bei historischen Aufgaben. Insgesamt bestätigt diese Studie damit den Befund aus Studie 6, dass sich zumindest bei historischen Aufgaben eine Verbesserung im Rahmen der angepassten Intervention einstellt.

### 5.5. Diskussion zu Interventionen im individuellen Urteilskontext

In dem zweiten empirischen Abschnitt wurde getestet, ob Menschen mit Hilfe von einfachen Interventionen ihr Weisheitsbewusstsein zur Verbesserung ihrer Urteile nutzen können. Hierfür wurde zunächst eine Intervention entwickelt, bei der Menschen im Sinne des Weisheitsbewusstseins ihre Urteile so weit korrigieren, bis sie das Gefühl haben, dass der wahre Wert weder eher in der einen noch eher in der anderen Richtung liegt. Auch wenn Menschen grundsätzlich in der Lage waren, ein Weisheitsbewusstsein auszubilden (zumindest in Studie 5), zeigte sich, dass diese Intervention wirkungslos blieb. Das Problem

war, dass Menschen ihre Urteile zu stark anpassten, auch wenn die Richtung dieser Korrekturen weitestgehend konsistent mit dem Weisheitsbewusstsein war. Auch wenn sie ihre Urteile mehrheitlich in die richtige Richtung korrigierten, schossen sie dabei teilweise über das Ziel hinaus und korrigierten ihre Urteile auch teilweise in die falsche Richtung. Insgesamt vergrößerten sie mit und ohne Intervention eher ihre Urteilsfehler anstatt sie zu verkleinern.

Auf Basis dieser Befunde wurde die Intervention dahingehend weiterentwickelt, dass Versuchspersonen bei einem expliziten Richtungsurteil aufgefordert wurden, ihre Urteile lediglich einen kleinen Schritt in die angegebene Richtung zu korrigieren. In einer präregistrierten Studie (Studie 6) erwies sich auch diese überarbeitete Intervention im teststarken Innersubjektvergleich ($1 – \beta = .97$ bei $d = 0.30$) als nicht erfolgreich. Sie kann somit nicht allgemein zur Korrektur von Urteilen empfohlen werden. Im Gegensatz zu der ursprünglichen Intervention zeigte sich bei der überarbeiteten Intervention jedoch zumindest deskriptiv ein positiver Effekt. Explorativ zeigte sich zudem, dass Menschen zumindest bei den historischen Schätzungen in der Lage waren, ihre Urteile zu verbessern und diese Verbesserung auch stärker ausfiel als in der Kontrollgruppe ohne Intervention. Dieser Effekt ließ sich in einer weiteren Studie mit neuen historischen Aufgaben im Innersubjektvergleich bestätigen. Ein kritischer und eindeutiger Zwischensubjektvergleich steht nach dem Stand dieser Arbeit jedoch noch aus, da sich der Mehrwert der Intervention gegenüber der Kontrollgruppe bisher nur unter Kontrolle der anfänglichen Akkuratheitsunterschiede zwischen den Versuchspersonen zeigte. Die bisherigen Ergebnisse lassen diesbezüglich jedoch optimistisch stimmen, da die Versuchspersonen in den Kontrollgruppen ihre Urteile in keiner der Studien verbessern konnten und der positive Effekt der angepassten Intervention bei historischen Schätzaufgaben im Innersubjektvergleich zwei Mal gezeigt werden konnte. Es ist trotzdem nicht auszuschließen, dass es sich bei diesen explorativen Ergebnissen um Zufallsbefunde handelt. Eine Arbeit von Müller-Trede (2011) zur Weisheit der Vielen in einer Person deutet allerdings ebenfalls darauf hin, dass es tatsächlich qualitative Unterschiede zwischen den Aufgabenkategorien geben könnte. So konnte Müller-Trede die Weisheit der Vielen in einer Person für historische Aufgaben replizieren (und auch für Prozentschätzungen), jedoch nicht für eine gemischte Kategorie von numerischen Aufgaben mit einem natürlichen Nullpunkt. Dies

entspricht auch dem Befundmuster von Studie 6, auch wenn Menschen in dieser Arbeit selbst ihre ersten Urteile korrigierten und keine Mittelwerte aus zwei Urteilen gebildet wurden.

In Anbetracht der Ergebnisse stellt sich die Frage, warum Menschen nur bei den historischen Aufgaben mit einer solchen Instruktion Akkuratheitsgewinne erzielen konnten und nicht auch bei der anderen Aufgabenkategorie. Nach Müller-Trede (2011) sprechen diese Ergebnisse gegen die Sampling-Erklärungen für die Weisheit der Vielen in einer Person (z.B. Vul & Pashler, 2008), nach denen sich in seinen Daten ein aufgabenübergreifender Effekt hätte zeigen müssen. Müller-Trede (2011) liefert jedoch keine alternative Erklärung, welche die positiven Effekte bei den anderen Aufgabenkategorien erklären könnte. Die Daten in dieser Arbeit sprechen gegen die Schlussfolgerung, dass das Sampling von neuen Informationen das Problem war. So besaßen Menschen auch bei den bei null begrenzten Aufgaben in allen Studien ein Weisheitsbewusstsein mit vergleichsweise hohen TDRs von ca. .60 (siehe Tabelle 2) und wussten daher, in welche Richtung sie ihre Urteile korrigieren sollten. Sie hatten jedoch offensichtlich Schwierigkeiten angemessene Korrekturen vorzunehmen, selbst wenn sie die Aufforderung erhielten, ihr Urteile nur „einen kleinen Schritt“ anzupassen. Ein Grund könnte sein, dass Menschen Schwierigkeiten hatten der Instruktion zu „kleinen Korrekturen“ bei bestimmten Urteilsaufgaben Folge zu leisten. Solch eine Korrektur erfordert eine gewisse Vertrautheit mit den Aufgaben und Wissen darüber, wie die wahren Werte üblicherweise verteilt sind. Hierbei handelt es sich um Wissen, dass von Brown und Siegler (1993) als *Metric Knowledge* bezeichnet wird, da es Menschen erlaubt, Urteile auf einen plausiblen Bereich zu beschränken. Stern et al. (2017) nennen als Beispiel für Metric Knowledge bei der Schätzung von Distanzen das Wissen, dass Deutschland eine Länge von ca. 900 km hat und der Äquator eine Länge von ca. 40.000 km, was Menschen erlaubt, anhand dieser Referenzen extreme Fehler (und auch extreme Korrekturen) bei anderen Distanzen zu vermeiden. Metric Knowledge sollte auch notwendig sein, damit die Aufforderung, Urteile „einen kleinen Schritt“ in eine bestimmte Richtung anzupassen, auch mit einer tatsächlich kleinen Korrektur auf der entsprechenden Skala beantwortet werden kann und nicht zu extrem ausfallen.

Auf Basis der Daten lässt sich jedoch nur spekulieren, ob bei den historischen Aufgaben ein besseres metrisches Wissen vorlag. Gerade bei der anderen

Aufgabenkategorie handelte es sich um Urteilsaufgaben mit sehr variierenden Einheiten (Cent, Kilometer, Stunden), für die teilweise sehr unterschiedliches metrisches Wissen erforderlich ist. Es lässt sich jedoch vermuten, dass Menschen gerade bei historischen Schätzungen ein besseres Gefühl für ein angemessenes Maß von kleinen Korrekturen haben, beziehungsweise welches Maß an Präzision bei welcher Frage gefordert ist. Bei jüngeren Ereignissen (z.B. „Wann gewann George Clooney seinen ersten Oscar“) ist meist ein höheres Maß an Präzision gefordert (die Ereignisse sollten nur um wenige Jahre über oder unterschätzt werden), weshalb bei der Aufforderung zu „kleine Korrekturen“ vermutlich nur Korrekturen um wenige Jahre vorgenommen werden. Bei älteren Ereignissen ist hingegen meist ein geringeres Maß von Präzision gefordert, weshalb kleine Korrekturen durchaus auch mehrere Jahrzehnte umfassen könnten. Die Daten aus Studie 6 und 7 stützen diese Überlegungen. So zeigte sich über die insgesamt 65 historischen Schätzaufgaben aus Studie 6 und 7 ein starker negativer Zusammenhang zwischen Höhe des wahren Werts (je höher desto jünger das Ereignis) und den mittleren Korrekturen (nicht standardisierte Opinion Shifts), $r = -.75$, $t(63) = -8.87$, $p < .001$ (zweiseitig). So korrigierten Menschen beispielsweise bei dem jüngsten Ereignis in Studie 6 und 7 („In welchem Jahr wurde China Mitglied der World Trade Organization (WTO)?“; wahrer Wert 2001; siehe Anhang A) ihre Urteile im Schnitt um acht Jahre, sofern sie eine Korrektur vornahmen. Bei dem ältesten Ereignis („In welchem Jahr wurde die Stimmgabel erfunden?“; wahrer Wert 1711) korrigierten Menschen hingegen ihre Urteile im Schnitt um 33 Jahre. Die Menschen könnten damit bei historischen Schätzaufgaben ein gutes Gefühl dafür haben, welches Maß an kleinen Korrekturen je nach historischem Ereignis angebracht ist.

Bei der anderen Aufgabenkategorie mit natürlichem Nullpunkt lässt sich jedoch nur schwer das Argument vorbringen, dass im Gegensatz zu den historischen Schätzungen *kein* hinreichendes metrisches Wissen vorlag, da es sich um sehr unterschiedliche Aufgaben handelte und die Größe der nicht standardisierten Opinion Shifts kaum vergleichbar ist (z.B. „Wie lang ist der Suez-Kanal?“ und „Wie viele Filialen hatte IKEA 2016?“; siehe Anhang A für eine Auflistung der Fragen mit wahren Werten). Bei näherer Betrachtung der Aufgaben stechen jedoch einige Aufgaben hervor, bei denen es Menschen vermutlich schwer fiel das richtige Maß an Präzision zu finden und besonders starke Korrekturen vorgenommen wurden. Bei der Frage „Wie viele Filialen hatte IKEA 2016?“ wurden beispielsweise häufig

große anstatt kleine Korrekturen vorgenommen. Der Mittelwert der absoluten Opinion Shifts lag hier bei Korrekturen um 9856 Filialen (*Mdn* = 200), obwohl der wahre Wert selbst nur bei 389 Filialen lag. Ein anderes Beispiel ist die Länge des Suezkanales (wahrer Wert 163 km), bei der im Mittel absolute Korrekturen von 194 km vorgenommen wurden (*Mdn* = 36 km). Charakteristisch ist für diese Aufgaben zudem, dass, anders als bei den historischen Schätzungen, der Median der Korrekturen meist deutlich unter dem Mittelwert der Korrekturen liegt (trotz Begrenzung extremer Urteile in Studie 6). Das heißt, dass es vermutlich einer Subgruppe von Menschen bei den jeweiligen Aufgaben besonders schwer fiel ein angemessenes Maß von Korrekturen zu finden; vermutlich, weil ihnen das notwendige Maß an metrischem Wissen fehlte.

Insgesamt zeigen diese Befunde, dass sich das Weisheitsbewusstsein zwar nicht generell durch die vorgeschlagenen Interventionen nutzen lässt, jedoch unter bestimmten Voraussetzungen zu einer Verbesserung von Urteilen führen kann. Problematisch sind hierbei insbesondere extreme Korrekturen, die vermutlich nur dann nicht systematisch eintreten, wenn eine explizite Aufforderung zu kleineren Korrekturen bei bestimmten Aufgaben erfolgt. Somit ist die Nutzbarkeit des Weisheitsbewusstseins im individuellen Urteilskontext nach bisherigem Stand noch relativ eingeschränkt.

In dieser Arbeit sollen Interventionen auf Basis des Weisheitsbewusstseins jedoch nicht nur in individuellen, sondern auch in einem sozialen Urteilskontext getestet werden (beim Umgang mit Ratschlägen). Es stellt sich die Frage, ob mit der bisherigen Interventionslogik in sozialen Kontexten bessere Ergebnisse erzielt werden können als im individuellen Kontext. Die Voraussetzungen stehen hierfür durchaus nicht ungünstig, da beim Umgang mit Ratschlägen tendenziell zu selten Korrekturen in Richtung des Ratschlags vorgenommen werden, beziehungsweise fallen diese in diesem Kontext tendenziell zu klein aus (Soll & Larrick, 2009; Soll & Mannes, 2011). Im folgenden Abschnitt werde ich herleiten, wie eine angepasste Intervention auf Basis des Weisheitsbewusstseins zu einer Verbesserung der Nutzung von Ratschlägen führen kann, und werde diese im Rahmen einer letzten Studie empirisch testen.

## 6. Intervention zur Verbesserung der Ratschlagsnutzung

In diesem Abschnitt werde ich untersuchen, inwiefern sich die bisher untersuchte Interventionslogik auf Basis des Weisheitsbewusstseins positiv auf den Umgang mit Ratschlägen auswirkt. Unabhängige Ratschläge sind aufgrund der Weisheit der Vielen eine sehr effektive Methode, um die Qualität von Urteilen zu verbessern (Soll & Larrick, 2009; Yaniv, 2004a), und können daher selbst als wirksame Intervention zur Verbesserung von Urteilen betrachtet werden. Doch auch beim Umgang mit Ratschlägen wird dieses Potential nur unzureichend von Menschen genutzt (Soll & Larrick, 2009; Soll et al., 2021). Im Folgenden werde ich detaillierter ausarbeiten, wie es bei dem Umgang mit Ratschlägen zu Defiziten kommen kann und inwiefern eine Intervention auf Basis des Weisheitsbewusstseins zu einer Verbesserung der Urteilsprozesse in diesem sozialen Kontext führen kann.

### 6.1. Defizite bei der Ratschlagsnutzung

Der Umgang mit Ratschlägen wird bei quantitativen Urteilen meist anhand des sogenannten *Judge-Advisor-Systems* untersucht (JAS; Sniezek & Buckley, 1995). Im JAS nehmen Versuchspersonen die Rolle der *Judges* ein, die ihre Urteile zunächst alleine treffen und im zweiten Schritt die Möglichkeit bekommen, diese Urteile zu revidieren. Für den zweiten Schritt bekommen die Judges zusätzlich zu ihren ursprünglichen Urteilen die Urteile einer oder mehrerer anderer Personen (*Advisors*) als Ratschläge präsentiert und erhalten die Möglichkeit, diese im Zuge ihrer Revision zu berücksichtigen. Über diesen Austausch der numerischen Werte hinaus erfolgt meist keine Kommunikation zwischen Judges und Advisors. Das JAS fokussiert damit primär auf *informationale Einflüsse*, das heißt, inwiefern Menschen Ratschläge von anderen Personen als Informationsquelle für ihre Urteilsbildung nutzen können. Deshalb werden *normative Einflüsse*, die auf dem Motiv basieren, in der sozialen Umwelt möglichst einen positiven Eindruck zu hinterlassen und Erwartungen von anderen gerecht zu werden (Deutsch & Gerard, 1955), weitestgehend kontrolliert (siehe auch Rader et al., 2017). Zentrale abhängige Variablen zur Ermittlung solcher informationaler Einflüsse sind im JAS zum einen die prozentuale Gewichtung der Ratschläge und die Verbesserung vom initialen zum finalen Urteil (Bonaccio & Dalal, 2006).

Die Forschung zum JAS zeigt, dass Menschen generell in der Lage sind, ihre Urteile mit Hilfe von Ratschlägen zu verbessern und das Potential einer unabhängigen Einschätzung einer anderen Person zu nutzen (Bonaccio & Dalal, 2006; Soll & Larrick, 2009). Auf der anderen Seite zeigen sich jedoch auch deutliche Defizite beim Umgang mit Ratschlägen. Dies zeigt sich insbesondere darin, dass Menschen im Schnitt Ratschläge weniger stark nutzen als sie sollten, was auch als *Egocentric (Advice) Discounting* (EAD) bezeichnet wird (Bonaccio & Dalal, 2006; Rader et al., 2017; Yaniv & Kleinberger, 2000). EAD lässt sich in vielen JAS Studien eindeutig quantifizieren, nämlich insbesondere dann, wenn Judge und Advisor im Schnitt die gleiche Akkuratheit haben und beide Urteile somit im Schnitt gleich stark (zu 50%) gewichtet werden sollten (Soll & Larrick, 2009). Empirisch zeigt sich jedoch, dass Menschen in diesen Situationen Ratschläge im Schnitt nur zu ca. 30% gewichten (Rader et al., 2017; Soll & Larrick, 2009; Yaniv & Kleinberger, 2000). Diese 30% kommen dadurch zustande, dass Menschen einerseits häufig Ratschläge mitteln (50% Gewichtung), jedoch andererseits mindestens ebenso häufig Ratschläge komplett ignorieren (0% Gewichtung). Somit können sie zwar häufig ihre Urteile verbessern, erzielen jedoch eine schwächere Leistung als es durch die konsequente Bildung eines Mittelwerts möglich wäre (Soll & Larrick, 2009). Es zeigt sich also auch in diesem Kontext, dass Menschen die Weisheit der Vielen (in diesem Fall die Weisheit eines Ratschlags) nur unzureichend nutzen.

Bisher wurden verschiedene Erklärungsansätze postuliert, um EAD zu erklären (für Überblicksarbeiten siehe Bonaccio & Dalal, 2006; Rader et al., 2017). Bisher gibt es jedoch keine Evidenz, die einen dieser Erklärungsansätze vollumfänglich stützen kann (Rader et al., 2017; Soll & Mannes, 2011). Entsprechend schwierig ist es, zielgerichtete psychologische Interventionen zu entwickeln, um die Gewichtung von Ratschlägen zu verbessern. Die bisherigen Interventionen zielen primär darauf ab, unspezifisch für eine Gleichgewichtung von Ratschlägen und eigenem Urteil zu sorgen (z.B., Yaniv & Choshen-Hillel, 2012a, 2012b; für eine Ausnahme siehe Sniezek et al., 2004). Zum einen sind diese Strategien in der Praxis oft nur schwer umsetzbar oder zumindest schwer vermittelbar, da sie zum Beispiel voraussetzen, dass Entscheidungsträger:innen ihr eigenes Wissen bei der Gewichtung nicht mit einbringen und sich einem sogenannten *Blindfolding* unterziehen (Yaniv & Choshen-Hillel, 2012a). Vermutlich gestaltet es sich als schwierig, Menschen auf anderen Wegen von einer Gleichgewichtung zu überzeugen. Zum anderen ist es so, dass eine Gleichgewichtung

häufig kontraindiziert ist und sich paradoxe Effekte ergeben können, zum Beispiel, wenn sowohl offensichtlich nutzlose als auch sehr hochwertige Ratschläge zu 50% gewichtet werden. Es handelt sich auch bei EAD keineswegs um ein kontextunabhängiges Phänomen. Vermutlich hängt die Stärke von EAD wesentlich von der Qualität von Ratschlägen ab (Schultze et al., 2017). So haben Menschen Schwierigkeiten Ratschläge zu ignorieren, die offensichtlich nutzlos oder irreführend sind (Fiedler et al., 2019; Schultze et al., 2017). In diesen Fällen sollten Interventionen im besten Fall dafür sorgen, dass die nutzlosen Ratschläge komplett ignoriert werden. Auf der anderen Seite sollten Ratschläge von Expert:innen häufig zu deutlich mehr als 50% gewichtet werden.

Mit einer Intervention auf Basis des Weisheitsbewusstseins sollen für die Ratschlagsnutzung spezifischere Effekte erzielt werden. Das heißt, dass mit der Intervention die Nutzung hochwertiger Ratschläge gesteigert werden soll und die Nutzung von Ratschlägen mit niedriger Qualität gesenkt werden soll. Im Folgenden werde ich zunächst kurz skizzieren, wie Menschen mit einem Weisheitsbewusstsein zu spezifischen Gewichtungsstrategien in der Lage sein könnten. Im Anschluss werde ich hierauf basierend eine Intervention für den Umgang mit Ratschlägen ableiten.

### 6.2. Das Weisheitsbewusstsein bei der Gewichtung von Ratschlägen

Wie bereits erwähnt, können Ratschläge für die Judges unterschiedlich nützlich sein. Für die Unterscheidung von Ratschlägen nach ihrer Nützlichkeit liefert Yaniv (2004b) eine Definition, die für die jeweiligen Judges unmittelbare Implikationen für die Gewichtung entsprechender Ratschläge aufzeigt. Yaniv unterscheidet zwischen *hilfreichen* und *nicht hilfreichen* Ratschlägen. Hilfreiche Ratschläge zeigen ausgehend von den Urteilen der Judges in die Richtung des wahren Werts (die richtige Richtung), während nicht hilfreiche Ratschläge in die entgegengesetzte Richtung zeigen (die falsche Richtung). Qualitativ hochwertige Ratschläge (mit einem kleinen Fehler) sollten wesentlich häufiger hilfreich sein als Ratschläge mit einem größeren Fehler. Trifft ein Ratschlag direkt den wahren Wert (Ratschlag: „Celle hat 69.399 Einwohner:innen“), so ist dieser für alle Menschen hilfreich, die mit ihrem ersten Urteil von diesem Wert abweichen. Aber auch Ratschläge mit einem größeren Fehler können für Judges hilfreich sein, wenn sie in die richtige Richtung zeigen. Wenn beispielsweise Celle leicht unterschätzt wird (65.000) sollte ein Ratschlag mit einem

größeren Fehler, der aber in der richtigen Richtung liegt (z.B. 110.000), zumindest zu einem geringen Anteil gewichtet werden. Die relative Nützlichkeit für die jeweiligen Judges (ist der Ratschlag hilfreich oder nicht) und die Qualität von Ratschlägen (wie groß ist der Fehler des Ratschlags) sind somit sehr eng verwandt, auch wenn sie sich, wie bereits erwähnt, konzeptuell etwas unterscheiden.

Für die Judges im JAS ist es ohne zusätzliche Informationen zunächst nur schwer erkennbar, ob Ratschläge eher hilfreich sind oder eher nicht hilfreich sind und stärker oder schwächer gewichtet werden sollten, zumindest wenn keine zusätzlichen Expertiseinformationen zur Verfügung stehen. Tatsächlich sollten die Ratschläge mehrheitlich hilfreich sein, sofern ihnen teilweise neue Informationen zugrunde liegen, genau wie das Weisheitsbewusstsein eher in die richtige Richtung zeigt, wenn es auf mehr neuen Informationen beruht (siehe Abschnitt 2.3.). So sind ca. zwischen 60% und 70% der Ratschläge hilfreich, wenn sie zufällig aus der gleichen Verteilung gezogen werden wie die Urteile der Judges (Soll & Larrick, 2009; Soll et al., 2021).[18] Nach der Naivitätsannahme (Fiedler & Juslin, 2006b) sollten Menschen jedoch nicht in der Lage sein zu abstrahieren, dass vielen Ratschlägen neue Informationen zugrunde liegen und deshalb im Zweifelsfall zumindest ein kleines Stück gewichtet werden sollten (siehe auch, Hütter & Ache, 2016; Minson et al., 2011).

Allerdings können Menschen mit dem Weisheitsbewusstsein anhand ihrer eigenen „neuen“ Informationen erkennen, ob ein Ratschlag eher hilfreich ist oder nicht. Liegt der Ratschlag im Einklang mit ihrem (häufig richtigen) Richtungsurteil, so sollte auch der Ratschlag eher hilfreich als nicht hilfreich sein und entsprechend stärker gewichtet werden. Voraussetzung für eine erfolgreiche Umsetzung ist, dass Menschen Ratschläge konsistent mit ihren Richtungsurteilen gewichten. Für eine genauere Erläuterung unterscheide ich zunächst zwischen *erwartungskonformen* Ratschlägen und *nicht erwartungskonformen* Ratschlägen. Ein Ratschlag ist nach meiner Definition erwartungskonform, wenn er mit dem

[18] In diesen Arbeiten wird nicht der Anteil hilfreicher Ratschläge berichtet, sondern die Bracketing Rates von Judge und Advisor (siehe Abschnitt 2.3.1.). Die Häufigkeit hilfreicher Ratschläge lässt sich direkt aus der Bracketing Rate berechnen, sofern die Urteile der Judges und die Urteile der Advisor zufällig aus der gleichen Verteilung gezogen werden (bspw., wenn Judges und Advisor ihre Rolle zufällig zugeteilt bekommen). Beispielsweise sind bei einer Bracketing Rate von 50% (in 50% der Fälle „klammern zwei Urteile den wahren Wert ein“) 75% der Ratschläge hilfreich, da die Ratschläge in allen Fällen mit Bracketing hilfreich sind und zusätzlich in der Hälfte der Fälle, bei denen kein Bracketing vorliegt.

expliziten Richtungsurteil übereinstimmt, also über dem ersten quantitativen Urteil liegt bei dem Richtungsurteil „drüber“ oder unter dem ersten Urteil liegt bei dem Richtungsurteil „drunter“. Ratschläge sind entsprechend nicht erwartungskonform, wenn sie in der entgegengesetzten Richtung liegen oder wenn kein explizites Richtungsurteil gewählt wurde und der Ratschlag vom eigenen Urteil abweicht. Aufgrund des Weisheitsbewusstseins sind erwartungskonforme Ratschläge eher hilfreich als nicht erwartungskonforme Ratschläge. Menschen sollten also hilfreiche Ratschläge stärker gewichten können als nicht hilfreiche, wenn sie sich konsistent mit ihren Erwartungen (bzw. Richtungsurteilen) verhalten. Mit anderen Worten kann das Weisheitsbewusstsein als „Diagnosetool“ eingesetzt werden, um den Nutzen von Ratschlägen zu bewerten. Entscheidend ist aber natürlich, dass auch bei dem Umgang mit Ratschlägen ein Weisheitsbewusstsein ausgebildet wird und bei der Gewichtung von Ratschlägen eingesetzt wird.

Die Erkennung und Gewichtung hilfreicher Ratschläge mit dem Weisheitsbewusstsein lässt sich noch einmal an dem Einführungsbeispiel verdeutlichen. In dem Beispiel hat eine Person aus Bayern die Bevölkerungszahl von Celle (69.399) auf 30.000 geschätzt (anhand von Coburg, Forchheim und Lichtenfels) und eine andere Person aus Schleswig-Holstein auf 70.000 (anhand von Norderstedt, Elmshorn und Neumünster). Die Person aus Bayern würde in diesem Beispiel das Urteil der Person aus Schleswig-Holstein als Ratschlag erhalten. Nach dem GSM könnte die Person aus Bayern genau wie beim Weisheitsbewusstsein aus ihrem privaten Informationspool teilweise *neue* Informationen heranziehen, die noch nicht in ihr erstes Urteil eingeflossen sind (siehe Abschnitt 2.3.2.) und damit eine Bewertung vornehmen, die von ihrem eigenen Urteil (teilweise) unabhängig ist. Würde sie beispielsweise „Bamberg“ (ca. 75.000) als neue Information mit Coburg (ca. 40.000) austauschen, könnte sie mit dieser Information erkennen, dass der *erwartungskonforme* Ratschlag (70.000) ausgehend vom eigenen Urteil (30.000) vermutlich in die richtige Richtung zeigt. Anhand dieser Information würde sie auch erkennen, dass *nicht erwartungskonforme* Ratschläge in der anderen Richtung (z.B. 20.000) eher nicht hilfreich sind.

In der Literatur zum Umgang mit Ratschlägen gibt es zumindest eine Arbeit, die darauf hindeutet, dass Menschen auch spontan ihr Weisheitsbewusstsein bei der Gewichtung von Ratschlägen einsetzen. So zeigte sich bei Yaniv (2004b), dass Menschen

hilfreiche Ratschläge stärker gewichten als nicht hilfreiche Ratschläge, auch wenn für die absolute Abweichung des Ratschlags vom ersten Urteil kontrolliert wird und es keine anderen Rückschlüsse auf die Qualität der Ratschläge gibt. Auf der anderen Seite gibt es auch deutliche Hinweise dafür, dass Menschen vergleichsweise selten oder nur unzureichend neue Informationen verwenden, um Ratschläge zu gewichten. Evidenz hierfür liefern die Befunde zu EAD (Bonaccio & Dalal, 2006; Rader et al., 2017; Soll & Larrick, 2009), die zeigen, dass Menschen sehr häufig Ratschläge ignorieren und ihr erstes Urteil beibehalten, obwohl die meisten Ratschläge hilfreich sind. Würden Menschen bei der Gewichtung von Ratschlägen vornehmlich neue Informationen verwenden, so sollten die Ratschläge bei ähnlicher Expertise der Judges im Schnitt näher an 50% gewichtet werden. Der Grund hierfür liegt darin, dass die neuen Informationen ähnlich häufig die Ratschläge und die ersten Urteile der Judges stützen sollten (siehe auch Yaniv, 2004b). Ich werde daher in einer letzten empirischen Studie die bisherige Interventionslogik auf den Ratschlagskontext anpassen. Die Intervention soll Menschen dazu anregen, ihr Weisheitsbewusstsein (häufiger) bei der Gewichtung von Ratschlägen einzusetzen und damit insbesondere die hilfreichen Ratschläge stärker zu gewichten.

### 6.3. Studie 8

Für den Test der Intervention beim Umgang mit Ratschlägen wurde eine weitere Onlinestudie mit Kalorienschätzungen durchgeführt (wie in Studie 3), die den vorherigen Interventionsstudien in Abschnitt 5 ähnelte und bei der zusätzlich zur Intervention (EG vs. KG) manipuliert wurde, ob die Versuchspersonen vor der Abgabe ihrer zweiten Urteile einen Ratschlag von einer früheren Versuchsperson erhielten (R+) oder nicht (R-). Damit entsprach der Ablauf dieses 2x2-Designs in den beiden Bedingungen mit Ratschlag dem JAS. Die Intervention bestand auch in dieser Studie aus einer Abfrage des Weisheitsbewusstsein. Diese Abfrage erfolgte immer auf einer eigenen Seite, bevor Ratschläge präsentiert wurden und ein zweites Urteil abgegeben wurde, damit sich das Weisheitsbewusstsein unabhängig vom Ratschlag ausbilden kann und nicht systematisch Informationen abgerufen werden, die eher für oder gegen den Ratschlag als das erste eigene Urteil sprechen (siehe z.B., Rader et al., 2015; Schultze et al., 2017). Das primäre Ziel dieser Studie lag darin zu untersuchen, inwiefern die Intervention Menschen hilft, Ratschläge besser zu gewichten. Entsprechend erfolgte in dieser Studie keine Bewegungsaufforderung in eine bestimmte Richtung. Im

letzten Schritt wurden lediglich Initialurteil und, je nach Bedingung, Richtungseinschätzung (EG) und Ratschlag (R+) präsentiert, und die Versuchspersonen wurden aufgefordert, ein neues Urteil zu treffen.

Um die Wirksamkeit der Intervention im Kontext der Ratschlagsnutzung zu untersuchen, wurden verschiedene Hypothesen präregistriert (siehe P8). Die erste präregistrierte Hypothese (H1) postulierte wieder ein Weisheitsbewusstsein als Grundvoraussetzung für den Erfolg der Intervention, wobei ähnlich wie in den Studien 5 und 6 über die Auswahl der Stimuli die Chancen deutlich erhöht wurden, dass sich ein Weisheitsbewusstsein zeigt. Hierfür wurden die Daten zu den Kalorienschätzungen aus Studie 3 herangezogen und aus den 42 Gerichten 15 ausgewählt, die die höchsten TDRs aufwiesen.

Die erste Reihe neuer Hypothesen für den JAS Kontext bezog sich allein auf die Bedingung, in der die Versuchspersonen die Intervention und Ratschläge erhielten (EG/R+). Anhand dieser Bedingung sollte getestet werden, ob Menschen mit dem Weisheitsbewusstsein in der Lage sind, die hilfreichen Ratschläge stärker zu gewichten als die nicht hilfreichen. Ich nehme an, dass Menschen sich grundsätzlich mit ihren Richtungsurteilen konsistent verhalten und *erwartungskonforme* Ratschläge stärker gewichten als *nicht erwartungskonforme*. Aufgrund des Weisheitsbewusstseins (H1) sollten Menschen damit auch in der Lage sein, *hilfreiche* Ratschläge stärker zu gewichten als *nicht-hilfreiche* Ratschlage. Für diese Zusammenhänge, beziehungsweise Effekte, wurden drei Hypothesen aufgestellt:

> **Hypothese 10 (H10; H2 in P8):** Menschen gewichten Ratschläge stärker, wenn sie mit der angegebenen Richtung des wahren Werts übereinstimmen (erwartungskonform) als wenn dies nicht der Fall ist (nicht erwartungskonform).
>
> **Hypothese 11 (H11; H3 in P8):** Menschen gewichten hilfreiche Ratschläge stärker als nicht hilfreiche Ratschläge (in der richtigen vs. falschen Richtung).
>
> **Hypothese 12 (H12; H4 in P8):** Der Effekt unter H11 wird durch den Effekt unter H10 mediiert.

Nach dem GSM ist es möglich, dass Menschen auch in der Ratschlagsbedingung ohne Intervention (KG/R+) hilfreiche Ratschläge stärker gewichten als nicht hilfreiche Ratschläge (Yaniv 2004b). Ich nehme jedoch an, dass diese Prozesse durch die Intervention verstärkt werden, sofern sie nicht hierdurch erst ausgelöst werden:

> **Hypothese 13 (H13; Abweichung von P8[19])**: Die Differenz in der Gewichtung hilfreicher und nicht hilfreicher Ratschläge (siehe H11) fällt mit Intervention stärker aus als ohne Intervention (EG/R+ vs. KG/R+).

Durch die positiven Effekte der Intervention auf die Gewichtung von Ratschlägen sollten Menschen auch in der Akkuratheit ihrer Urteile stärker von Ratschlägen profitieren. Im Einklang mit der bisherigen Forschung erwarte ich, dass Menschen unabhängig von der Intervention in der Lage sind, ihre Urteile mit Hilfe von Ratschlägen zu verbessern (Bonaccio & Dalal, 2006). Mit Intervention sollten die Versuchspersonen jedoch stärker von den Ratschlägen profitieren als ohne Intervention. Aufgrund des Ausbleibens bisheriger Interventionseffekte im individuellen Urteilskontext wurde keine Hypothese für einen Effekt der Intervention in dieser Bedingung aufgestellt, sondern nur eine Interaktionshypothese bezüglich der spezifischen Wirksamkeit der Intervention beim Umgang mit Ratschlägen:

> **Hypothese 14 (H14; H5 in P8):** Menschen können ihre Urteile stärker verbessern, wenn sie Ratschläge erhalten, als wenn sie keine Ratschläge erhalten.

> **Hypothese 15 (H15; H6 in P8):** Der positive Effekt auf die Verbesserung der Urteile durch den Erhalt von Ratschlägen sollte beim Vergleich der Bedingungen mit Intervention (EG/R+ vs. EG/R-) größer ausfallen als ohne Intervention (KG/R+ vs. KG/R-).

Zuletzt wurde eine Serie von Hypothesen für den Fall aufgestellt, dass sich Evidenz für den Interaktionseffekt unter H15 zeigt. Da sich die Intervention spezifisch auf die

[19] Anstelle dieser Hypothese wurden ursprünglich zwei Hypothesen aufgestellt, nämlich „Menschen gewichten *hilfreiche* Ratschläge (in der richtigen Richtung) mit Intervention *stärker* als ohne Intervention (EG/R+ vs. KG/R+)“ und „Menschen gewichten *nicht hilfreiche* Ratschläge (in der falschen Richtung) mit Intervention *schwächer* als ohne Intervention (EG/R+ vs. KG/R+)“ (H8 und H9 in P8). Bei diesen Hypothesen kann jedoch der Fall auftreten, dass sich Evidenz für die eine aber nicht die andere Hypothese zeigt und somit unklar ist, ob generell in der einen Bedingung Ratschläge stärker gewichtet werden oder spezifisch die hilfreichen. Entsprechend wurden die Hypothesen durch die kombinierte Hypothese ersetzt (H13 im Text).

Gewichtung von Ratschlägen auswirken sollte, gehe ich davon aus, dass sich ein Vorteil der Intervention beim Vergleich der Verbesserungen der beiden Ratschlagsbedingungen zeigt (EG/R+ vs. KG/R+). Sofern dies der Fall ist, gehe ich davon aus, dass dieser bedingte Haupteffekte über die Unterschiede in der Gewichtung hilfreicher und nicht hilfreicher Ratschläge (H13) mediiert wird:

> **Hypothese 16 (H16; H7 in P8):** Menschen erzielen mit Ratschlag und Intervention größere Verbesserungen ihrer Urteile als Menschen mit Ratschlag aber ohne Intervention (EG/R+ vs. KG/R+).

> **Hypothese 17 (H17; Abweichung von P8[20]):** Der Effekt unter H16 wird über die Unterschiede in der Gewichtung hilfreicher und nicht hilfreicher Ratschläge unter H13 mediiert.

### 6.3.1. Design & Stichprobe

Bei Studie 8 handelte es sich um ein 2x2-Zwischensubjektdesign mit orthogonaler Manipulation der Intervention und der Präsentation von Ratschlägen: Intervention/Experimentalgruppe (EG) vs. Kontrollgruppe (KG) x mit Ratschlag (R+) vs. ohne Ratschlag (R-). Die Stichprobengröße wurde vorab auf mindestens *N* = 400 festgelegt (100 Personen pro Zelle; siehe P8). Hierfür wurde keine Testplanung durchgeführt, sondern auf Basis von Vorerfahrungen angenommen, dass es sich bei 100 Personen pro Zelle um eine angemessen große Stichprobengröße handelt, um neue Hypothesen teststark mit geplanten Kontrasten zu testen, die für die zentralen Hypothesen (H14 und H15) geplant waren. Hierfür wurden 500 Personen zur Teilnahme bei MTurk angefragt, unter der Annahme, dass nach den vorherigen Onlineerhebungen der Datenverlust durch Ausschlusskriterien bis zu 20% betragen könnte. Rekrutierung und Ausschlusskriterien entsprachen denen von Studie 3 (Geoblocking, Ablehnungsrate bei MTurk unter 5%, Aufmerksamkeits- und Verständnischecks, keine doppelten Teilnahmen), aus welcher auch die Ratschläge für Studie 8 gewonnen wurden. Abgesehen von dem Kriterium zu doppelten Teilnahmen wurden alle

[20] Ursprüngliche wurde eine doppelte Mediation über die höhere Gewichtung hilfreicher Ratschläge und die niedrigere Gewichtung nicht hilfreicher Ratschläge in EG/R+ vs. KG/R+ präregistriert (H10 und H11 in P8). Die Hypothesen bezüglich der angenommenen Mediatoren wurden jedoch durch eine kombinierte Hypothese ersetzt (siehe Fußnote 19), weshalb auch die Mediationshypothesen entsprechend angepasst wurden. Die Interpretation der Ergebnisse änderte sich hierdurch nicht.

Ausschlusskriterien präregistriert beziehungsweise direkt bei der Einladung berücksichtigt (siehe P8).[21]

Die Studie wurde *N* = 497 Mal beendet. In weiteren neun Fällen wurde die Studie bis zur Zuteilung zu den Versuchsbedingungen bearbeitet, aber vor Beendigung abgebrochen (siehe P8). Von den 497 Datensätzen wurden 102 aufgrund wiederholter Falschantworten bei den Verständnisfragen ausgeschlossen, und weitere 20 von Personen, welche die Studie mehrfach bearbeitet hatten. Der finale Datensatz bestand damit aus *N* = 375 Personen (224 Männer, 139 Frauen, drei nichtbinär/divers, neun ohne Angabe) im Alter von *M* = 35.39 Jahren (*Mdn* = 33, *SD* = 9.83) mit folgender Zellbesetzung: 95 EG/R+, 94 EG/R-, 93 KG/R+, und 93 KG/R-. Die Versuchspersonen erhielten für die Teilnahme an der Studie einen Grundbetrag von 2$ und für Phase 1 und 2 insgesamt Boni von bis zu 2$ (*M* = 0.88, *SD* = 0.55).

### 6.3.2. Methode

Auch Studie 8 bestand aus zwei Phasen. Die Methodik für Phase 1 entsprach in allen Bedingungen der Methodik von Studie 3, abgesehen von der Berechnung und Beschreibung der Boni. Für ein besseres Verständnis wird Phase 1 im Folgenden noch einmal kurz zusammengefasst (für weitere Details siehe Studie 3). Die Studie startete mit einer reCAPTCHA-Aufgabe, einer Einverständniserklärung und einfachen Verständnis- und Aufmerksamkeitschecks, nach denen sich die Ausschlusskriterien richteten. Im weiteren Verlauf wurden die Kalorienschätzungen und mögliche Bonuszahlungen erläutert. Im Anschluss folgte Phase 1, bei der für mehrere Gerichte Kalorien geschätzt wurden. Die Besonderheit in Studie 8 bestand lediglich darin, dass immer die gleichen 15 Gerichte in der gleichen Reihenfolge präsentiert wurden (anstatt 14 von 42 in Studie 3). Zudem wurden die Versuchspersonen wie in den Interventionsstudien im vorherigen Abschnitt bereits zu Phase 1 darauf hingewiesen, dass sie auch in Phase 2 einen Bonus verdienen könnten, der nicht von den Antworten in Phase 1 abhing und erst zu Phase 2 detaillierter erläutert werden

[21]Zum Zeitpunkt der Erhebung wurde auf Basis der Analysen der bisherigen Onlinestudien davon ausgegangen, dass doppelte Teilnahmen durch MTurk komplett verhindert werden. Diese Annahme beruhte jedoch auf einem ursprünglich fehlerhaften Analysecode, der erst nach der Erhebung korrigiert wurde. Weitere Teilnahmen wurden von MTurk nur verhindert, sofern die Versuchspersonen die Studie vorher komplett bearbeitet hatten.

würde. Die Berechnung des Bonus wurde im Vergleich zu Studie 3 leicht abgeändert, da die Versuchspersonen in Studie 3 höhere Bonuszahlungen erhielten als auf Basis der Daten von Soll et al. (2021) mit dem gleichen Aufgabentyp erwartet wurde, wodurch Studie 3 teurer als geplant war. In dieser neuen Studie wurde für Phase 1 ein Gericht mit Schätzung zufällig ausgewählt, für das die Versuchspersonen 1$ bekamen. Von diesem Bonus wurden fünf Cent für alle 20 Kalorien abgezogen, mit welchen der wahre Wert verfehlt wurde. Nach Bearbeitung der ersten Phase wurden die Versuchspersonen randomisiert den vier unterschiedlichen Versuchsbedingungen zugeordnet, wobei alle Zellen gleichmäßig besetzt wurden. Eine perfekt gleichmäßige Zellbesetzung konnte jedoch nicht erreicht werden, da in Einzelfällen auch Abbrüche nach der Zuteilung zu den Bedingungen erfolgten. Versuchspersonen, die aufgrund falscher Antworten auf die Verständnisfragen ausgeschlossen wurden, konnten die Studie dennoch beenden und auch einen Bonus für Phase 2 verdienen. Sie bearbeiteten die Studie dabei in der KG ohne Ratschläge, gingen aber nicht in die Zellbesetzung dieser Bedingung ein.

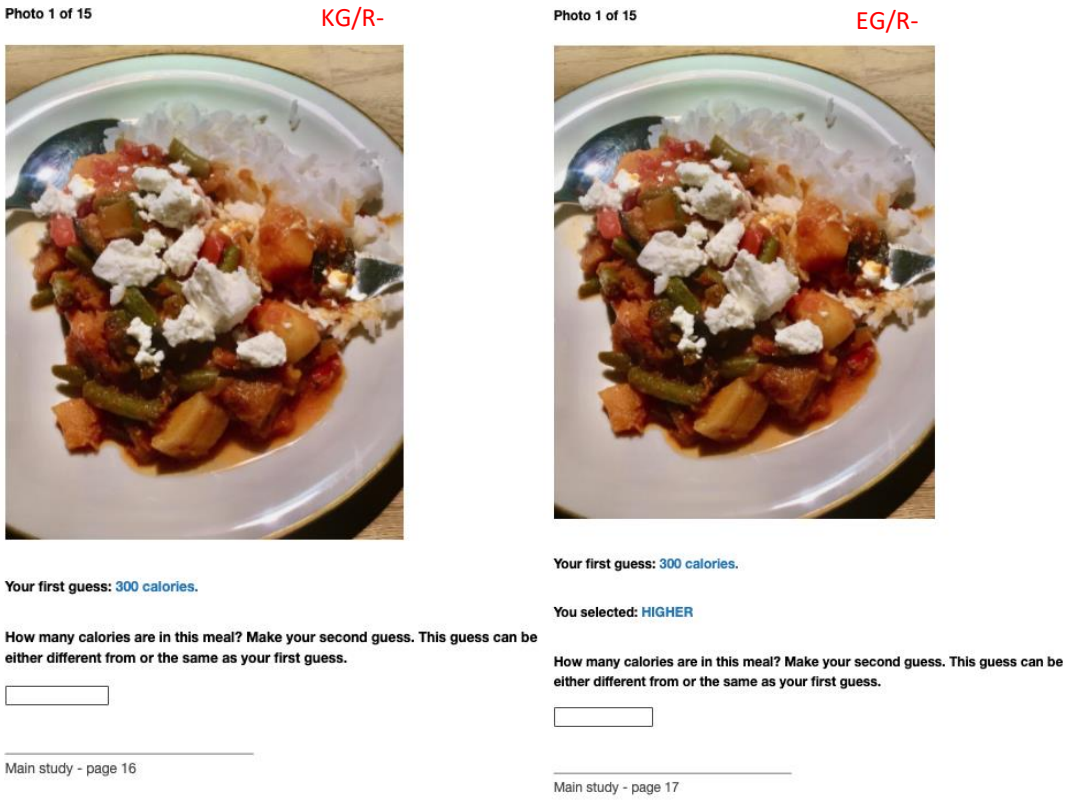
Die zweite Phase wurde auch in Studie 8 mit der Eingabe eines Codes initiiert. Der Ablauf zu Phase 2 entsprach in den beiden Bedingungen ohne Ratschlag weitestgehend KG und EG der vorherigen Interventionsstudien. In der EG/R- folgten nach der kurzen allgemeinen Einleitung zum Ablauf die detaillierten Instruktionen zur Abgabe der Richtungsurteile in der englischen Variante, die schrittweise präsentiert wurden, wobei keine explizite Aufforderung zur Korrektur erfolgte. Die Versuchspersonen wurden lediglich darüber informiert, dass sie nach ihren Richtungsurteilen zusätzlich nach einer zweiten Schätzung gefragt werden würden und dass sie hierfür wahlweise die gleiche Schätzung wie vorher abgeben könnten oder eine andere. Im Anschluss folgten die Instruktionen zum Bonus für Phase 2, der genau wie in Phase 1 berechnet wurde. Zuletzt folgte wie in Phase 1 eine kurze Zusammenfassung zur Bearbeitung der Urteilsaufgaben und eine Erinnerung, dass erneut ein Bonus für ein zufällig ausgewähltes Gericht gezahlt werden würde.

Auf den nächsten Seiten wurden die gleichen Aufgaben wie in Phase 1 in der gleichen Reihenfolge bearbeitet. In der EG/R- wurde auch in dieser Studie erst ein Richtungsurteil abgegeben und auf einer neuen Seite ein neues quantitatives Urteil. Hierfür wurden das Gericht, das erste Urteil und das Richtungsurteil präsentiert, und die Versuchspersonen wurden aufgefordert, ein neues Urteil (oder auch das gleiche erneut) abzugeben (siehe

Abbildung 14 rechte Bildhälfte). In der KG/R- erfolgten äquivalente Instruktionen, die um die Instruktionen zu den Richtungsurteilen reduziert waren. Zudem wurde nur die Seite präsentiert, bei der ein neues quantitatives Urteil abgegeben werden konnte (oder das alte erneut; siehe Abbildung 14 linke Bildhälfte).

**Abbildung 14**

*Abgabe der zweiten Urteile in den Versuchsbedingungen ohne Ratschlag*



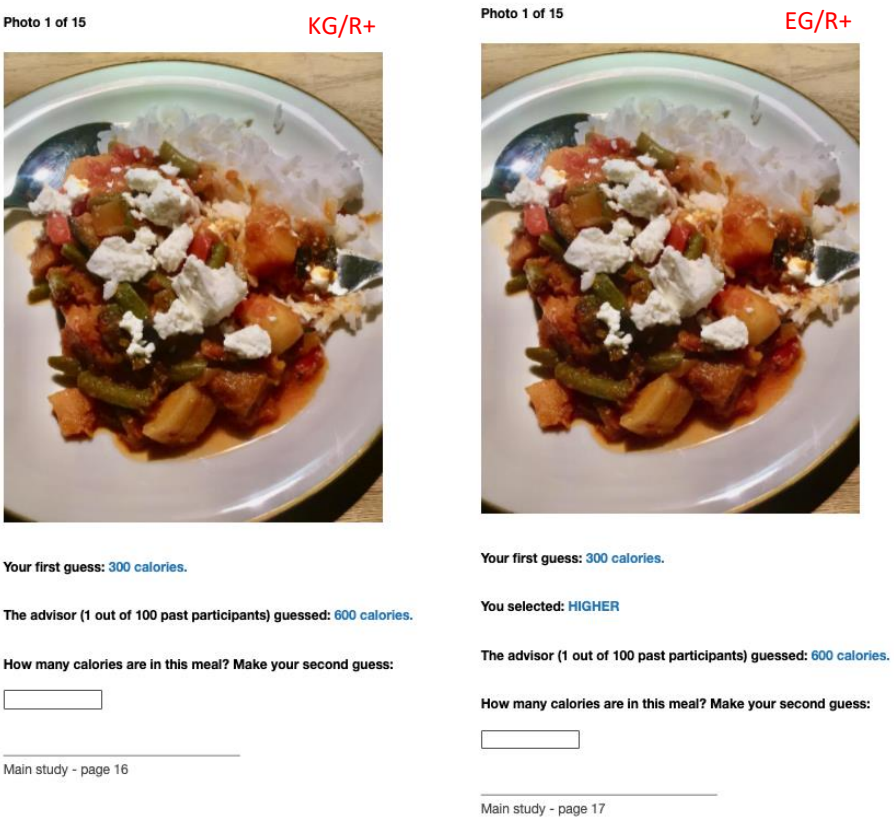

*Anmerkung*. KG/R- = Kontrollgruppe ohne Ratschlag. EG/R- = Experimentalgruppe ohne Ratschlag.

Die Bedingungen mit Ratschlag unterschieden sich bei der Bearbeitung der Aufgabe von den jeweiligen anderen Bedingungen nur darin, dass auf der jeweiligen Seite zur Abgabe eines zweiten quantitativen Urteils auch ein quantitatives Urteil einer anderen Versuchsperson als Ratschlag angezeigt wurde (siehe Abbildung 15). Zur Generierung dieser Ratschläge wurden zufällig 100 Ratschläge pro Aufgabe aus dem finalen Datensatz von Studie 3 gezogen, wobei noch nicht die Versuchspersonen ausgeschlossen wurden, die die Studie vorher einmal begonnen und abgebrochen hatten (siehe Fußnote 21). Aus diesen

Ratschlägen wurden zufällig 100 Sets mit je einem Ratschlag pro Aufgabe gebildet, wobei jeder Ratschlag nur einmal verwendet wurde (Ziehen ohne Zurücklegen). Bei der Zuteilung zu den Bedingungen wurde den Versuchspersonen in den Bedingungen mit Ratschlag zufällig eines der 100 Sets zugewiesen. Innerhalb der Bedingungen wurden hierbei Sets bevorzugt, die in der jeweiligen Bedingung noch nicht verwendet wurden. In der Einleitung zu Phase 2 wurden die Versuchspersonen bereits darauf hingewiesen, dass im Folgenden die Schätzung einer früheren Versuchsperson angezeigt werden würde (der/die Ratgeber:in für dieses Gericht) und es die Aufgabe der Versuchsperson sei, eine zweite (möglicherweise korrigierte) Schätzung abzugeben. Die Zusammenfassung der Instruktionen wurde in beiden Bedingungen um einen weiteren Punkt ergänzt, bei dem betont wurde, dass für jedes Gericht zufällig eine neue Person aus dem Pool von 100 früheren Versuchspersonen ausgewählt werden würde.

**Abbildung 15**

*Abgabe der zweiten Urteile in den Versuchsbedingungen mit Ratschlag*




*Anmerkung*. KG/R+ = Kontrollgruppe mit Ratschlag. EG/R+ = Experimentalgruppe mit Ratschlag.

Nach der Bearbeitung der letzten Aufgabe folgten in allen Bedingungen wie in Studie 3 der Abschlussfragebogen, der Urhebervermerk zu dem Bildmaterial, Informationen über den erreichten Bonus, ein kurzes Debriefing und die automatisierte Auszahlung der Versuchspersonen mit nachträglicher Überweisung der Boni.

### 6.3.3. Abhängige Maße

Als abhängige Maße wurden in diesem Abschnitt größtenteils die gleichen Maße wie in den vorherigen Studien verwendet. Genau wie in Studie 4 wurde der IMP Score als Verbesserungsmaß präregistriert, der nicht standardisiert wurde (siehe P8)[22]. Zudem wurde als neue Variable der *Advice Taking Koeffizient* (AT; Harvey & Fischer 1997) präregistriert (siehe P8), der die Gewichtung von Ratschlägen als relatives Maß operationalisiert und wie folgt definiert ist:

$$AT = \frac{Urteil_{Phase\,2} - Urteil_{Phase\,1}}{Ratschlag - Urteil_{Phase\,1}} \quad (3)$$

Der AT entspricht hierbei der prozentualen Gewichtung des Ratschlags im Rahmen der Korrektur des ersten Urteils und ist nur für Fälle definiert, bei denen erstes Urteil und Ratschlag nicht übereinstimmen. Entsprechend wurden bei allen Analysen mit dem AT in beiden Bedingungen je 2% der Beobachtungen ausgeschlossen, bei denen erstes Urteil und Ratschlag übereinstimmten. Bei den definierten Fällen entspricht ein AT von 1 einer kompletten Übernahme des Ratschlags, ein AT von 0.5 einer Gleichgewichtung und ein AT von 0 einer Beibehaltung des ersten Urteils. Im Einklang mit der bisherigen Forschung (z.B. Soll & Larrick, 2009) wurden ATs kleiner 0 durch 0 und ATs größer 1 durch 1 ersetzt (siehe P8).

### 6.3.4. Ergebnisse

**Weisheitsbewusstsein.** Als Vorabanalyse wurde für die beiden Experimentalgruppen untersucht, wie häufig sich die Versuchspersonen für ein explizites Richtungsurteil

[22] Auch bei den Kalorienschätzungen können sich grundsätzlich Ausmaß der Fehler und IMP Scores zwischen den Gerichten stark unterscheiden, was für eine Standardisierung spräche, jedoch bei der Präregistrierung nicht bedacht wurde. Für diese Auswahl von Gerichten hatte eine Standardisierung der IMP Scores jedoch kaum einen Einfluss, beziehungsweise zeigte sich mit Standardisierung qualitativ das gleiche Ergebnismuster. Im Fließtext wird deshalb nur die nicht standardisierten Variante der IMP Scores berichtet, die auch ursprünglich präregistriert wurde.

entschieden haben. Es zeigte sich, dass Menschen sowohl in der EG/R+ mit einer DJR von *M* = .68 (*SD* = .28), als auch in der EG/R- mit einer DJR von *M* = .55 (*SD* = .27) mehrheitlich explizite Richtungsurteile wählten. Fünf Personen in der EG/R+ und vier in der EG/R- entscheiden sich in keinem der 15 Fälle für ein explizites Richtungsurteil. Zudem zeigte sich, dass die Versuchspersonen in der EG mit Ratschlägen häufiger ein explizites Richtungsurteil wählten als in der EG ohne Ratschläge, $t(186.70) = 3.08$, $p = .002$ (zweiseitig), $d = 0.45$. Ein Grund könnte hierfür sein, dass Menschen durch vorherige Ratschläge indirekt neue Informationen erhielten, die sie später zur Einschätzung der Richtung des wahren Werts nutzen konnten. Zum Beispiel könnten Menschen durch Ratschläge lernen, dass sie bei ihren ersten Urteilen meist zu hoch geschätzt haben (Stern, 2017) und hierdurch später häufiger den wahren Wert in einer bestimmten Richtung vermuten (bspw. unter dem ersten Urteil). In diesem Fall sollte sich auch ein ausgeprägteres Weisheitsbewusstsein in der EG mit Ratschlägen zeigen, und die Richtungsurteile sollten auch häufiger in der Richtung der Populationsmittelwerte liegen, sofern diese neuen Informationen wiederum aus dem geteilten Pool aller Informationen stammen (was nach dem GSM der Fall sein sollte).

Für den Test des Weisheitsbewusstseins wurde in dieser Studie ein teststarker Einstichproben-*t*-Test präregistriert (P8), bei welchem die Daten aus beiden Interventionsbedingungen zusammen gegen .50 getestet wurden. Hierbei zeigte sich, dass die TDRs über beide Interventionsbedingungen hinweg im Schnitt über .50 lagen ($M = .61$, $SD = .23$), $t(179) = 6.03$, $p < .001$ (einseitig; P8), $d = 0.45$. In einer präregistrierten Folgeuntersuchung zeigte sich jedoch, dass die TDRs in der EG/R+ ($M = .65$, $SD = .23$) höher lagen als die TDRs in der EG/R- ($M = .56$, $SD = .24$), $t(177.63) = 2.38$, $p = .019$ (zweiseitig; P8), $d = 0.35$. Für diesen Fall wurden zusätzlich separate einseitige Tests des Weisheitsbewusstseins präregistriert (siehe P8). Auch bei den separaten Analysen lagen die TDRs sowohl in der EG/R+ statistisch signifikant über .50, $t(89) = 6.14$, $p < .001$ (einseitig; P8), $d = 0.65$, als auch in der EG/R-, $t(89) = 2.58$, $p = .011$ (einseitig; P8), $d = .27$. Somit zeigte sich auch bei beiden einzelnen Bedingungen Evidenz für H1, beziehungsweise ein Weisheitsbewusstsein in beiden Bedingungen. Ergänzend zu den TDRs wurden in dieser Studie auch die CDRs analysiert, wobei direkt separate Analysen für die zwei Interventionsbedingungen gerechnet wurden. Die CDRs lagen sowohl in der EG/R+ ($M = .67$, $SD = .22$) über .50, $t(89) = 7.28$, $p < .001$ (zweiseitig), $d = 0.77$, als auch in der EG/R- ($M = .57$,

*SD* = .24), *t*(89) = 2.85, *p* = .005 (zweiseitig), *d* = 0.30, wobei sich auch die CDRs zwischen den Bedingungen unterschieden, *t*(175.29) = 2.68, *p* = .008 (zweiseitig), *d* = 0.40. Insgesamt zeigte sich auch in Studie 8 Evidenz für H2a. Auch in dieser Studie wurde demzufolge mit den expliziten Richtungsurteilen im Schnitt der Populationsmittelwert detektiert. Die höheren TDRs und CDRs in der EG/R+ stehen zudem in Einklang mit der oben erläuterten Annahme, dass Menschen durch frühere Ratschläge neue Informationen gewinnen. Solch ein Prozess kann sich positiv auf die spezifische Wirksamkeit der Intervention im Ratschlagskontext auswirken. Konkret könnte durch *frühere* Ratschläge mit Intervention das Weisheitsbewusstsein gesteigert werden, wodurch bei *späteren* Ratschlägen besser zwischen hilfreichen und nicht hilfreichen Ratschlägen unterschieden werden kann. Unabhängig von der Erklärung dieses explorativen Befunds, zeigten sich mit einem ausprägten Weisheitsbewusstsein beim Umgang mit Ratschlägen gute Voraussetzungen für den Erfolg der Intervention im JAS.

**Gewichtung von Ratschlägen.** Für die Gewichtung von Ratschlägen wurden mehrere Hypothesen aufgestellt. Die ersten drei bezogen sich auf die Gewichtung von Ratschlägen in der EG/R+ (H10-H12). Als erstes wurde die Hypothese aufgestellt, dass Ratschläge besonders stark gewichtet werden, wenn diese erwartungskonform sind, das heißt, wenn sie mit der Richtung der Richtungsurteile übereinstimmen (H10). Diese Hypothese lässt sich mit Hilfe von *t*-Tests für abhängige Stichproben über gemittelte ATs der EG/R+ testen. Die hierfür untersuchten Fälle traten jedoch nicht bei allen Versuchspersonen auf, weshalb nicht immer alle Versuchspersonen bei den entsprechenden Tests berücksichtigt werden konnten. Zum Beispiel konnten die Ratschläge bei den fünf Versuchspersonen in der EG/R+, die sich nie für eine explizite Richtung entschieden, auch in keinem Fall erwartungskonform sein. Bei dem ersten *t*-Test zeigte sich, dass die mittleren ATs bei Ratschlägen in der Richtung des Richtungsurteils (*M* = 0.59, *SD* = 0.23) höher ausfielen als bei Ratschlägen in entgegengesetzter Richtung oder bei Ratschlägen ohne explizites Richtungsurteil (*M* = 0.34, *SD* = 0.30), *t*(89) = 9.43, *p* < .001 (zweiseitig), *d* = 0.99. Es zeigte sich also eindeutige Evidenz für H10.[23] Die erwartungskonformen Ratschläge wurden fast doppelt so stark gewichtet wie

[23] Für den Test von H10 wurde für eine besser verständliche Darstellung von der Präregistrierung abgewichen. Ursprünglich wurde für den Test von H10 ein Test im Rahmen einer Multi-Level Mediationsanalyse präregistriert (siehe P8). Qualitativ zeigten sich bei dieser Analyse die gleichen Ergebnisse (eine höhere Gewichtung erwartungskonformer Ratschläge).

Ratschläge, die nicht mit den Richtungsurteilen übereinstimmten. Zwischen den Fällen mit Ratschlägen in entgegengesetzter Richtung ($M$ = 0.32, $SD$ = 0.35) und ohne explizites Richtungsurteil ($M$ = 0.37, $SD$ = 0.31) zeigte sich nur deskriptiv ein Unterschied, $t(76) = -1.81$, $p = .075$ (zweiseitig), $d = -0.21$ und in beiden Fällen wurde auch bei separaten Vergleichen schwächer gewichtet als bei erwartungskonformen Ratschlägen, beide $p$s < .001 (zweiseitig).

Aufgrund des Weisheitsbewusstseins und erwartungskonformer Gewichtungen sollten die Versuchspersonen in der EG/R+ in der Lage sein, hilfreiche Ratschläge, die in die Richtung des wahren Werts zeigen, stärker zu gewichten als nicht hilfreiche Ratschläge, die in die falsche Richtung zeigen (H11). Um diese Hypothese zu testen, wurden für jede Versuchsperson in der EG/R+ mittlere ATs für die hilfreichen und nicht hilfreichen Ratschläge berechnet (in Richtung des wahren Werts und entgegengesetzter Richtung) und in einem $t$-Test für abhängige Stichproben verglichen. Hierbei zeigten sich keine höheren ATs für hilfreiche Ratschläge ($M$ = 0.40, $SD$ = 0.25) als für nicht hilfreiche Ratschläge ($M$ = 0.38, $SD$ = 0.31), $t(79) = 0.88$, $p = .380$ (zweiseitig), $d = 0.10$.[24] Die Versuchspersonen waren trotz Weisheitsbewusstsein nicht in der Lage, hilfreiche Ratschläge stärker zu gewichten als nicht hilfreiche Ratschläge. Somit konnte H11 nicht angenommen werden, weshalb auch die Testung der Mediationshypothese (H12) obsolet war. Formal entfällt hierdurch auch der Test von H13, nach denen der in H11 postulierte Effekt in der EG/R+ größer ausfällt als in der KG/R+. Im Sinne der Vollständigkeit sei jedoch erwähnt, dass auch in der KG/R+ hilfreiche Ratschläge ($M$ = 0.46, $SD$ = 0.28) nicht stärker gewichtet wurden als nicht hilfreiche ($M$ = 0.45, $SD$ = 0.29) und die Gewichtung hilfreicher und nicht hilfreicher Ratschläge nicht statistisch signifikant von der EG/R+ verschieden war (H13), beide $p$s > .640 (zweiseitig). Auch insgesamt zeigte sich bei der Gewichtung von Ratschlägen in EG/R+ ($M$ = 0.43, $SD$ = 0.26) und KG/R+ ($M$ = 0.48, $SD$ = 0.28) kein statistisch signifikanter Unterschied, $t(184.68) = -1.31$, $p = .192$ (zweiseitig), $d = -0.19$, wobei die Ratschläge in der KG tendenziell etwas stärker gewichtet wurden. Zur Vollständigkeit wird im nächsten Abschnitt der kritische Test der Intervention berichtet, bei dem die Verbesserung der Urteile im Vergleich aller Versuchsbedingungen analysiert wurde. Die wesentlichen Voraussetzungen für eine

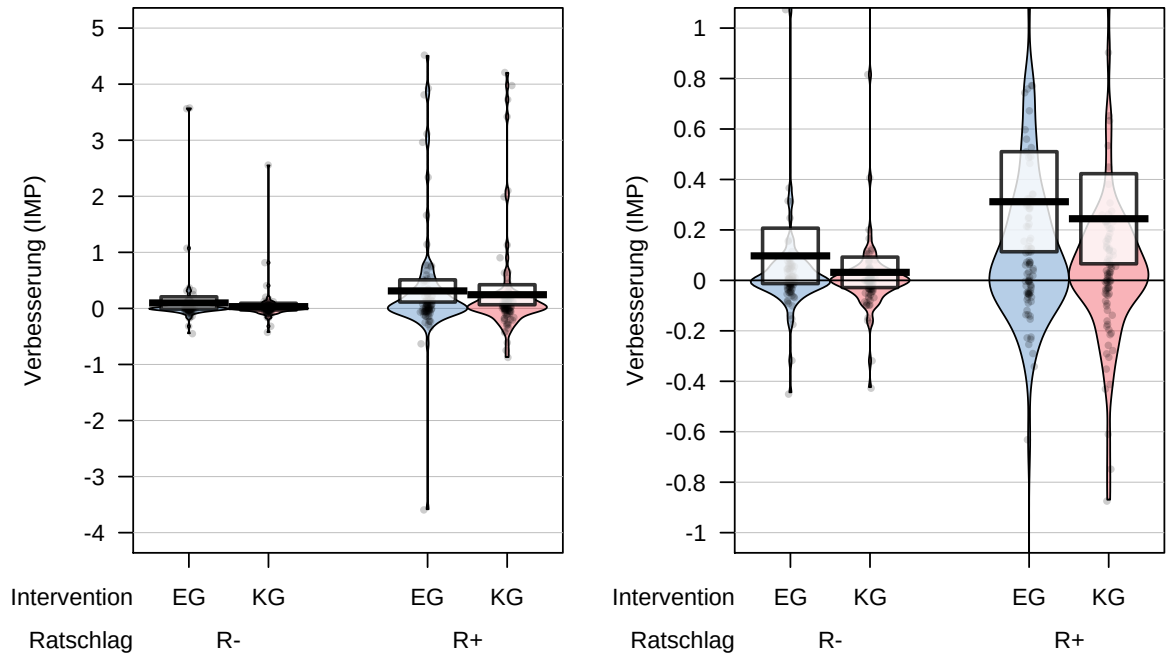
[24] Bei 15 Versuchspersonen waren alle Ratschläge hilfreich und keine nicht hilfreich, weshalb diese Personen bei der Analyse nicht berücksichtigt wurden. Dies entspricht jedoch der Annahme, dass die meisten repräsentativen Ratschläge hilfreich sind (in Studie 8 waren es 71% der Ratschläge). Zudem sollte auch H11 ursprünglich im Rahmen einer Multi-Level Mediationsanalyse getestet werden (siehe P8), bei der sich qualitativ die gleichen Ergebnisse zeigten.

Wirksamkeit der Intervention im JAS sind jedoch nach der Analyse der Ratschlagsnutzung nicht gegeben.

**Urteilsverbesserung.** Für einen direkten Test der Wirksamkeit von Ratschlägen (H14) und der spezifischen Wirksamkeit der Intervention (H15) wurden geplante Kontraste auf Basis der IMP Scores präregistriert (siehe P8). Für die Kontrastanalyse wurden drei orthogonale Kontraste für das 2 x 2-Design definiert: Ein Kontrast für den Haupteffekt der Ratschlagsmanipulation, ein Kontrast für den Haupteffekt der Intervention und ein dritter für deren Interaktion. Für den ersten und dritten Kontrast wurden mit H14 und H15 explizite Hypothesen aufgestellt, die jeweils postulieren, dass Ratschläge zu einer stärkeren Verbesserung von Urteilen führen und dieser Vorteil mit Intervention größer ausfällt. Diese beiden Kontraste wurden einseitig getestet (siehe P8). Alle drei Kontraste wurden im Rahmen einer einfachen OLS Regression mit den IMP Scores als abhängige Variable getestet. Die Kontrastanalyse zeigte eine signifikant größere Verbesserung in den Bedingungen mit Ratschlägen (EG/R+: $M$ = 91.32, $SD$ = 287.33; KG/R+: $M$ = 74.17, $SD$ = 262.22) im Vergleich zu den Bedingungen ohne Ratschlag (EG/R-: $M$ = 29.41, $SD$ = 162.19; KG/R-: $M$ = 10.06, $SD$ = 87.54), $t(371) = 2.83$, $p = .002$ (einseitig; P8), $d$ = 0.29, jedoch keinen Haupteffekt für die Intervention, $t(371) = 0.82$, $p = .413$ (zweiseitig; P8), $d$ = 0.09, und auch keinen Interaktionseffekt, $t(371) = -0.05$, $p = .520$ (einseitig; P8), $p = .961$ (zweiseitig), $d$ = -0.01. H14 kann damit angenommen werden, wobei sich keine Evidenz für H15 zeigte. Bei einer deskriptiven Betrachtung der Daten zeigte sich, dass es relativ viele Ausreißer bei dieser Analyse gab, wobei bei einer näheren Betrachtung der Verteilungen sichtbar ist, dass diese das Ergebnis nicht wesentlich beeinflusst haben (siehe Abbildung 16).

**Abbildung 16**

*Urteilsverbesserung (IMP) der Bedingungen von Studie 8*




*Anmerkung.* Mittlere Urteilsverbesserung (schwarze Linien) nach Interventions- und Ratschlagsbedingung mit Begrenzung der Y-Achse bei -1 und 1 (rechte Bildhälfte) und ohne Begrenzung (linke Bildhälfte). Abgebildet sind zudem 95% Konfidenzintervalle (weiße Rechtecke), Verteilungsdichte („Geigenplots“) und IMP Scores einzelner Versuchspersonen (schwarze Punkte). EG = Intervention/Experimentalgruppe. KG = Kontrollgruppe. R+ = Bedingung mit Ratschlag. R- = Bedingung ohne Ratschlag. IMP = Improvement Score auf Basis der absoluten Fehlerdifferenzen.

Auch in Studie 8 konnte die Intervention damit nicht zur Verbesserung der Urteilsprozesse genutzt werden, und zwar weder im individuellen Kontext noch im Ratschlagskontext. Für den individuellen Kontext war dieses Ergebnis erwartbar, da hier keine explizite Korrekturaufforderung erfolgte. Für die Intervention beim JAS entspricht dies dem Befund, dass mit dem Weisheitsbewusstsein die hilfreichen Ratschläge nicht stärker gewichtet wurden als die nicht hilfreichen Ratschläge. Mit dem Ausbleiben der Interventionseffekte muss auch formal H16 abgelehnt werden, die besagt, dass die Verbesserung in der EG/R+ größer ausfällt als in der KG/R+. Entsprechend entfällt auch die Testung der Mediation dieses Effekts (H17).

Zur besseren Einordnung dieser Effekte wurde zuletzt getestet, ob sich die Versuchspersonen innerhalb ihrer jeweiligen Bedingungen im Schnitt verbesserten oder verschlechterten. Diese Tests erfolgten analog zu den Tests der Verbesserungen im vorherigen Abschnitt, wobei im Einklang mit den Hypothesen zu Studie 8 nur für die Ratschlagsbedingungen gerichtete Tests präregistriert wurden (siehe P8). In den Bedingungen ohne Ratschlag zeigte sich keine statistisch signifikante Verbesserung in der EG/R-, $t(93) = 1.76$, $p = .082$ (zweiseitig; P8), $d = 0.18$, und auch nicht in der KG/R-, $t(92) = 1.11$, $p = .270$ (zweiseitig; P8), $d = 0.11$. In den Bedingungen mit Ratschlag zeigte sich hingegen eine Verbesserung sowohl in der EG/R+, $t(94) = 3.10$, $p = .001$ (einseitig; P8), $d = 0.32$, als auch in der KG/R+, $t(92) = 2.73$, $p = .004$ (einseitig; P8), $d = 0.28$. Damit zeigte sich insgesamt, dass die Versuchspersonen ihre Urteile nur mit Ratschlägen verbessern konnten, sich jedoch in keiner der Bedingungen systematisch verschlechterten.

### 6.4. Diskussion

In diesem empirischen Abschnitt wurde eine Intervention auf Basis des Weisheitsbewusstseins auf ihre spezifische Wirksamkeit beim Umgang mit Ratschlägen getestet. Im Einklang mit der bisherigen Forschung zeigte sich, dass Menschen insgesamt ihre Urteile mit Hilfe von Ratschlägen verbessern konnten (Bonaccio & Dalal, 2006). Über diesen Effekt hinaus zeigte sich jedoch - trotz deutlichem Weisheitsbewusstsein im JAS - keinerlei Wirksamkeit der Intervention in der Verbesserung der Urteile, weder spezifisch für die Interventionsbedingung mit Ratschlägen noch insgesamt. Eine wesentliche fehlende Voraussetzung für die Wirksamkeit der Intervention war, dass das Weisheitsbewusstsein nicht in eine stärkere Gewichtung hilfreicher Ratschläge übersetzt werden konnte. Bei dem Ausbleiben dieses Effekts spielen vermutlich mehrere Faktoren eine Rolle. So bildet sich das Weisheitsbewusstsein nicht immer, sondern nur mehrheitlich aus; es weist in diesen Fällen auch nicht perfekt, sondern nur mehrheitlich in die Richtung des wahren Werts, und zuletzt verhalten sich Menschen zwar überwiegend, jedoch nicht in allen Fällen, konsistent mit dem Weisheitsbewusstsein. Insgesamt gewichteten die Versuchspersonen damit weder in der Ratschlagsbedingung mit Intervention noch in der ohne Intervention die hilfreichen Ratschläge stärker als die nicht hilfreichen Ratschläge. Dieser Befund steht im Kontrast zu dem Befund von Yaniv (2004b), der eine stärkere Gewichtung hilfreicher Ratschläge berichtete. Ein entscheidender Unterschied könnte hierbei sein, dass Yaniv sowohl die

Abweichung der Ratschläge als auch deren Richtung direkt manipulierte, während in Studie 8 repräsentative Ratschläge verwendet wurden. Anhand dieser manipulierten Ratschläge (keine von echten Versuchspersonen) ist es vermutlich leichter zu erkennen, ob Ratschläge hilfreich sind oder nicht, insbesondere wenn diese stark in die falsche Richtung abweichen. In einem repräsentativen Set von echten Ratschlägen (wie in Studie 8), sollten solche Ratschläge vergleichsweise selten vorkommen, da wie bereits erwähnt, die meisten Ratschläge hilfreich sind (siehe Abschnitt 6.2) und die nicht hilfreichen vermutlich nicht besonders extrem in die falsche Richtung abweichen. In Studie 8 waren beispielsweise 71% der Ratschläge hilfreich.

Als positiver explorativer Befund zeigte sich auf der anderen Seite, dass das Weisheitsbewusstsein mit Ratschlägen deutlich ausgeprägter ausfiel als ohne Ratschläge. Ein Grund hierfür könnte sein, dass Menschen durch Ratschläge neue Informationen erhalten, die sie auch bei späteren Richtungsurteilen verwenden können. Die Person aus dem Einführungsbeispiel aus Bayern könnte beispielsweise, nach der Interkation mit der Person aus Schleswig-Holstein, ein früheres eigenes Urteil zu einer weiteren Stadt aus Niedersachsen bewerten (z.B. Göttingen). Hierfür würden ihr dann nicht mehr nur die Städte aus Oberfranken zur Verfügung stehen, die sie vor der Interaktion bereits kannte, sondern auch, dass Celle von einer anderen Person auf 70.000 geschätzt wurde. Im Einklang hiermit zeigte sich bei Stern (2017), dass durch Ratschläge Lerneffekte auftreten, die über die jeweilige Interaktion mit Ratgeber:innen hinausgehen, und die dafür sorgen, dass Menschen im Schnitt auch bei neuen Aufgaben aus der gleichen Domäne akkuratere quantitative Urteile abgegeben.

Trotz der positiven Effekte von Ratschlägen auf das Weisheitsbewusstsein führt die Intervention auf Basis des Weisheitsbewusstseins nach derzeitigem Kenntnisstand nicht zu einer Verbesserung der Ratschlagsnutzung. Wie bereits erwähnt, ist ein Problem, dass Menschen mit dem Weisheitsbewusstsein nicht die hilfreichen Ratschläge stärker gewichten können als die nicht hilfreichen. Eine weitere wesentliche Annahme zur Ableitung der Wirksamkeit der Intervention im JAS war, dass Menschen ohne Intervention weniger neue Informationen aus ihrem Langzeitgedächtnis oder auf Basis der Bilder herzanziehen als mit Intervention. Die Daten in Studie 8 deuten jedoch darauf hin, dass dies nicht der Fall ist. So wurden die Ratschläge in der Kontrollgruppe insgesamt zu einem ähnlichen Maße gewichtet

wie in der Experimentalgruppe, deskriptiv sogar etwas stärker (ein mittlerer AT von 43% in der EG/R+ und 48% in der KG/R+). Hätten die Versuchspersonen in der Kontrollgruppe tatsächlich primär alte und nur in der Experimentalgruppe auch neue oder häufiger neue Informationen verwendet, sollte die Nutzung der Ratschläge in der Kontrollgruppe im Vergleich zur Experimentalgruppe deutlich geringer ausfallen. Der Grund hierfür ist, dass die neuen Informationen weniger das Initialurteil und eher den Ratschlag stützen sollten als die alten Informationen (siehe auch, Yaniv, 2004b; Yaniv & Kleinberger, 2000). Es lässt sich spekulieren, dass es keine qualitativen Unterschiede zwischen Interventionsbedingung und Kontrollgruppe gab und in beiden Bedingungen in ähnlichem Maße neue und alte Informationen bei der Bewertung und Gewichtung verwendet wurden. Das heißt, dass die Intervention vermutlich selbst dann nicht wirksam wäre, wenn die hilfreichen Ratschläge stärker gewichtet worden wären als die nicht hilfreichen, weil sich vermutlich in der Kontrollgruppe der gleiche Effekt gezeigt hätte.

Für den Umgang mit Ratschlägen bedarf es aus diesen Gründen gegebenenfalls einen anderen Interventionsansatz als einen auf Basis des Weisheitsbewusstseins. Das GSM liefert hierfür theoretisch neue Ansatzpunkte. So wurde mit dem Weisheitsbewusstsein bisher mit einem kognitionspsychologischen Ansatz versucht, die Nutzung von Ratschlägen zu verbessern, indem Menschen neue Informationen aus ihrem privaten Informationspool bei der Bewertung und Gewichtung von Ratschlägen einsetzen. Gegebenenfalls ist es jedoch notwendig, eine Ebene höher anzusetzen (auf der Ebene des privaten Informationspools), da auch bei der bestmöglichen Nutzung des privaten Pools gewisse Grenzen bestehen. So kann beispielsweise die Person aus Bayern nicht erkennen, dass der Ratschlag der Person aus Schleswig-Holstein hilfreich ist, wenn sie nicht die notwendigen Informationen aus ihrem privaten Informationspool ziehen kann (wenn sie z.B. nicht weiß, dass Bamberg 75.000 Einwohner:innen hat und zudem überhaupt keine entsprechenden Städte in anderen Regionen kennt). Hierfür wären gegebenenfalls sozialpsychologische Interventionen notwendig, bei denen Menschen in einen direkten Austausch treten und ihre privaten Informationen zielführend teilen (siehe auch Minson et al., 2011). Bevor im Rahmen zukünftiger Projekte weitere Interventionen untersucht werden, ist es jedoch wichtig, für das zugrundeliegende Rahmenmodell eine solide empirische Basis zu schaffen. Wie in

Abschnitt 4 angekündigt, werde ich den empirischen Teil dieser Arbeit mit einem kritischen metaanalytischen Test des GSM schließen.

## 7. Metaanalytischer Test des GSM

Mit dem GSM wurde von mir ein Modell entwickelt, mit welchem erklärt werden kann, wann Menschen die Richtung ihrer eigenen Fehler erkennen können und inwiefern im Anschluss (theoretisch) mit Korrekturen Verbesserungen der eigenen Urteile erzielt werden können. Wie eingangs erwähnt, handelt es sich hierbei in gewisser Hinsicht um einen ersten Anwendungsfall eines Modells, welches über seine hierarchische Sampling-Struktur verschiedene Annahmen aus unterschiedlichen Forschungsbereichen vereint (insbesondere zu Metakognitionen und zur Weisheit der Vielen). Es macht zudem neue Annahmen dahingehend, inwiefern Menschen diese Sampling-Prozesse bewusst auslösen können, was insbesondere für die Bewertung eigener Fehler eine kritische Annahme ist. Dieser Ansatz kann aber auch auf andere Formen von Urteilen angewendet werden, bei denen Teile dieser hierarchischen Struktur häufig nicht berücksichtigt werden. So wird beispielsweise das Resampling neuer Informationen auf der untersten Ebene des GSM (Resampling neuer Infos aus dem eigenen privaten Informationspool) vornehmlich im Kontext der Weisheit der Vielen in einer Person untersucht (z.B., Vul & Pashler, 2008), während Beschränkungen auf höheren Ebenen (die Verteilung des Wissens in der Population) primär in metakognitiven Ansätzen untersucht werden (z.B., Koriat, 2012a, 2019). Für ein Gesamtverständnis von Urteilsprozessen ist es jedoch wichtig, alle Ebenen zu berücksichtigen und damit lokalisieren zu können, auf welchen Ebenen die Fehler entstehen und inwiefern diese mit Interventionen korrigiert werden können. Bevor ich jedoch vertiefend diskutiere, welche Konsequenzen dies für die Forschung mit sich bringt (siehe hierfür Abschnitt 8.1), sollte zunächst anhand des Weisheitsbewusstseins kritisch gefragt werden, inwiefern sich zu diesem Zeitpunkt überhaupt Evidenz für die postulierten Beschränkungen der unterschiedlichen Ebenen zeigt.

Nach Stand der acht Studien zeigt sich mit dem Weisheitsbewusstsein (H1) deutliche Evidenz für die Fähigkeit, neue (vorher ungenutzte) Informationen aus einem privaten Informationspool heranzuziehen (Evidenz für Beschränkung von Ebene 4; siehe Abbildung 1). Zudem zeigen die Ergebnisse von Studie 2, dass dieser private Informationspool selbst begrenzt ist und das Weisheitsbewusstsein beispielsweise nicht so stark ausgeprägt ist wie die Erkennung der Richtung von Fehlern bei einer anderen Person (Evidenz für Beschränkung

von Ebene 3). Es fehlt jedoch noch ein kritischer und vor allem statistisch starker Test der Vorhersage, dass alle diese Urteile durch den geteilten Informationspool beschränkt sind, auf denen alle Urteile zurückgehen (Evidenz für Beschränkung von Ebene 2). Wie bereits in Abschnitt 2 und 4 erwähnt, ließe sich diese Vorhersage mit dem Weisheitsbewusstsein daran testen, dass Menschen, wenn sie ein Urteil zwischen wahrem Wert und Populationsmittelwert treffen, eher in Richtung des Populationsmittelwerts tendieren (Hypothese H2b). Die ersten Studien in Abschnitt 4 liefern insgesamt Hinweise, dass dies der Fall ist. Da diese Fälle jedoch vergleichsweise selten auftreten, kann diese Frage erst jetzt mit der notwendigen Teststärke im Rahmen einer Metaanalyse beantwortet werden (wie in Abschnitt 4 angekündigt).

Ziel der Metaanalyse ist es, die wesentlichen Vorhersagen des GSM (H1, H2a und H2b) studien- und aufgabenübergreifend mit großer statistischer Power zu testen. Die metaanalytischen Tests für H1 und H2a werden lediglich im Sinne einer vollständigen Darstellung berichtet, da sich für diese Hypothesen bereits bei der Analyse der einzelnen Studien deutliche und vor allem sehr konsistente Evidenz zeigte (siehe Tabelle 2). Kritisch ist insbesondere der metaanalytische Test von Hypothese 2b, die bisher in den einzelnen Studien nur mit einer vergleichsweisen geringen statistischen Power untersucht werden konnte (siehe Tabelle 3 für die Ergebnisse der einzelnen Studien). Sofern sich metaanalytisch Evidenz für das Modell zeigt, kann ein grundlegender theoretischer Beitrag zur Erkennung und Nutzung eigener Korrekturpotentiale (und deren Limitationen) geleistet werden. Zudem zeigt sich damit Evidenz für einen Ansatz, aus dem später (in Abschnitt 8) weitere Implikationen für Theorie und Praxis abgeleitet werden können.

Für die metaanalytische Untersuchung der Richtungsurteile wurden mehrere logistische Multilevel-Modelle gefittet, bei denen nur die expliziten Richtungsurteile (ohne „weder noch“-Urteile) untersucht wurden. Entsprechend wurden auch keine Kontrollgruppen ohne Richtungsurteile in die Analyse aufgenommen. Zudem wurde diejenige Bedingung in Studie 2 ausgeschlossen, bei der die Richtungsurteile relativ zu fremden Urteilen getroffen wurden. Gleiches gilt für die Bedingung mit Richtungsurteilen und Ratschlägen in Studie 8, da es Hinweise dafür gab, dass Menschen aus den Ratschlägen zusätzliche Informationen für ihre Richtungsurteile gewinnen konnten (siehe Abschnitt 6). Abgesehen hiervon wurden alle Daten berücksichtigt. Der Test der Hypothesen erfolgte über

den Intercept ($b_0$) der logistischen Multilevel-Modelle. In allen Multilevel-Modellen wurden Random Intercepts modelliert mit den Studien, den Versuchspersonen genestet in den Studien und den unterschiedlichen Stimuli als Gruppierungsvariablen. Zuletzt wurden immer maximale Modelle angestrebt, bei denen auch alle möglichen Random Slopes modelliert wurden (Barr et al., 2013), wobei die Komplexität dieser Modelle reduziert wurde, wenn die Daten eine entsprechend komplexe Modellstruktur nicht zuließen (Matuschek et al., 2017).

Für einen aufgaben- und studienübergreifenden Test des Weisheitsbewusstseins (H1) wurde zunächst mit einem logistischen Intercept-Only-Modell vorhergesagt, ob die Versuchspersonen im Schnitt mit ihren expliziten Richtungsurteilen eher die Richtung des wahren Werts (1) oder die andere Richtung wählten (0). Äquivalent zu der TDR wurden hierbei auch alle Richtungsurteile als falsch (0) kodiert, bei denen das erste Urteil mit dem wahren Wert übereinstimmte und trotzdem eine explizite Richtung gewählt wurde. Anders als bei den TDRs wurden die kodierten Werte vor der Analyse nicht pro Person aggregiert. Ein Fixed Intercept größer null würde nach diesem Modell indizieren, dass Menschen eher die Richtung des wahren Werts (1) als die andere Richtung (0) wählten. Insgesamt umfasste diese Analyse 21,009 Beobachtungen von $N$ = 1,254 Versuchspersonen über 158 unterschiedliche Stimuli aus verschiedenen Urteilsdomänen. Über alle Beobachtungen hinweg wurde in 57% der Fälle die Richtung des wahren Werts gewählt. Der Fixed Intercept des Modells lag über 0, $b_0 = 0.27$, z = 5.74, $p < .001$ (zweiseitig). Im Einklang mit den Ergebnissen der einzelnen Studien zeigte sich deutliche Evidenz für H1.

Als nächstes wurde H2a anhand der gleichen Daten mit einer strukturell äquivalenten Analyse getestet, bei der vorhergesagt wurde, ob die expliziten Richtungsurteile eher in der Richtung der (gerundeten) Populationsmittelwerte der jeweiligen Studie lagen (1) oder eher in der anderen Richtung (0). Urteile, die mit dem Populationsmittelwert übereinstimmten, wurden äquivalent zu den CDRs als 0 kodiert. Entsprechend indiziert ein Fixed Intercept größer 0, dass eher die Richtung des Populationsmittelwerts (1) als die andere Richtung (0) gewählt wurde. Über alle Beobachtungen hinweg wurde im Schnitt in 60% der Fälle die Richtung des Populationsmittelwerts gewählt. Auch in dieser Analyse lag der Fixed Intercept über 0, $b_0 = 0.40$, $z = 11.30$, $p < .001$ (zweiseitig), womit sich auch deutliche aufgabenübergreifende Evidenz für H2a zeigte.

Zum Test von H2b wurde in das letzte Intercept-Only-Modell ein zusätzlicher dichotomer Prädiktor ($b_1$) mit aufgenommen, mit dem kodiert wurde, ob Populationsmittelwert und wahrer Wert ausgehend vom ersten Urteil in der gleichen Richtung lagen (als 1 kodiert; 74% der Beobachtungen) oder in unterschiedlichen Richtungen lagen (als 0 kodiert; 22% der Beobachtungen). Hierfür wurden zusätzlich alle Beobachtungen ausgeschlossen, bei denen das ursprüngliche quantitative Urteil entweder mit dem wahren Wert oder dem Populationsmittelwert übereinstimmte (4%). In diesem Modell wurden zunächst alle möglichen Random Slopes modelliert (Barr et al., 2013; mit Studien, Versuchspersonen genestet in Studien und den Stimuli als Gruppierungsvariablen), wobei die Komplexität des Modells aufgrund eines singulären Fits (Indikator für Overfitting) reduziert werden musste. Hierfür wurde der Random Slope für die Studien entfernt, womit sich die Ergebnisse nicht qualitativ änderten. In dem reduzierten Modell zeigte sich für den Fixed Intercept, dass dieser über 0 lag, $b_0 = 0.27$, $z = 4.77$, $p < .001$ (zweiseitig), was konkret heißt, dass Menschen auch zum Populationsmittelwert tendierten, wenn der wahre Wert in der anderen Richtung lag ($b_1 = 0$). Im Schnitt lagen in diesen Fällen 42% der Richtungsurteile in Richtung des wahren Werts und 58% in der Richtung des Populationsmittelwerts. H2b kann auf Basis dieser eindeutigen Evidenz angenommen werden. Es zeigte sich in diesem Modell jedoch auch ein statistisch signifikanter positiver Effekt für den neuen Prädiktor, $b_1 = 0.24$, $z = 3.99$, $p < .001$ (zweiseitig), nach dem die Richtungsurteile noch häufiger in der Richtung des Populationsmittelwerts lagen, wenn auch der wahre Wert in dieser Richtung lag. Insgesamt lagen hier 63% der Richtungsurteile in der Richtung von wahrem Wert und Populationsmittelwert.

Eine Erklärung für den Effekt des neuen Prädiktors könnte ein „Regression zur Mitte“-Effekt auf Informationsebene sein (Fiedler & Krueger, 2012), der in diesem Zusammenhang bereits in Abschnitt 4 kurz diskutiert wurde. Hiernach sollten neue Informationen häufiger in die Richtung des Populationsmittelwert zeigen, je größer die Abweichung des ersten Urteils vom Populationsmittelwert ist. Entsprechend sollte auch mit steigender Abweichung die Wahrscheinlichkeit steigen, die Richtung des Populationsmittelwerts mit einem expliziten Richtungsurteil zu wählen. Nach dieser Erklärung tendieren Menschen bei Fällen mit Populationsmittelwert und wahrem Wert in unterschiedlichen Richtungen weniger stark zum Populationsmittelwert, weil der Abstand zum Populationsmittelwert in diesen Fällen

relativ gering ist. Die Urteile liegen in diesen Fällen *zwischen* wahrem Wert und Populationsmittelwert und sind deshalb besonders akkurat, beziehungsweise liegen sie nah am Populationsmittelwert. Eine Voraussetzung ist natürlich, dass Populationsmittelwert und wahrer Wert relativ nah beieinander liegen, was bekanntermaßen häufig der Fall ist und mit der Weisheit der Vielen mehrfach dokumentiert wurde (Herzog et al. 2019; Larrick et al., 2012). Um diese Erklärung zu testen, wurde als weiterer Prädiktor ($b_2$) in das letzte Modell die *z*-standardisierte absolute Abweichung der ersten Urteile vom Populationsmittelwert mit aufgenommen. Die Standardisierung erfolgte anhand des Mittelwerts und der Standardabweichung der Abweichung pro Studie und Stimulus. Zunächst wurden auch für dieses Modell alle möglichen Random Intercepts und Slopes modelliert, was in einem singulären Fit resultierte. Um die Komplexität des Modells zu reduzieren, wurden deshalb im ersten Schritt die beiden Random Slopes für die Studien entfernt (Abweichung vom Populationsmittelwert und Richtung des wahren Werts) und im zweiten Schritt zudem der Random Slope der standardisierten Abweichung vom Populationsmittelwert für die Stimuli. Die Ergebnisse dieses reduzierten Modells unterschieden sich nicht qualitativ von den beiden komplexeren Modellen. Bei dem reduzierten Modell zeigte sich im Einklang mit den vorherigen Ergebnissen, dass der Fixed Intercept positiv von null abwich, $b_0 = 0.41$, $z = 6.79$, $p < .001$ (zweiseitig). In diesem Modell zeigte sich jedoch kein Effekt für die Richtung des wahren Werts (gleiche vs. unterschiedliche Richtung), $b_1 = 0.08$, $z = 1.35$, $p = .176$ (zweiseitig), aber ein positiver Effekt für die Abweichung der Initialurteile vom Populationsmittelwert, $b_2 = 0.27$, $z = 10.80$, $p < .001$ (zweiseitig). Damit verschwindet der Effekt für die Richtung des wahren Werts, wenn statistisch für die Abweichung vom Populationsmittelwert kontrolliert wird, genau wie bei einer Mediation der Effekt der unabhängigen Variable verschwindet, wenn für den Mediator kontrolliert wird. Diese Ergebnisse sind damit konsistent mit den Annahmen des GSM, dass mit den Richtungsurteilen die Richtung des Populationsmittelwert detektiert wird, wobei dies umso eher gegeben ist, je größer die Abweichung vom Populationsmittelwert ist. Darüber hinaus zeigte sich kein Effekt für die zusätzlich kodierte Richtung des wahren Werts.

Die Ergebnisse der metaanalytischen Tests untermauern und präzisieren insgesamt die Ergebnisse der individuellen Studien. In einer kritischen und teststarken Analyse konnten die aus dem GSM abgeleiteten Hypothesen vollumfänglich gestützt werden. Die Menschen

detektierten mit ihren Richtungsurteilen eher den Populationsmittelwert aller Urteile als den wahren Wert, wobei diese meist in der gleichen Richtung lagen. Damit bekommt der Name des Weisheitsbewusstseins seine zweite Bedeutung, da hiermit die Weisheit der Vielen detektiert wird, auch wenn explizit nach der Richtung des wahren Werts gefragt wird. Es ist jedoch auch im eigentlichen Wortsinn selbst *weise*, da damit insgesamt mehrheitlich die Richtung des wahren Werts erkannt wird. Wesentlich zentraler ist jedoch, dass sich hiermit Evidenz für die angenommene hierarchische Struktur des GSM zeigt, die besagt, dass Menschen bei der Urteilsbildung und -bewertung indirekt auf den gleichen geteilten Informationspool zugreifen, auf den auch andere Entscheidungsträger:innen der Population zugreifen. Mit einem solchen Modell lässt sich beschreiben, wann Menschen ihre Fehler erkennen können und wann nicht. Hierzu sind Menschen nach dem GSM in der Lage, wenn der geteilte Informationspool nicht systematisch verzerrt ist, beziehungsweise wenn sich die Weisheit der Vielen zeigt. Zudem werden wesentliche Annahme bisheriger hierarchischer Sampling-Modelle gestützt (z.B., Hourihan & Benjamin, 2010; Koriat, 2012a), die in dem GSM zusammengefasst werden und zu einem umfassenden Verständnis der Urteilsbildung beitragen können. Im Rahmen einer allgemeinen Diskussion meiner Ergebnisse werde ich detaillierter auf die theoretischen Implikationen dieser Erkenntnisse eingehen.

## 8. Übergreifende Diskussion

Quantitative Urteile sind eine wesentliche Grundlage für alltägliche und größere Entscheidungen. Deshalb ist es wichtig, dass Fehler bei solchen Urteilen rechtzeitig erkannt und korrigiert werden. Das erste Ziel dieser Arbeit war zu untersuchen, ob Menschen in der Lage sind, selbst zu erkennen, in welche Richtung sie ihre eigenen Urteile korrigieren sollten. Das zweite Ziel war hierauf basierend neue Interventionen zu entwickeln, die Menschen helfen, dieses Potential (sofern es sich nachweisen lässt) zur Verbesserung ihrer quantitativen Urteile zu nutzen. Zur Untersuchung dieser Fragestellungen wurde auf Basis vorheriger Arbeiten (z.B., Juslin et al., 2007; Koriat, 2012a; Vul & Pashler, 2008) ein hierarchisches Sampling-Modell zur Urteilsbildung entwickelt (das GSM), nach dem Menschen mehrfach Informationen aus einer geteilten Informationsverteilung ziehen, hierbei jedoch auf unterschiedlichen Ebenen in der Menge der verfügbaren (oder verarbeitbaren) Informationen limitiert sind. Als wesentliche Grundlage dieser Arbeit wurde aus dem GSM ein Phänomen abgeleitet, das von mir als Weisheitsbewusstsein bezeichnet

wird. Damit ist gemeint, dass Menschen den wahren Wert häufig eher über oder eher unter ihrem eigenen ersten Urteil vermuten, und, sofern sie solch ein explizites Richtungsurteil treffen, überzufällig häufig damit richtig liegen. Das Weisheitsbewusstsein lässt sich aus den Beschränkungen auf der untersten Ebene des GSM ableiten, indem Menschen zur Bildung eines ersten quantitativen Urteils und zur Bewertung dieses Urteils jeweils eine begrenzte und teilweise unterschiedliche Menge von Informationen aus einem privaten Informationspool ziehen. Sofern die gezogenen Informationen nicht systematisch verzerrt sind, sollten Menschen hierbei ein Weisheitsbewusstsein ausbilden. Das Weisheitsbewusstsein wurde in dieser Arbeit in acht Studien und zuletzt metaanalytisch getestet und im Rahmen mehrerer Interventionsstudien hinsichtlich seiner praktischen Nutzbarkeit untersucht. Die hieraus resultierenden Erkenntnisse werde ich in den folgenden Abschnitten hinsichtlich theoretischer und praktischer Implikationen diskutieren, wobei ich zunächst stärker auf die Implikationen des GSM und später stärker auf die Implikationen für Interventionen zur Korrektur von Urteilen eingehen werde.

### 8.1. Implikationen des GSM

In dieser Arbeit zeigte sich Evidenz für wesentliche Vorhersagen des GSM, nach dem Menschen indirekt Informationen aus einem geteilten Informationspool ziehen, um Urteile zu treffen und auch ihre eigenen oder fremde Urteile zu bewerten. Der wesentliche Test des Modells erfolgte in dieser Arbeit anhand des Weisheitsbewusstseins. Diese Fähigkeit wurde anhand eines Studiendesigns getestet, bei dem die Versuchspersonen nach der Abgabe einer Reihe quantitativer Urteile in einer zweiten Reihe von Richtungsurteilen angeben sollten, ob sie den wahren Wert (1) eher über, (2) eher unter oder (3) weder eher über noch eher unter ihrem ersten Urteil vermuteten. Dabei wählten Menschen zum einen meist in deutlich mehr als der Hälfte der Fälle eine der beiden Richtungen und lagen zudem, im Sinne des Weisheitsbewusstseins, überzufällig häufig (in mehr als 50% der Fälle) mit ihren expliziten Richtungsurteilen richtig. Das Weisheitsbewusstsein erwies sich als sehr robustes Phänomen, das sich in sieben von acht Studien und zuletzt metaanalytisch nachweisen ließ.

Das Weisheitsbewusstsein geht damit in einem wesentlichen Schritt über die bisherigen Befunde hinaus, indem es zeigt, dass Menschen auch bewusst die Richtung von Fehlern erkennen können. Dieser Schritt ist deshalb bemerkenswert, weil in der bisherigen

Literatur häufig davon ausgegangen wird, dass Menschen *naiv* gegenüber den Ursprüngen ihrer Fehler sind (Fiedler & Juslin, 2006b) und bei der Beurteilung der Richtung eines wahren Werts keine neuen Informationen heranziehen können, um die Richtung ihrer eigenen Fehler einzuschätzen (Gaertig & Simmons, 2021). Das Weisheitsbewusstsein steht im Einklang mit der bisherigen Literatur zur Weisheit der Vielen in einer Person (Herzog & Hertwig, 2014a), die zeigt, dass Menschen selbst zwei teilweise unabhängige quantitative Urteile zum gleichen Sachverhalt abgeben können, die gemittelt akkurater sind als die beiden Einzelurteile. Dieser Befund lässt sich mit dem GSM damit erklären, dass Menschen mehrfach auf ihren privaten Informationspool zugreifen können und damit teilweise neue unabhängige Informationen über den wahren Wert gewinnen. Das GSM geht darüber hinaus, weil Menschen nach dem GSM diese Sampling-Prozesse zur Urteilsbewertung bewusst selbst auslösen können und damit auch die Richtung ihrer eigenen Fehler erkennen können.

Eine Fähigkeit wie das Weisheitsbewusstsein lässt sich grundsätzlich auch mit anderen Erklärungsansätzen ableiten. Eine weitere Möglichkeit wäre zum Beispiel, dass der Urteilsprozess einem Ankerprozess gleicht, bei dem Menschen von einem meist nicht sonderlich informativen Referenzwert starten (dem Anker), der in die (meist korrekte) Richtung nur so weit angepasst wird, bis ihnen ihr Urteil hinreichend akkurat erscheint (Epley & Gilovich, 2006). Sofern dieser Prozess bewusst abläuft, könnte im nächsten Schritt eingeschätzt werden, in welcher Richtung der wahrer Wert eher liegt (Krueger & Chen, 2014). Solche Prozesse sind grundsätzlich vereinbar mit dem GSM, wobei ich bei der Herleitung des Modells diesbezüglich zunächst nur vereinfachende Annahmen gemacht habe (Bildung eines Mittelwerts über eine begrenzte Anzahl von Informationen). So könnten Menschen beispielsweise als Anker eine erste Information abrufen, die anhand von neuen Informationen immer weiter angepasst wird. Der entscheidende Beitrag des GSM liegt jedoch darin, dass die dabei herangezogenen Informationen einem hierarchischen Sampling-Prozess folgen, der auf mehreren Ebenen limitiert ist. Damit kann mit dem Modell genau beschrieben werden, welche Art von Informationsdefizit mit dem Weisheitsbewusstsein erkannt wird und inwiefern weiterhin deutliche Grenzen bestehen (bspw., wenn der geteilte Informationspool systematische verzerrt ist). Die konkreten Annahmen des GSM werde ich diesbezüglich in einem eigenen Unterabschnitt anhand des Weisheitsbewusstseins

diskutieren. Bei dem Weisheitsbewusstsein handelt es sich, wie erwähnt, um einen ersten Anwendungsfall, anhand dessen die hierarchische Struktur des GSM erstmalig getestet werden kann. Das GSM erhebt jedoch den Anspruch als generelles Modell der Urteilsbildung zu fungieren, anhand dessen auch andere Urteilsprozesse und insbesondere Interventionen zur Verbesserung dieser Urteilsprozesse untersucht werden können. In einer einzelnen Arbeit kann eine solche Annahme kaum umfassend überprüft werden. Zudem würde ein umfassendes Review der bisherigen Urteilsforschung bezüglich dieser Annahmen zu weit von den zentralen Fragestellungen dieser Arbeit wegführen. Dennoch werde ich nach der Diskussion der Ebenen in einem weiteren Unterabschnitt zumindest kurz skizzieren, inwiefern sich diese Annahmen auf andere Urteilsphänomene und Interventionen übertragen lassen, um die Relevanz des GSM und der Evidenz für das GSM für bestehende und zukünftige Fragestellungen der Urteilsforschung zu unterstreichen.

### 8.1.1. Ebenen der Fehlerentstehung

Nach dem GSM entstehen Urteilsfehler durch Informationsverluste auf drei verschiedenen Ebenen, die maßgeblich dafür sind, ob Menschen ihre Fehler erkennen und korrigieren können (siehe Abbildung 1 in Abschnitt 2). Diese Ebenen spiegeln erstens wider, dass Wissen generell nur begrenzt in der Welt vorhanden ist (Ebene 2); zweitens, dass hiervon Einzelpersonen nur eine begrenzte Menge zugänglich ist (Ebene 3), die sie, drittens, nicht immer alle zeitgleich nutzen können (Ebene 4). Mit dem GSM lässt sich ableiten, dass Menschen eigene Fehler erkennen können, die entstehen, weil sie nicht immer alle Informationen nutzen, die sie zur Verfügung haben (Ebene 4). Evidenz hierfür zeigte sich mit dem Weisheitsbewusstsein, dass bereits ausführlicher diskutiert wurde. Sie können damit aber keine Fehler erkennen, die beispielsweise entstehen, weil bestimmte Informationen nicht gelernt und im Langzeitgedächtnis abgespeichert wurden (Ebene 3). Evidenz hierfür zeigte sich mit Studie 2, bei der sich zeigte, dass fremde Menschen besser in der Lage sind die Richtung der Fehler einer Person zu erkennen (bzw. die Richtung des wahren Werts). Fremde Menschen sollten nach dem GSM in ihrem Langzeitgedächtnis anteilig mehr neue Informationen zur Verfügung haben, die Menschen bei der Erkennung ihrer eigenen Fehler schlicht nicht zugänglich sind.

Zuletzt können Menschen nach dem GSM keine Fehler erkennen, die dadurch entstehen, dass die Gesamtmenge der verfügbaren Informationen (Ebene 2) teilweise verzerrt ist, worauf letztlich alle Urteile (auch das Weisheitsbewusstsein) beruhen. So zeigte sich, dass Menschen mit dem Weisheitsbewusstsein weniger den wahren Wert detektierten als vielmehr den Populationsmittelwert aller Urteile, mit dem sich das bestmögliche Urteil des geteilten Informationspools approximieren lässt. In allen Studien lagen die expliziten Richtungsurteile der Versuchspersonen überzufällig häufig in der Richtung des Populationsmittelwerts (H2a), in der auch meist der wahre Wert lag. Vor allem konnte zuletzt in einer Metaanalyse gezeigt werden, dass Menschen bei ihren Richtungsurteilen auch dann eher den Populationsmittelwert aller Urteile detektierten, wenn der wahre Wert in der entgegengesetzten Richtung lag (H2b). Die robusten positiven Befunde zur Weisheit der Vielen (Herzog et al., 2019; Larrick et al., 2012) zeigen, dass der Populationsmittelwert vieler Urteile meist sehr nah am wahren Wert liegt, was darauf hindeutet, dass bei den meisten Aufgaben der geteilte Informationspool weder besonders stark nach oben oder unten systematisch verzerrt ist. Dennoch handelt es sich hierbei um eine zentrale Limitation für die Erkennung und Korrektur von Fehlern, die nicht nur das Weisheitsbewusstsein, sondern auch die Weisheit der Vielen betrifft.

Insgesamt konnten mehrere wesentliche Annahmen des GSM zur Fehlerentstehung empirisch gestützt werden, die auch in metakognitiven Modellen zur Urteilsbildung (z.B., Juslin et al., 2007; Koriat, 2012a) und in der Literatur zur Weisheit der Vielen (z.B., Vul & Pashler, 2008) getroffen werden. Mit dem GSM konnte jedoch auch deutlich gemacht werden, welche Formen von Fehlern von Menschen selbst erkannt werden können. Wesentlich war hierfür die neue Annahme, dass Menschen die Sampling-Prozesse zur Urteilsbildung bewusst neu auslösen können und damit die Richtung ihrer Fehler erkennen können. Zusammen mit den bisherigen Befunden zu solchen Ansätzen kann das GSM zu einem umfassenden Verständnis von kognitiven und metakognitiven Urteilsprozessen beitragen und dem Anspruch gerecht werden, *generalisierende* Aussagen zu treffen, die auf unterschiedliche Formen von Urteilen und Urteilsphänomene zutreffen. Damit liefert das GSM auch konkrete Ansatzpunkte für die zukünftige Interventionsforschung zu Urteilsprozessen.

### 8.1.2. Implikationen für andere Formen von Urteilen und Interventionen

Das GSM liefert mit seiner hierarchischen Struktur und seinem Anspruch generelle Aussagen zu treffen, zunächst eine Taxonomie für die Ursprünge von Urteilsfehlern. Hiermit lässt sich sehr allgemein definieren, welche Formen von psychologischen Interventionen Urteilsprozesse verbessern können, je nach Ursprung der Fehler. So können Menschen beispielsweise mit dem Sampling neuer Informationen aus ihrem Langzeitgedächtnis auch nur die Informationsverluste kompensieren, die dadurch entstehen, dass genau diese Informationen vorher nicht angemessen berücksichtigt wurden. Üblicherweise kommen hierfür kognitionspsychologische Interventionen zum Einsatz, die Menschen anregen, die entsprechende Evidenz für die Korrektur dieser Fehler auch tatsächlich abzurufen und einzusetzen (siehe z.B. Mussweiler et al., 2000). Informationsverluste auf der nächsthöheren Ebene (durch die Begrenzung des privaten Informationspools), können hingegen häufig nur in sozialen Kontexten korrigiert werden, beispielsweise indem Menschen Ratschläge anderer Personen einholen oder Urteile und Entscheidungen in Gruppen treffen. Für diese Ebene bedarf es primär sozialpsychologischer Interventionen (siehe z.B. Schulz-Hardt et al., 2006), aber auch kognitionspsychologischer Interventionen (z.B. zur verbesserten Bewertung der Urteile anderer), um Informationsverluste im sozialen Austausch bestmöglich kompensieren zu können. Informationsverluste, die jedoch durch einen generellen Mangel an Informationen entstehen, können nicht ohne weiteres durch einen Abruf oder den Austausch vorhandener Informationen korrigiert werden. Dieses Wissen muss erst generiert werden, damit es auch zur Verbesserung von Urteilsprozessen genutzt werden kann.

Eine solche Taxonomie ist deshalb wichtig, weil die Interventionen, die im Rahmen der psychologischen Forschung zur Behebung bestimmter Probleme entwickelt werden, auch zu der Ebene passen müssen, auf der die Probleme entstehen. Teilweise werden jedoch die Ursprünge von Fehlern in der Forschung falsch zugeordnet, weshalb auch keine passenden Interventionen entwickelt werden können. Ein prominentes Beispiel ist hierfür die Literatur zur Urteilssicherheit (Confidence), bei der diese Passung in den Anfängen dieses Forschungsbereichs vernachlässigt wurde. Frühere Arbeiten legten nahe, dass Menschen unabhängig von der Form der Urteile und Abfrageformat der Urteilssicherheit zu Overconfidence neigen und sich bei ihren Urteilen generell sicherer sind als es gerechtfertigt ist (Lichtenstein et al., 1982). Neuere Arbeiten zeigen jedoch, dass Overconfidence massiv

von der Frage nach der Urteilssicherheit und der Auswahl von Aufgaben abhängt (Klayman et al., 2006). So zeigt sich insbesondere bei *dichotomen Urteilen* (bzw. Entscheidungen), dass Overconfidence sich primär bei der Auswahl von besonders irreführenden Aufgaben zeigt (bei denen die meisten Menschen falsch liegen), und nicht, wie früher angenommen, generell bei allen Aufgaben (Gigerenzer et al., 1991; Klayman et al., 2006; Koriat, 2012a). Somit wurde früher fälschlich angenommen, dass Overconfidence bei dichotomen Urteilen auf der untersten Ebene bei der Urteilsbildung durch verzerrte Kognitionen entsteht, wobei nach heutiger Sicht diese Fehler tatsächlich primär auf Ebene der geteilten Informationen entstehen, die bei manchen (irreführenden) Aufgaben verzerrt sein können. Bei Konfidenzintervallen zeigt sich jedoch auch bei einer repräsentativen Auswahl von Aufgaben ein robuster Overconfidence-Effekt, der vermutlich zumindest teilweise auf eine verzerrte Auswahl und Verarbeitung von Informationen auf unterster Ebene zurückzuführen ist (Klayman et al., 2006).

Das GSM liefert jedoch nicht nur eine allgemeine Taxonomie für Fehlerursachen, sondern auch sehr konkrete Ansatzpunkte für neue Interventionen und für die Verbesserung bestehender Interventionen. Wesentlich ist hierfür das Herausstellungsmerkmal des GSM, dass Menschen das Sampling von Informationen bewusst und zielgerichtet auslösen können. Ein Beispiel aus der Literatur zur Urteilssicherheit wäre das Phänomen der Overconfidence bei Konfidenzintervallen zu quantitativen Urteilen („In welchem Intervall liegt die Bevölkerungszahl von Celle mit 80% Wahrscheinlichkeit?“), bei denen sich, wie bereits erwähnt, auch bei einer repräsentativen Auswahl von Aufgaben Overconfidence zeigt (Klayman et al., 2006). Die Ursachen von Overconfidence lassen sich bei diesen Urteilen gut mit Sampling-Ansätzen erklären, nach denen Menschen durch die Beschränkungen ihres Arbeitsgedächtnisses zu wenige Informationen abrufen, um ihre Sicherheitsurteile zu treffen, die zudem in Richtung der anfänglichen Meinung verzerrt sein können und naiv verarbeitet werden (Juslin et al., 2007; Klayman et al., 2006; Soll & Klayman, 2004). Schätzt beispielsweise die Person aus Bayern die Bevölkerungszahl von Celle auf 30.000, sollten extreme Informationen, die für eine besonders niedrige oder eine besonders hohe Zahl sprechen, zu selten abgerufen werden, wodurch sich die Person sicherer ist als es gerechtfertigt wäre. Nach dem GSM können Menschen jedoch theoretisch auch zielgerichtet für das obere Limit ihres Konfidenzintervalls besonders hohe Informationen abrufen und in

einem zweiten Schritt besonders niedrige für das untere Limit, wobei es hierfür gegebenenfalls Interventionen mit expliziten Instruktionen bedarf (siehe auch Moore et al., 2015; Soll & Klayman, 2004). Das GSM zeigt jedoch auch Grenzen für solche Formen von Interventionen auf, da Menschen nicht Informationen gezielt abrufen können, die sie nicht in ihrem privaten Informationspool zur Verfügung haben. Ein Beispiel hierfür sind die Strategien auf Basis der Weisheit der Vielen in einer Person, bei denen Menschen üblicherweise per Incentivierung oder mit sehr expliziten Instruktionen angeregt werden, abweichende Urteile abzugeben (z.B., Gaertig & Simmons, 2021; Herzog & Hertwig, 2014b; Vul & Pashler, 2008). Nach den Annahmen des GSM können sich solche Maßnahmen als kontraproduktiv herausstellen, wenn Menschen schlicht keine zusätzlichen Informationen haben, um ein abweichendes Urteil zu treffen. Solche Maßnahmen wären nach dem GSM auch nicht notwendig, da Menschen ihre Sampling-Prozesse bewusst selbst auslösen können und hierzu nicht erst „gezwungen werden müssen".

Dieser Exkurs sollte kurz skizzenhaft darstellen, inwiefern sich das GSM auch auf andere Urteilsphänomene und Interventionen anwenden lässt. Im Zentrum dieser Arbeit standen jedoch andere Anwendungsfälle, nämlich die selbstständige Korrektur von Urteilen und der Umgang mit Ratschlägen mit Hilfe des Weisheitsbewusstseins, die ich in den nächsten beiden Abschnitten separat diskutieren werde.

### 8.2. Implikationen für Korrekturinterventionen im individuellen Urteilskontext

In dieser Arbeit wurde in einer Reihe von Studien systematisch getestet, inwiefern Menschen das Weisheitsbewusstsein mit Hilfe von Interventionen zur selbständigen Korrektur ihrer Urteile nutzen können. Die Interventionslogik bestand in allen diesen Studien darin, dass die Versuchspersonen nach Abgabe eines ersten Urteils im Rahmen der Abfrage des Weisheitsbewusstseins zu einer ausgeglichenen Informationssuche instruiert wurden und nach der Einschätzung der Richtung des wahren Werts zu einer Korrektur in die entsprechende Richtung aufgefordert wurden. Die konkreten Korrekturinstruktionen unterschieden sich hierbei teilweise zwischen den Studien. So wurde zunächst eine Variante getestet, bei der die Versuchspersonen ihre Urteile in die entsprechende Richtung anpassen sollten, bis sie das Gefühl hätten, dass der Wert weder eher in der einen noch in der anderen Richtung liege (Studie 4 & 5). Bei dieser Variante zeigte sich jedoch keinerlei Wirksamkeit der

Intervention, und zwar weder in einem besonders teststarken Innersubjektvergleich noch im Zwischensubjektvergleich zu einer Kontrollgruppe ohne Intervention. Tendenziell zeigte sich eher, dass sich die Akkuratheit der Urteile mit und ohne Intervention verschlechterte. Auch wenn diese Korrekturen zumindest in Studie 5 meist in die richtige Richtung erfolgten, zeigte sich, dass Menschen ihre Urteile zu stark korrigierten und damit eher verschlechterten. Diese Befunde stehen im Einklang mit der Literatur zur Weisheit der Vielen in einer Person, bei der die zweiten Urteile von Personen tendenziell weniger akkurat sind als die ersten (z.B., Fraundorf & Benjamin, 2014; Gaertig & Simmons, 2021; Vul & Pashler, 2008). Es lässt sich jedoch nicht abschließend klären, warum die Korrekturen tendenziell zu groß ausfielen, zumal bisherige Arbeiten zur Overconfidence (Klayman et al., 2006) oder Ratschlagsnutzung (z.B., Soll & Larrick, 2009) vermuten lassen, dass Menschen sehr überzeugt von ihren eigenen Urteilen sind und daher eher zu geringe Korrekturen vornehmen sollten. Eine mögliche Erklärung wäre ein Demand-Effekt oder *Good-Subject*-Effekt (Nichols & Maner, 2008), der sich dadurch ergibt, dass Menschen in einem experimentellen Setting aufgefordert werden, zwei Mal die gleiche Frage zu beantworten. Menschen könnten dies zumindest teilweise als Aufforderung verstehen, ein deutlich abweichendes zweites Urteil abzugeben, auch wenn explizit auf die Möglichkeit hingewiesen wird, dass auch das erste Urteil erneut eingegeben werden kann. So wurden insbesondere in Studie 4 häufig auch Urteile korrigiert, wenn Menschen den wahren Wert weder eher in der einen noch der anderen Richtung vermuteten.

In zwei weiteren Studien (Studie 6 & 7) wurde deshalb eine zweite Variante der Interventionslogik getestet, bei der die Versuchspersonen explizit aufgefordert wurden, ihre Urteile nur einen kleinen Schritt in die vorher eingeschätzte Richtung zu korrigieren. In Studie 6 zeigte sich zunächst, dass auch mit dieser Variante weder im Inner- noch im Zwischensubjektvergleich eine systematische Verbesserung der Urteile erzielt werden konnte. Auf der anderen Seite verschlechterten sich die Versuchspersonen auch nicht, anders als in der Kontrollgruppe ohne Intervention. In explorativen Folgeuntersuchungen stellte sich jedoch heraus, dass mit dieser Intervention zumindest bei historischen Schätzaufgaben, sowohl im Inner- als auch im Zwischensubjektvergleich, Erfolge in den Verbesserungen der Urteile erzielt werden konnten. Im Zwischensubjektvergleich jedoch nur, wenn für Unterschiede in der anfänglichen Akkuratheit kontrolliert wurde. Eine

Einschränkung ist daher, dass die finale Akkuratheit mit Intervention nicht auf einem höheren Niveau lag als in der Kontrollgruppe, wobei bei diesem Zwischensubjektvergleich von einer geringeren Testpower im Vergleich zu den Analysen der Urteilsverbesserungen auszugehen ist.

Drei Umstände lassen jedoch optimistisch stimmen, dass Menschen mit einer solchen Intervention tatsächlich ihre historischen Schätzungen verbessern können. Erstens konnte die Verbesserung im Innersubjektvergleich in einem einfachen Messwiederholungsdesign anhand neuer historischer Schätzaufgaben repliziert werden. Zweitens zeigte sich in keiner der Kontrollgruppen, dass Menschen ohne Intervention ihre Urteile verbessern können, sondern tendenziell ohne Intervention eher verschlechtern. Drittens stehen diese Ergebnisse im Einklang mit der Arbeit von Müller-Trede (2011), bei der die gleichen Aufgabenkategorien verwendet wurden und bei der sich die Weisheit der Vielen in einer Person als vergleichbarer Effekt auch nur bei den historischen Schätzungen zeigte (und auch bei Prozentschätzungen, die zusätzlich untersucht wurden). Nichtdestotrotz kann auf Basis der derzeitigen Datenlage noch keine Praxisempfehlung ausgesprochen werden. Hierfür bedarf es zum einen weiterer Replikationen, die eindeutige Evidenz bezüglich der Wirksamkeit der Intervention im Zwischensubjektvergleich liefern, und zudem eines besseren theoretischen Verständnisses, unter welchen Umständen hiermit eine Verbesserung der Urteile erreicht werden kann. Dieses Verständnis ist zwingend notwendig, damit die Intervention nicht nur für historische Schätzungen, sondern auch andere Aufgabentypen empfohlen werden kann, die in der Praxis häufiger unter Unsicherheit bearbeitet werden müssen (z.B. Prognosen).

Konkreter wurde bereits in Abschnitt 5 diskutiert, dass Menschen bei anderen Aufgaben als historischen Schätzungen teilweise nur schwer in der Lage sind, die Aufforderung zu kleinen Korrekturen auch adäquat umzusetzen. Hierfür ist nämlich ein gewisses Maß an Wissen über die Antwortmetrik notwendig, beziehungsweise Wissen über den Bereich plausibler Werte bei Aufgaben aus der entsprechenden Domäne (Brown & Siegler, 1993). Direkt ließe sich diese Erklärung in zukünftigen Studien anhand von mehreren Aufgaben mit größeren metrischen Fehlern aus der gleichen Domäne untersuchen, bei denen orthogonal zu der Intervention manipuliert wird, ob zu Beginn der Studie Referenzwerte zur Kalibrierung und Reduzierung der metrischen Fehler präsentiert werden (siehe z.B., Bonner et al., 2007; Laughlin et al., 1999). Ein geeigneter Aufgabentyp wäre zum

Beispiel die Schätzung von Distanzen europäischer Hauptstädte (siehe Stern et al., 2017). Zu Beginn der Studie könnte den Versuchspersonen eine kurze Luftliniendistanz (z.B. „Brüssel und Amsterdam“, 173 km)[25] und eine längere Luftliniendistanz (z.B. „Lissabon und Warschau“, 2.760 km) als Referenzwerte präsentiert werden, wobei es die Aufgabe der Versuchspersonen ist, neue Distanzen zu schätzen (z.B. „Paris und Berlin“, „Madrid und Wien“, etc.). Vor der Intervention (zu Phase 1) sollten Menschen in den Bedingungen mit Referenzwerten akkuratere Urteile abgeben als in den Bedingungen ohne Referenzwerte, beziehungsweise sollten sich die metrischen Fehler reduzieren. Nach der Intervention (zu Phase 2) sollten jedoch nur die Versuchspersonen mit Intervention zu kleinen Korrekturen *und* Referenzwerten ihre Urteile im Vergleich zu Phase 1 noch weiter verbessern können oder zumindest stärker als in den anderen drei Bedingungen. Damit würde sich Evidenz dafür zeigen, dass sowohl ein grundlegendes metrisches Wissen als auch die entsprechende Intervention notwendig sind, damit Menschen ihre Urteile mit kleinen Schritten selbst verbessern können. Die Erklärung müsste jedoch abgelehnt werden, wenn allein die Intervention oder allein die Referenzwerte hinreichend sind, um eine Verbesserung der Urteile von Phase 1 zu Phase 2 zu erzielen oder sogar beides nicht notwendig ist. Eine solche Studie würde zum theoretischen Verständnis beitragen, unter welchen Umständen die Intervention zu Erfolgen führt. Auf Basis solcher Ergebnisse könnte gegebenenfalls auch eine besonders robuste kombinierte Intervention entwickelt werden, bei der die Intervention zusammen mit Referenzwerten angewendet wird und hierdurch aufgabenübergreifend Verbesserungen bewirkt werden. Eine solche Intervention wäre sehr praxistauglich, da Referenzwerte für viele Aufgabentypen leicht zu finden sind (z.B. alte Aktienkurse) und die Intervention selbst keinerlei zusätzlicher Informationen bedarf und eigenhändig von Menschen umgesetzt werden kann.

### 8.3. Implikationen für den Umgang mit Ratschlägen

In einer letzten Studie wurde die Interventionslogik auf Basis des Weisheitsbewusstseins auf einen Kontext angepasst, in dem Menschen Ratschläge anderer Personen erhielten. Durch die Intervention sollten hilfreiche Ratschläge (die in die richtige Richtung zeigen) stärker genutzt werden als nicht hilfreiche Ratschläge (die in die falsche

[25] Für diese und alle kommenden Distanzbeispiele wurden die wahren Werte über https://www.luftlinie.org/ (Georg, 2021) abgerufen.

Richtung zeigen). Es zeigte sich jedoch, dass Menschen auch mit der Intervention hierzu nicht systematisch in der Lage waren. Entsprechend konnte der Intervention keine Wirksamkeit konstatiert werden. In Einklang mit der bisherigen Forschung konnten Menschen zwar mit Ratschlägen ihre Urteile stärker verbessern als ohne (Bonaccio & Dalal, 2006); sie konnten jedoch mit Intervention nicht stärker von den Ratschlägen profitierten als ohne Intervention.

Damit stellt sich unmittelbar die Frage, ob und inwiefern eine Intervention auf Basis des Weisheitsbewusstseins zukünftig weiterentwickelt werden kann, um die Gewichtungsprozesse von Ratschlägen zu verbessern. Auf Basis der derzeitigen Erkenntnisse gestaltet es sich jedoch als schwierig, direkt neue Interventionsvarianten abzuleiten. Das Problem ist, dass wesentliche Annahmen nicht bestätigt werden konnten, die bei der Vorhersage der Wirksamkeit der Intervention zugrunde gelegt wurden. So zeigte sich zum einen keine Evidenz dafür, dass Menschen ohne Intervention wesentlich weniger neue Informationen heranziehen, um Ratschläge zu gewichten (Informationen, die noch nicht in das erste Urteil eingeflossen sind). Die Gewichtung der Ratschläge in der Kontrollgruppe hätte in diesem Fall insgesamt unter dem Niveau der Interventionsbedingung liegen sollen, doch deskriptiv war das Gegenteil der Fall. Zum anderen konnten diese „neuen“ Informationen, wie bereits erwähnt, in keiner der beiden Bedingungen zur stärkeren Gewichtung hilfreicher Ratschläge im Vergleich zu nicht hilfreichen Ratschlägen genutzt werden.

Deshalb bedarf es zunächst weiterer theoretischer Erkenntnisse, bevor weitere Implikationen für die Praxis abgeleitet werden können. Ein wichtiger erster Schritt wäre zu untersuchen, inwiefern neue Informationen bei der spontanen Gewichtung von Ratschlägen zum Einsatz kommen. Hierbei könnte sich herausstellen, dass die Bewertungsprozesse beim Umgang mit Ratschlägen durchaus funktional ablaufen und die Probleme eher auf andere Faktoren zurückzuführen sind. Beispielsweise wäre es möglich, dass Menschen bei der Bewertung von Ratschlägen durchaus vorher ungenutzte Informationen abrufen, um diese zu bewerten und zu gewichten (siehe auch Schultze et al., 2017). Teilweise sind hierfür nach dem GSM allerdings keine zusätzlichen Informationen vorhanden, weil der private Informationspool von Menschen begrenzt ist. In diesen Fällen können Menschen tatschlich nur die „alten“ Informationen nutzen, um Ratschläge zu bewerten, die bereits in das erste

Urteil eingeflossen sind, wodurch auch hilfreiche Ratschläge ignoriert werden würden (siehe auch Yaniv, 2004b). Für diese Fälle wären andere Interventionsansätze notwendig, um den Umgang mit Ratschlägen noch weiter zu verbessern, da mit dem Weisheitsbewusstsein keine Informationen abgerufen werden können, die nicht vorhanden sind. Vermutlich müssten Judges und Advisor in einen direkten Austausch treten, um ihre privaten Informationen zu teilen und die Finalurteile der Judges zu verbessern (siehe z.B. Minson et al., 2011). Diese Prozesse könnten mit neuen Interventionen noch weiter verbessert werden. Beispielsweise könnte ein künstlicher Dissens zwischen Judge und Advisor dafür sorgen, dass die Diskussionsintensität gesteigert wird und mehr private Informationen geteilt werden (Schulz-Hardt et al., 2006).

Anhand dieser Diskussion soll noch einmal deutlich werden, wie wichtig es ist, die Probleme bei Urteilsprozessen richtig zu lokalisieren und passgenaue Interventionen zu entwickeln (siehe Abschnitt 8.1.2.). Gegebenenfalls müssten in zukünftigen Untersuchungen zum JAS vermehrt sozialpsychologische Interventionen (wie die Erzeugung von Dissens) anstelle von kognitionspsychologischen Interventionen (wie die Intervention zum Weisheitsbewusstsein) zum Einsatz kommen.

### 8.4. Limitationen

Ausgangspunkt dieser Arbeit war die Frage, inwiefern Menschen ihre Fehler erkennen und korrigieren können. Zur Beantwortung dieser Frage wurde auf Basis bisheriger Arbeiten das GSM als generelles Modell zur Urteilbildung entwickelt und wesentliche Vorhersagen zur Modelltestung abgeleitet. Wie bei jeder theoretischen und empirischen Arbeit ergeben sich natürlich auch in dieser Arbeit Limitationen, die ich im Folgenden diskutieren werde. Zunächst werde ich kurz skizzieren, inwiefern sich Limitationen für das GSM als generelles Modell der Urteilsbildung ergeben. Diese Limitationen betreffen teilweise weniger die wesentlichen Fragestellungen in dieser Arbeit (z.B. das Weisheitsbewusstsein). Da ich jedoch in kurzer Form den Wert des GSM für andere Urteilsprozesse hervorgehoben habe, muss ich auch in kurzer Form auf die Limitationen des Modells eingehen, beziehungsweise, inwiefern das Modell für andere Urteilsprozesse noch weiter spezifiziert werden muss. Im Anschluss werde ich detaillierter auf die Limitationen

bezüglich des Weisheitsbewusstseins als zentrales neues Phänomen eingehen, welches nach dem GSM der Erkennung und Korrektur von Urteilsfehlern zugrunde liegt.

### 8.4.1. Limitationen des GSM

Zur Beschreibung der Urteilsprozesse mit dem GSM habe ich mehrere vereinfachende Annahmen gemacht. Diese betreffen zum einen die Ziehung von Informationen auf allen Ebenen (für die ich eine zufällige Ziehung angenommen habe) und die Integration dieser Informationen (für die ich einen einfachen gleichgewichteten Mittelwert angenommen habe). Mit diesen Annahmen ließ sich das Weisheitsbewusstsein ableiten und auch empirisch nachweisen. Es lassen sich jedoch andere Phänomene der Urteilsforschung nicht erklären, die diesbezüglich spezifischere Annahmen machen. Dies betrifft zum einen Ankereffekte, bei denen angenommen wird, dass selektiv Informationen herangezogen werden, die dem Anker entsprechen (Mussweiler & Strack, 1999), oder angenommen wird, dass bei der Integration nur unzureichend von Ankern adjustiert wird (Epley & Gilovich, 2006). Auch bei dem Phänomen der Overconfidence bei Konfidenzintervallen wird angenommen, dass diese teilweise durch einen selektiven Abruf nicht extremer Informationen zustande kommen (d.h., Informationen, die weder für besonders hohe noch besonders niedrige Werte sprechen; Klayman et al., 2006). Das GSM lässt sich jedoch grundsätzlich diesbezüglich spezifizieren. So wurde beispielsweise in Abschnitt 8.1 kurz skizziert, inwiefern die Integration von Informationen auch sequentiell ablaufen kann, wodurch sich auch Ankereffekte ergeben können. Dennoch steht eine genaue Spezifikation der Vorhersagen des Modells bezüglich dieser Phänomene zu diesem Zeitpunkt noch aus.

Eine weitere Limitation des GSM ist, dass es nach derzeitigem Stand nicht erklären kann, inwiefern extreme Urteilsfehler bei einzelnen Versuchspersonen entstehen, die sich wesentlich häufiger zeigen als sich bei einer zufälligen Ziehung gleichwertiger Informationen erwarten ließe. So gibt es beispielsweise vermutlich nicht wenige Menschen, die verschiedene Städtedistanzen generell massiv überschätzen. Vermutlich überschätzen sie diese auch, wenn sie hierfür unterschiedliche Überlegungen und Informationen heranziehen. Ihre Schätzungen sind schlicht falsch skaliert. Dadurch können sie gegebenenfalls gut einschätzen, ob die Distanz zwischen London und Berlin kürzer ist als die Distanz zwischen

London und Rom (sogenanntes *Mapping Knowledge*), wobei beide Distanzen massiv überschätzt werden (ein Mangel an *Metric Knowledge*; Brown & Siegler, 1993). Solche Urteilsfehler haben sich in dieser Arbeit vermutlich insbesondere bei der Korrektur von Urteilen als problematisch herausgestellt. Bei den Korrekturen nahmen Menschen tendenziell eher zu große Schritte vor, selbst wenn sie zu „kleinen Schritten" instruiert wurden. Wer davon ausgeht, dass die Distanz von London und Berlin 10,000 km beträgt, für die oder den wären vermutlich eine Korrektur von 500 km nur ein kleiner Schritt, obwohl die wahre Distanz bei 935 km liegt. Eine wichtige Erweiterung des GSM wäre die Integration solcher Skalierungsfehler. Es könnten beispielsweise zusätzliche Verzerrungen auf Ebene des privaten Informationspools eingebaut werden (Ebene 3), die nicht allein durch die zufällige Ziehung gleichwertiger Informationen aus dem geteilten Informationspool entstehen.

### 8.4.1. Limitationen des Weisheitsbewusstseins

Ein wesentlicher Beitrag dieser Arbeit liegt darin, dass gezeigt werden konnte, dass Menschen in der Lage sind, die Richtung ihrer Fehler überzufällig häufig richtig einzuschätzen. Dieses „Weisheitsbewusstsein" zeigte sich über insgesamt 158 unterschiedliche Urteilsfragen (Stimuli) aus verschiedenen Urteilsdomänen, was für die Generalisierbarkeit des Phänomens spricht. Dennoch ergeben sich für dieses Phänomen gewisse Limitationen, da auch in dieser Arbeit nur ein Ausschnitt der in der Realität möglichen Arten von Aufgaben untersucht werden konnte. Eine besonders wichtige und bisher nicht repräsentierte Gruppe von Aufgaben sind Vorhersagen, bei denen Urteilsgegenstände beurteilen werden, die in der Zukunft liegen. Solche Vorhersagen sind sehr häufig Gegenstand wichtiger politischer und wirtschaftlicher Entscheidungen, stellen jedoch größere Herausforderungen an kontrollierte Studien, bei denen die Urteile von Personen meist unmittelbar bewertet werden müssen. Eine Möglichkeit für zukünftige Untersuchungen des Weisheitsbewusstseins sind jedoch simulierte Prognoseaufgaben wie simulierte Aktienkurse. Eine andere Möglichkeit sind Feldstudien, die zu einem späteren Zeitpunkt evaluiert werden, jedoch einen wesentlich höheren Erhebungsaufwand mit sich bringen (z.B., Ben-David et al., 2013). Nach dem GSM sollten sich die Befunde grundsätzlich auch auf Vorhersageaufgaben generalisieren lassen. Eine wichtige Voraussetzung ist hierbei lediglich, dass Menschen auch bei diesen Aufgaben mehr (unabhängige) Informationen zur Verfügung haben als sie in einem ersten Urteil verarbeiten können. Dennoch müsste das

Weisheitsbewusstsein direkt anhand dieser Aufgaben untersucht werden, um Fragen nach der Generalisierbarkeit zu beantworten.

Weitere Limitationen betreffen die Generalisierbarkeit der Befunde zum Weisheitsbewusstsein auf neue Populationen von Versuchspersonen. In dieser Arbeit konnte das Weisheitsbewusstsein bereits bei zwei unterschiedlichen Populationen von Versuchspersonen in unterschiedlichen Sprachräumen gezeigt werden, wobei es sich in beiden Fällen um sogenannte *Convenience Samples* handelte und keine repräsentativen Bevölkerungsgruppen. Zudem werden meist wichtige Urteile von Expert:innen in den jeweiligen Domänen getroffen, womit die Frage einhergeht, ob sich die Befunde auch auf diese Population generalisieren lässt. So sind Expert:innen beispielsweise in der Lage, ihre Arbeitsgedächtniskapazitäten effizienter zu nutzen, da sie Informationen zu größeren Informationspaketen zusammenfassen können (Gobet et al., 2001). Expert:innen sollten somit auch häufiger in der Lage sein, alle verfügbaren Informationen bereits in einem ersten Urteil zu verarbeiten, womit sich später kein Weisheitsbewusstsein ausbilden könnte. Auf der anderen Seite sollten Expert:innen mehr relevante Informationen zur Verfügung haben, weshalb auch hier grundsätzlich mehr Informationen verfügbar sein könnten als diese verarbeiten können – das wiederum würde auch bei dieser Personengruppe die Ausbildung eines Weisheitsbewusstseins ermöglichen.

Zumindest teilweise offen bleibt auch die Frage, inwiefern Menschen spontan ein Weisheitsbewusstsein ausbilden, wenn sie nicht explizit danach gefragt werden, oder ob hierfür bestimmte Instruktionen notwendig sind. In dieser Arbeit wurden zur Abfrage des Weisheitsbewusstseins Instruktionen verwendet, die dem Dialectical Bootstrapping ähneln (Herzog & Hertwig, 2009, 2014b). Im Rahmen dieser Instruktionen sollten Menschen zuerst annehmen, dass sie mit ihrem ersten Urteil den wahren Wert verfehlten und im nächsten Schritt Gründe abrufen, die eher für die eine oder andere Richtung sprechen, und zuletzt die Richtung des wahren Werts einschätzen. Alle diese Schritte könnten theoretisch notwendig sein, damit Menschen ein Weisheitsbewusstsein ausbilden. Weiteren Aufschluss über diese Fragen nach den Voraussetzungen könnten Untersuchungen liefern, bei denen die Instruktionen zur Abfrage des Weisheitsbewusstseins manipuliert werden, beziehungsweise schrittweise reduziert werden (in einem sogenannten *Dismantling-Design*). Ich gehe jedoch davon aus, dass diese Schritte nicht notwendig sind, auch wenn sie dafür sorgen könnten,

dass sich dadurch häufiger ein Weisheitsbewusstsein ausbildet. So zeigte sich in den Kontrollgruppen der Interventionsstudien, dass Menschen durchaus häufig (jedoch etwas seltener als in den Experimentalbedingungen) ihre Urteile in eine bestimmte Richtung korrigierten und in diesen Fällen überzufällig häufig die richtige Richtung wählten. Hierbei könnten theoretisch andere Prozesse ablaufen als bei der Richtungsabfrage (vgl. Gaertig & Simmon, 2021), wobei es zumindest die sparsamere Erklärung ist, hierfür die gleichen Prozesse anzunehmen. In diesem Zusammenhang stellt sich auch die Frage, wie unmittelbar Menschen ein Weisheitsbewusstsein ausbilden. Die zeitlichen Abstände zwischen erstem Urteil und Richtungsurteil waren in den Studien dieser Arbeit vergleichsweise kurz, wobei sich durchaus schwächere Effekte ergeben könnten, wenn Menschen direkt nach dem ersten Urteil ein Richtungsurteil treffen. Zumindest in den Arbeiten zur Weisheit der Vielen in einer Person zeigte sich, dass die Unabhängigkeit zweier Urteile teilweise durch einen größeren zeitlichen Abstand gesteigert werden kann (Vul & Pashler, 2008; siehe aber auch Steegen et al., 2014). Ein wesentlicher Faktor bei der spontanen Bewertung eigener Urteile könnte zudem sein, wie sicher sich Menschen bei ihren jeweiligen Urteilen sind. Bei größerer Unsicherheit könnten Menschen ihre eigenen Urteile wesentlich häufiger spontan kritisch bewerten und gegebenenfalls ändern. So zeigt sich beispielsweise auch, dass eine hohe Urteilssicherheit dazu führt, dass Menschen ihre Urteile eher nicht ändern, wenn sie Ratschläge von anderen Personen erhalten (Rader et al., 2017). Gleiches sollte für die spontane Bewertung und Korrektur von Urteilen gelten.

Zum Schluss ergeben sich Limitationen bezüglich der Prozesse, die im Rahmen des GSM bezüglich des Weisheitsbewusstseins und auch der Korrekturinterventionen angenommen wurden. In dieser Arbeit wurde das Weisheitsbewusstsein bisher nur gemessen und bestimmte Prozesse mit Hilfe von Interventionen verstärkt, wobei eine direkte Manipulation der angenommenen Prozesse hinsichtlich des Abrufs neuer Informationen noch aussteht. Im Klartext heißt das, dass die bisherigen Ergebnisse konsistent mit dem GSM sind, aber bisher noch keine Evidenz für das Modell generiert werden konnte, welche das GSM von anderen Alternativerklärungen abgrenzt. Eine konkrete Alternativerklärung für das Weisheitsbewusstsein wäre zum Beispiel, dass Menschen bei ihren Richtungsurteilen lediglich die Richtung wählen, in der, unabhängig von dem Inhalt der konkreten Frage, mehr (plausible) Antwortmöglichkeiten liegen (siehe auch Gaertig &

Simmons, 2021). Die Anzahl der Möglichkeiten hängt hierbei von der Begrenzung der Antwortskala und der relativen Position des Initialurteils ab. Schätzt beispielsweise eine Person, dass 20% der Bevölkerung Eigenschaft X aufweist, so liegen grundsätzlich mehr Antwortmöglichkeiten über 20% als unter 20%, da die Schätzung solcher Proportionen bei 0% und 100% begrenzt sind. Diese Erklärung kann auch auf Aufgaben ausgeweitet werden, die nur nach oben oder unten begrenzt sind. Werden beispielsweise die Kosten eines Gegenstandes Y geschätzt (bei 0€ begrenzt und nach oben theoretisch unbegrenzt), so gibt es theoretisch mehr Möglichkeiten, nach denen Y teurer sein kann als ein ursprünglicher Referenzwert, im Vergleich zu Möglichkeiten, nach denen er günstiger wäre als dieser Referenzwert. Es ist jedoch auch möglich und plausibel, dass Menschen andere psychologische Grenzwerte ansetzen und mit ihren Urteilen zum Mittelpunkt dieser Grenzwerte tendieren. Wird beispielsweise der Preis eines gebrauchten Smartphones auf 700€ geschätzt, so liegen theoretisch mehr Antwortmöglichkeiten über 700€, aber mehr *plausible* Möglichkeiten unter 700€. Anhand der erhobenen Daten lässt sich jedoch nicht zwischen dieser Erklärung und der des GSM unterscheiden, da der psychologische Mittelpunkt einer Skala vermutlich nah an dem Populationsmittelwert aller Urteile liegen sollte, sofern er sich überhaupt post hoc definieren lässt. Es lässt sich jedoch bereits sagen, dass sich zumindest das Gesamtbild der Befunde mit dieser Alternative nicht erklären lässt. So kann dieser Ansatz beispielsweise nicht erklären, warum Menschen eher in der Lage sind, die Richtung des wahren Werts relativ zu einem fremden Urteil einzuschätzen als zu einem eigenen Urteil. Direkt ließe sich diese Alternative im Rahmen einer Studie testen, bei der manipuliert wird, ob Menschen bei der Abfrage des Weisheitsbewusstseins erneut Einsicht in ihre privaten Informationen erhalten oder nur wissen, im Rahmen welcher Domäne und auf welcher Skala das Urteil getroffen wurde (siehe auch Soll & Mannes, 2011). Beispielsweise könnten die Aufgaben mit den Kalorienschätzungen verwendet werden. In diesem Fall würde den Versuchspersonen zur Bewertung ihrer Urteile in einer Bedingung erneut die Bilder mit dem jeweiligen Gericht präsentiert werden und in einer anderen nicht. Nach dem GSM sollten Menschen primär dann ein Weisheitsbewusstsein ausbilden, wenn sie Zugriff auf ihre ursprünglichen Informationen erhalten (wenn sie die Bilder sehen), während sich nach der Alternativerklärung keine Bedingungsunterscheide zeigen sollten.

### 8.5. Fazit

Menschen sind bei der Bewertung ihrer eigenen quantitativen Urteile in der Lage zu erkennen, in welche Richtung sie ihre Urteile korrigieren sollten. Dieses Phänomen wurde in dieser Arbeit als Weisheitsbewusstsein bezeichnet und empirisch nachgewiesen. Nach derzeitigem Stand können Menschen jedoch noch nicht aufgabenübergreifend in die Lage versetzt werden, dieses Potential selbstständig zur Korrektur ihrer eigenen Urteile oder zur besseren Gewichtung von Ratschlägen zu nutzen. Insofern kann diesbezüglich noch keine Praxisempfehlung ausgesprochen werden. Es konnte jedoch eine Strategie entwickelt werden, bei der Menschen kleine Korrekturen im Einklang mit ihrem Weisheitsbewusstsein vornehmen, für die sich bei bestimmten Aufgabenkategorien (historische Schätzungen) erste positive Evidenz zeigte. Darüber hinaus konnten mit dem Weisheitsbewusstsein zentrale Annahmen eines allgemeinen Sampling-Modells gestützt werden, nach welchem Menschen Verbesserungspotentiale in ihren eigenen Urteilen erkennen können, indem sie bei den entsprechenden Bewertungsprozessen teilweise neue Informationen aus einem geteilten Informationspool ziehen, der der Gesamtpopulation zur Verfügung steht. Neben den hier untersuchten Implikationen kann dieses Modell zu einem umfassenden Verständnis (meta-)kognitiver Prozesse und deren Limitationen bei der Urteilsbildung und gegebenenfalls auch bei der Gewichtung von Ratschlägen beitragen.

## 9. Literaturverzeichnis


Allaire, J. J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., Wickham, H., Cheng, J., Chang, W., & Iannone, R. (2021). *rmarkdown: Dynamic documents for R* (Version 2.9) [Computer software]. https://github.com/rstudio/rmarkdown

Aust, F. (2019). *Prereg: R markdown templates to preregister scientific studies* (Version 0.4.0) [Computer software]. https://CRAN.R-project.org/package=prereg

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown* (Version 0.1.0.9997) [Computer software]. https://github.com/crsh/papaja

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-David, I., Graham, J. R., & Harvey, C. R. (2013). Managerial miscalibration. *The Quarterly Journal of Economics*, *128*(4), 1547–1584. https://doi.org/10.1093/qje/qjt023

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*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

Bonner, B. L., Sillito, S. D., & Baumann, M. R. (2007). Collective estimation: Accuracy, expertise, and extroversion as sources of intra-group influence. *Organizational Behavior and Human Decision Processes*, *103*(1), 121–133. https://doi.org/10.1016/j.obhdp.2006.05.001

Brown, N. R., & Siegler, R. S. (1993). Metrics and mappings: A framework for understanding real-world quantitative estimation. *Psychological Review*, *100*(3), 511–534. https://doi.org/10.1037/0033-295x.100.3.511

Champely, S. (2018). *Pwr: Basic functions for power analysis*. (Version 1.2.2) [Computer software]. https://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd Edition). Routledge. https://doi.org/10.4324/9780203771587

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*(1), 87–185. https://doi.org/10.1017/s0140525x01003922

*Das ultimative Studenten-Kochbuch* (2018). Naumann & Göbel.

Deutsch, M., & Gerard, H. B. (1955). A study of normative and informational social influences upon individual judgment. *Journal of Abnormal and Social Psychology*, *51*(3), 629–636. https://doi.org/10.1037/h0046408

*30-Minuten-Rezepte* (2017). Naumann & Göbel. https://www.naumann-goebel.de/produkt-details/30-minuten-rezepte-9783815569078/show/

Edwards, W. (1982). Conservatism in human information processing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 359–369). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.026

*1000 Rezepte für zwei* (2014). Naumann & Göbel. https://www.naumann-goebel.de/produkt-details/1000-rezepte-fuer-zwei-9783815584903/show/

*1000 vegetarische Gerichte* (2014). Naumann & Göbel. https://www.naumann-goebel.de/produkt-details/1000-vegetarische-gerichte-9783815584897/show/

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science*, *17*(4), 311–318. https://doi.org/10.1111/j.1467-9280.2006.01704.x

Fiedler, K. (2000). Beware of samples! A cognitive-ecological sampling approach to judgment biases. *Psychological Review*, *107*(4), 659–676. https://doi.org/10.1037/0033-295x.107.4.659

Fiedler, K., Hütter, M., Schott, M., & Kutzner, F. (2019). Metacognitive myopia and the overutilization of misleading advice. *Journal of Behavioral Decision Making*, *32*(3), 317–333. https://doi.org/10.1002/bdm.2109

Fiedler, K., & Juslin, P. (2006a). *Information sampling and adaptive cognition*. Cambridge University Press. https://doi.org/10.1017/CBO9780511614576

Fiedler, K., & Juslin, P. (2006b). Taking the interface between mind and environment seriously. In K. Fiedler & P. Juslin (Eds.), *Information sampling and adaptive cognition*

(pp. 3–29). Cambridge University Press. https://doi.org/10.1017/CBO9780511614576.001

Fiedler, K., & Krueger, J. I. (2012). More than an artifact: Regression as a theoretical construct. In J. I. Krueger (Ed.), *Social judgment and decision making* (pp. 171–189). Psychology Press. https://psycnet.apa.org/record/2011-26824-010

Fischer, I., & Harvey, N. (1999). Combining forecasts: What information do judges need to outperform the simple average? *International Journal of Forecasting*, *15*(3), 227–246. https://doi.org/10.1016/S0169-2070(98)00073-9

Fischhoff, B. (1982). Debiasing. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 422–444). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.032

Fraundorf, S. H., & Benjamin, A. S. (2014). Knowing the crowd within: Metacognitive limits on combining multiple judgments. *Journal of Memory and Language*, *71*(1), 17–38. https://doi.org/10.1016/j.jml.2013.10.002

Gaertig, C., & Simmons, J. P. (2021). The psychology of second guesses: Implications for the wisdom of the inner crowd. *Management Science*, *67*(9), 5921–5942. https://doi.org/10.1287/mnsc.2020.3781

Georg, S. (2021, October 26). In *Luftlinie.org*. https://www.luftlinie.org/

Gigerenzer, G., Hoffrage, U., & Kleinbölting, H. (1991). Probabilistic mental models: A Brunswikian theory of confidence. *Psychological Review*, *98*(4), 506–528. https://doi.org/10.1037/0033-295X.98.4.506

Gobet, F., Lane, P. C. R., Croker, S., Cheng, P. C.-H., Jones, G., Oliver, I., & Pine, J. M. (2001). Chunking mechanisms in human learning. *Trends in Cognitive Sciences*, *5*(6), 236–243. https://doi.org/10.1016/s1364-6613(00)01662-4

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114–125. https://doi.org/10.1007/s40881-015-0004-4

Harvey, N., & Fischer, I. (1997). Taking advice: Accepting help, improving judgment, and sharing responsibility. *Organizational Behavior and Human Decision Processes*, *70*(2), 117–133. https://doi.org/10.1006/obhd.1997.2697

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual

judgments with dialectical bootstrapping. *Psychological Science*, *20*(2), 231–237. https://doi.org/10.1111/j.1467-9280.2009.02271.x

Herzog, S. M., & Hertwig, R. (2013). The crowd within and the benefits of dialectical bootstrapping: A reply to White and Antonakis (2013). *Psychological Science*, *24*(1), 117–119. https://doi.org/10.1177/0956797612457399

Herzog, S. M., & Hertwig, R. (2014a). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*(10), 504–506. https://doi.org/10.1016/j.tics.2014.06.009

Herzog, S. M., & Hertwig, R. (2014b). Think twice and then: Combining or choosing in dialectical bootstrapping? *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *40*(1), 218–232. https://doi.org/10.1037/a0034054

Herzog, S. M., Litvinova, A., Yahosseini, K. S., Tump, A. N., & Kurvers, R. H. J. (2019). The ecological rationality of the wisdom of crowds. In R. Hertwig, T. J. Pleskac, & T. Pachur (Eds.), *Taming Uncertainty* (pp. 245–262). The MIT Press. https://doi.org/10.7551/mitpress/11114.003.0019

Hourihan, K. L., & Benjamin, A. S. (2010). Smaller is better (when sampling from the crowd within): Low memory-span individuals benefit more from multiple opportunities for estimation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*(4), 1068–1074. https://doi.org/10.1037/a0019694

Hütter, M., & Ache, F. (2016). Seeking advice: A sampling approach to advice taking. *Judgment and Decision Making*, *11*(4), 401–415. http://journal.sjdm.org/15/151110a/jdm151110a.pdf

Johnson, P. (2020). *devEMF: EMF graphics output device* (Version 4.0-2) [Computer software]. https://CRAN.R-project.org/package=devEMF

Juslin, P., Olsson, H., & Olsson, A.-C. (2003). Exemplar effects in categorization and multiple-cue judgment. *Journal of Experimental Psychology. General*, *132*(1), 133–156. https://doi.org/10.1037/0096-3445.132.1.133

Juslin, P., Winman, A., & Hansson, P. (2007). The naïve intuitive statistician: A naïve sampling model of intuitive confidence intervals. *Psychological Review*, *114*(3), 678–703. https://doi.org/10.1037/0033-295X.114.3.678

Klayman, J., Soll, J. B., Juslin, P., & Winman, A. (2006). Subjective confidence and the sampling of knowledge. In K. Fiedler & P. Juslin (Eds.), *Information sampling and*

*adaptive cognition* (pp. 153–182). Cambridge University Press. https://doi.org/10.1017/CBO9780511614576.008

Krueger, J. I., & Chen, L. J. (2014). The first cut is the deepest: Effects of social projection and dialectical bootstrapping on judgmental accuracy. *Social Cognition*, *32*(4), 315–336. https://doi.org/10.1521/soco.2014.32.4.315

Koriat, A. (2012a). The self-consistency model of subjective confidence. *Psychological Review*, *119*(1), 80–113. https://doi.org/10.1037/a0025648

Koriat, A. (2012b). When are two heads better than one and why? *Science*, *336*(6079), 360–362. https://doi.org/10.1126/science.1216549

Koriat, A. (2019). Confidence judgments: The monitoring of object-level and same-level performance. *Metacognition and Learning*, *14*(3), 463–478. https://doi.org/10.1007/s11409-019-09195-7

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). LmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Larrick, R. P., Mannes, A. E., & Soll, J. B. (2012). The social psychology of the wisdom of crowds. In J. I. Krueger (Ed.), *Frontiers of social psychology: Social judgment and decision making* (pp. 227–242). Psychology Press. https://psycnet.apa.org/record/2011-26824-013

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459

Laughlin, P. R., Bonner, B. L., Miner, A. G., & Carnevale, P. J. (1999). Frames of reference in quantity estimations by groups and individuals. *Organizational Behavior and Human Decision Processes*, *80*(2), 103–117. https://doi.org/10.1006/obhd.1999.2848

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

Liste der größten Städte in Bayern (2021, October 26). In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Liste_der_größten_Städte_in_Bayern&oldid=216053914

Liste der größten Städte in Niedersachsen (2021, October 26). In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Liste_der_gr%C3%B6%C3%9Ften_St%C3%A4dte_in_Niedersachsen&oldid=216053868

Liste der größten Städte in Schleswig-Holstein (2021, October 26). In *Wikipedia*. https://de.wikipedia.org/w/index.php?title=Liste_der_größten_Städte_in_Schleswig-Holstein&oldid=216054577

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime.com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior Research Methods*, *49*(2), 433–442. https://doi.org/10.3758/s13428-016-0727-z

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Milkman, K. L., Chugh, D., & Bazerman, M. H. (2009). How can decision making be improved? *Perspectives on Psychological Science*, *4*(4), 379–383. https://doi.org/10.1111/j.1745-6924.2009.01142.x

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Minear, M., & Park, D. C. (2004). A lifespan database of adult facial stimuli. *Behavior Research Methods, Instruments, & Computers*, *36*(4), 630–633. https://doi.org/10.3758/BF03206543

Minson, J. A., Liberman, V., & Ross, L. (2011). Two to tango: Effects of collaboration and disagreement on dyadic judgment. *Personality & Social Psychology Bulletin*, *37*(10), 1325–1338. https://doi.org/10.1177/0146167211410436

Moore, D. A., Tenney, E. R., & Haran, U. (2015). Overprecision in judgment. In G. Keren & G. Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making II* (pp. 182–209). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118468333.ch6

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making*, *6*(4), 283–294. http://journal.sjdm.org/11/101217a/jdm101217a.pdf

Mussweiler, T., & Strack, F. (1999). Comparing is believing: A selective accessibility model of judgmental anchoring. *European Review of Social Psychology*, *10*(1), 135–167. https://doi.org/10.1080/14792779943000044

Mussweiler, T., Strack, F., & Pfeiffer, T. (2000). Overcoming the inevitable anchoring effect: Considering the opposite compensates for selective accessibility. *Personality & Social Psychology Bulletin*, *26*(9), 1142–1150. https://doi.org/10.1177/01461672002611010

Navarro, D. (2015). *Learning statistics with R: A tutorial for psychology students and other beginners* (Version 0.5) [Computer software]. http://ua.edu.au/ccs/teaching/lsr

Nichols, A. L., & Maner, J. K. (2008). The good-subject effect: Investigating participant demand characteristics. *The Journal of General Psychology*, *135*(2), 151–165. https://doi.org/10.3200/GENP.135.2.151-166

Pachur, T., & Bröder, A. (2013). Judgment: A cognitive processing perspective. *WIREs Cognitive Science*, *4*(6), 665–681. https://doi.org/10.1002/wcs.1259

Peterson, C. R., & Beach, L. R. (1967). Man as an intuitive statistician. *Psychological Bulletin*, *68*(1), 29–46. https://doi.org/10.1037/h0024722

Phillips, N. (2017). *yarrr: A Companion to the e-Book "YaRrr!: The Pirate's Guide to R"* (Version 0.1.5) [Computer software]. https://CRAN.R-project.org/package=yarrr

Rader, C. A., Larrick, R. P., & Soll, J. B. (2017). Advice as a form of social influence: Informational motives and the consequences for accuracy. *Social and Personality Psychology Compass*, *11*(8). e12329. https://doi.org/10.1111/spc3.12329

Rader, C. A., Soll, J. B., & Larrick, R. P. (2015). Pushing away from representative advice: Advice taking, anchoring, and adjustment. *Organizational Behavior and Human Decision Processes*, *130*, 26–43. https://doi.org/10.1016/j.obhdp.2015.05.004

R Core Team. (2019). *R: A Language and environment for statistical computing* (Version 3.6.2) [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/

RStudio Team (2019). *RStudio: Integrated development for R* (Version 1.2.5033) [Computer software]. RStudio, Inc. http://www.rstudio.com/

*Schnelle Gerichte - Die beliebtesten Rezepte* (2015). Naumann & Göbel. https://www.naumann-goebel.de/produkt-details/schnelle-gerichte-die-beliebtesten-rezepte-9783815588710/show/

Schultze, T., Gerlach, T. M., & Rittich, J. C. (2018). Some people heed advice less than others: Agency (but not communion) predicts advice taking. *Journal of Behavioral Decision Making*, *31*(3), 430–445. https://doi.org/10.1002/bdm.2065

Schultze, T., Mojzisch, A., & Schulz-Hardt, S. (2017). On the inability to ignore useless advice: A case for anchoring in the judge-advisor-system. *Experimental Psychology*, *64*(3), 170–183. https://doi.org/10.1027/1618-3169/a000361

Schultze, T., Rakotoarisoa, A.-F., & Schulz-Hardt, S. (2015). Effects of distance between initial estimates and advice on advice utilization. *Judgment and Decision Making*, *10*(2), 144–171. http://journal.sjdm.org/14/141112a/jdm141112a.pdf

Schulz-Hardt, S., Brodbeck, F. C., Mojzisch, A., Kerschreiter, R., & Frey, D. (2006). Group decision making in hidden profile situations: Dissent as a facilitator for decision quality. *Journal of Personality and Social Psychology*, *91*(6), 1080–1093. https://doi.org/10.1037/0022-3514.91.6.1080

Sniezek, J. A., & Buckley, T. (1995). Cueing and cognitive conflict in judge-advisor decision making. *Organizational Behavior and Human Decision Processes*, *62*(2), 159–174. https://doi.org/10.1006/obhd.1995.1040

Sniezek, J. A., Schrah, G. E., & Dalal, R. S. (2004). Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making*, *17*(3), 173–190. https://doi.org/10.1002/bdm.468

Soll, J. B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(2), 299–314. https://doi.org/10.1037/0278-7393.30.2.299

Soll, J. B., & Larrick, R. P. (2009). Strategies for revising judgment: How (and how well) people use others' opinions. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *35*(3), 780–805. https://doi.org/10.1037/a0015145

Soll, J. B., & Mannes, A. E. (2011). Judgmental aggregation strategies depend on whether the self is involved. *International Journal of Forecasting*, *27*(1), 81–102. https://doi.org/10.1016/j.ijforecast.2010.05.003

Soll, J. B., Milkman, K. L., & Payne, J. W. (2015). A User's Guide to Debiasing. In G. Keren & G.

Wu (Eds.), *The Wiley Blackwell Handbook of Judgment and Decision Making II* (pp. 924–951). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118468333.ch33

Soll, J. B., Palley, A. B., & Rader, C. A. (2021). The bad thing about good advice: Understanding when and how advice exacerbates overconfidence. *Management Science*. Advance online publication. https://doi.org/10.1287/mnsc.2021.3987

*Soul Food - Veggie-Blitzrezepte* (2015). Naumann & Göbel. https://www.naumann-goebel.de/produkt-details/soul-food-veggie-blitzrezepte-9783815588499/show/

Stasser, G., & Dietz-Uhler, B. (2001). Collective choice, judgment, and problem solving. In M. A. Hogg & R. S. Tindale (Eds.), *Blackwell Handbook of Social Psychology: Group Processes* (pp. 31–55). Blackwell Publishers Ltd. https://doi.org/10.1002/9780470998458.ch2

Steegen, S., Dewitte, L., Tuerlinckx, F., & Vanpaemel, W. (2014). Measuring the crowd within again: A pre-registered replication study. *Frontiers in Psychology*, *5*, 1–8. https://doi.org/10.3389/fpsyg.2014.00786

Stern, A. (2017). *Sozial vermittelte Lernprozesse bei quantitativen Schätzaufgaben* [Doctoral dissertation, University of Goettingen]. eDiss. http://hdl.handle.net/11858/00-1735-0000-0023-3E3D-9

Stern, A., Schultze, T., & Schulz-Hardt, S. (2017). How much group is necessary? Group-to-individual transfer in estimation tasks. *Collabra. Psychology*, *3*(1), 1–17. https://doi.org/10.1525/collabra.95

Stewart, T. R. (2001). Improving reliability of judgmental forecasts. In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 81–106). Springer. https://doi.org/10.1007/978-0-306-47630-3_5

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*(5), 550–562. https://doi.org/10.1037/h0070482

Surowiecki, J. (2004). *The wisdom of crowds: Why the many are smarter than the few and how collective wisdom shapes business, economies, societies and nations*. Doubleday.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L., & Imai, K. (2014). mediation: R Package for causal mediation analysis. *Journal of Statistical Software, 59*(5), 1–38. https://doi.org/10.18637/jss.v059.i05

Treffenstädt, C., & Wiemann, P. (2018). *Alfred–A library for rapid experiment development*

(Version 0.2b5) [Computer software]. https://doi.org/10.5281/zenodo.1437220

Vul, E., & Pashler, H. (2008). Measuring the crowd within: Probabilistic representations within individuals. *Psychological Science*, *19*(7), 645–647. https://doi.org/10.1111/j.1467-9280.2008.02136.x

Welsh, M. B., & Begg, S. H. (2018). More-or-less elicitation (MOLE): Reducing bias in range estimation and forecasting. *EURO Journal on Decision Processes*, *6*(1-2), 171–212. https://doi.org/10.1007/s40070-018-0084-5

White, C. M., & Antonakis, J. (2013). Quantifying accuracy improvement in sets of pooled judgments: Does dialectical bootstrapping work? *Psychological Science*, *24*(1), 115–116. https://doi.org/10.1177/0956797612449174

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., … Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*. *4*(43), 1–6. https://doi.org/10.21105/joss.01686

Wilson, T. D., & Brekke, N. (1994). Mental contamination and mental correction: Unwanted influences on judgments and evaluations. *Psychological Bulletin*, *116*(1), 117–142. https://doi.org/10.1037/0033-2909.116.1.117

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *30*(6), 1167–1175. https://doi.org/10.1037/0278-7393.30.6.1167

Xie, Y. (2021). *knitr: A general-purpose package for dynamic report generation in R* (Version 1.33) [Computer software]. https://yihui.org/knitr/

Yaniv, I. (1997). Weighting and trimming: Heuristics for aggregating judgments under uncertainty. *Organizational Behavior and Human Decision Processes*, *69*(3), 237–249. https://doi.org/10.1006/obhd.1997.2685

Yaniv, I. (2004a). The benefit of additional opinions. *Current Directions in Psychological Science*, *13*(2), 75–78. https://doi.org/10.1111/j.0963-7214.2004.00278.x

Yaniv, I. (2004b). Receiving other people's advice: Influence and benefit. *Organizational Behavior and Human Decision Processes*, *93*(1), 1–13. https://doi.org/10.1016/j.obhdp.2003.08.002

Yaniv, I., & Choshen-Hillel, S. (2012a). Exploiting the wisdom of others to make better decisions: Suspending judgment reduces egocentrism and increases accuracy. *Journal of Behavioral Decision Making*, *25*(5), 427–434. https://doi.org/10.1002/bdm.740

Yaniv, I., & Choshen-Hillel, S. (2012b). When guessing what another person would say is better than giving your own opinion: Using perspective-taking to improve advice-taking. *Journal of Experimental Social Psychology*, *48*(5), 1022–1028. https://doi.org/10.1016/j.jesp.2012.03.016

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy of judgmental estimation. *Journal of Behavioral Decision Making*, *10*(1), 21–32. https://doi.org/10.1002/(SICI)1099-0771(199703)10:1<21::AID-BDM243>3.0.CO;2-G

Yaniv, I., I., & Kleinberger, E. (2000). Advice taking in decision making: Egocentric discounting and reputation formation. *Organizational Behavior and Human Decision Processes*, *83*(2), 260–281. https://doi.org/10.1006/obhd.2000.2909

## 10. Anhang

Anhang A: Übersicht über alle Stimuli aus Studie 1, 2, 5, 6 und 7

Anhang B: Übersicht zum Weisheitsbewusstsein mit Ausschluss extremer Fehler

Anhang C: Digitaler Anhang

**Anhang A: Übersicht über alle Stimuli aus Studie 1, 2, 5, 6 und 7**

**Tabelle Anhang A**

*Übersicht über alle Stimuli aus Studie 1, 2, 5, 6 & 7*

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 1, 2 | Wie viel Prozent aller Touristen weltweit stammten 2014 aus Europa? | 51% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52511/tourismus | 18.10.18 |
| 1, 2 | Wie viel Prozent der 10 Mio. am häufigsten genutzten Internetseiten sind auf Englisch (Stand 2015)? | 55% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52515/weltsprache | 18.10.18 |
| 1, 2 | Wie viel Prozent des grenzüberschreitenden Luftfrachtaufkommens wurden 2015 zwischen Europa und Nordamerika ausgetauscht? | 10% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52528/luftfracht | 18.10.18 |
| 1, 2 | Wie viel Prozent des weltweiten Primärenergie-Verbrauchs wurde 2014 mit Öl und Gas gedeckt? | 56% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/239756/themengrafik-handelsstroeme-erdoel-und-erdgas | 18.10.18 |
| 1, 2 | Wie viel Prozent des weltweit verbrauchten Öls werden grenzüberschreitend transportiert (Stand 2014)? | 66% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/239756/themengrafik-handelsstroeme-erdoel-und-erdgas | 18.10.18 |
| 1, 2 | Wie viel Prozent der Weltbevölkerung waren 2012 von Armut betroffen (Kaufkraft von weniger als 3,10 US-Dollar pro Tag)? | 35% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52680/armut | 18.10.18 |
| 1, 2 | Wie viel Prozent der Weltbevölkerung lebte 2015 in einer Stadt? | 54% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52720/megacitys | 18.10.18 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 1, 2 | Wie viel Prozent der Weltbevölkerung hatte 2015 keinen Zugang zu Trinkwasser? | 9% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52696/trinkwasser-und-sanitaereinrichtungen | 18.10.18 |
| 1, 2 | Wie viel Prozent der Weltbevölkerung hatte 2015 keinen Zugang zu Sanitäreinrichtungen bzw. geregelter Abwasserentsorgung? | 32% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52696/trinkwasser-und-sanitaereinrichtungen | 18.10.18 |
| 1, 2 | Wie viel Prozent der Weltbevölkerung machen die Europäer aus (Stand 2018)? | 10% | https://de.statista.com/statistik/daten/studie/1738/umfrage/verteilung-der-weltbevoelkerung-nach-kontinenten/ | 18.10.18 |
| 1, 2, 5, 6 | Wann wurde das erste McDonald's Restaurant in Europa eröffnet? | 1971 | https://www.mcdonalds.de/uber-uns/geschichte | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr gründete sich die globalisierungskritische Bewegung Attac erstmals? | 1998 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52525/globalisierungskritik | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr wurde die erste H&M-Filiale außerhalb Europas eröffnet? | 2000 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52789/mode | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr wurde China Mitglied der World Trade Organization (WTO)? | 2001 | https://www.wto.org/english/thewto_e/whatis_e/tif_e/org6_e.htm | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr fuhr die erste Dampflokomotive in Deutschland? | 1835 | http://www.dhm.de/lemo/rueckblick/erste-eisenbahnen-in-deutschland.html | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr gelang der erste Nonstopflug über den Atlantik? | 1919 | https://www.welt.de/kultur/history/article13361494/Der-erste-Atlantikueberflieger-hiess-nicht-Lindbergh.html | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr wurde der Suez-Kanal erstmals eröffnet? | 1869 | https://www.wissen.de/lexikon/suezkanal | 18.10.18 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 1, 2, 5, 6 | In welchem Jahr fand der erste Transatlantikflug der Lufthansa von Berlin nach New York statt? | 1938 | https://www.austrianwings.info/2013/08/erster-transatlantikflug-der-lufthansa-vor-75-jahren/ | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr wurde die erste Fluggesellschaft der Welt gegründet? | 1909 | http://www.airships.net/delag-passenger-zeppelins/ | 18.10.18 |
| 1, 2, 5, 6 | In welchem Jahr fuhr die erste U-Bahn der Welt in London? | 1863 | https://www.br.de/fernsehen/ard-alpha/sendungen/schulfernsehen/meilensteine-london-ubahn-100.html | 18.10.18 |
| 1, 2, 5, 6 | Wie lang ist die Strecke der Transsibirischen Eisenbahn? | 9288 km | https://www.planet-wissen.de/technik/verkehr/geschichte_der_eisenbahn/pwietranssibirischeeisenbahn100.html | 18.10.18 |
| 1, 2, 5, 6 | Wie lang ist das längste Containerschiff der Welt (Stand 2017)? | 400 m | https://www.verkehrsrundschau.de/mediathek/galerie/die-zehn-groessten-containerschiffe-der-welt-2017-1984817.html | 18.10.18 |
| 1, 2, 5, 6 | Wie lang ist der Suez-Kanal? | 163 km | https://www.wissen.de/lexikon/suezkanal | 18.10.18 |
| 1, 2, 5, 6 | Wie viel Pfennig kostete 1885 ein fünf-minütiges Ferngespräch mit dem Telefon in Deutschland? | 100 Pfennig | http://www.bayern-online.com/v2261/artikel.cfm/203/Gebuehren-Entwicklung-in-Deutschland.html | 18.10.18 |
| 1, 2, 5, 6 | Wie viele Städte gab es 2015 weltweit mit mehr als einer Million Einwohnern? | 501 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52720/megacitys | 18.10.18 |
| 1, 2, 5, 6 | Wie viele U-Bahn-Stationen gibt es in Shanghai (Stand 2017)? | 364 | https://www.ingenieur.de/technik/fachbereiche/verkehr/die-10-groessten-u-bahnnetze-welt/ | 18.10.18 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 1, 2, 5, 6 | Wie lange dauerte der erste Transatlantikflug der Lufthansa von Berlin nach New York? | 25 Std. | https://www.austrianwings.info/2013/08/erster-transatlantikflug-der-lufthansa-vor-75-jahren/ | 18.10.18 |
| 1, 2, 5, 6 | Wie viele Restaurants betrieb McDonald's Ende 2015 weltweit? | 36525 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52774/fast-food | 18.10.18 |
| 1, 2, 5, 6 | Wie viele Filialen hatte IKEA 2016? | 389 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/256162/wohnkultur | 18.10.18 |
| 1, 2, 5, 6 | Wie viele Filialen betreut H&M in Deutschland (Stand 2017)? | 455 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52789/mode | 18.10.18 |
| 2 | Wie viel Prozent der Gesamtausgaben für Forschung und Entwicklung wurden 2015 von Unternehmen aus den USA geleistet? | 39% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52641/forschung-und-entwicklung-fue | 15.03.19 |
| 2 | Wie viel Prozent der Europäer waren 2015 60 Jahre alt oder älter? | 24% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52811/demografischer-wandel | 15.03.19 |
| 2 | Wie viel Prozent aller Staaten weltweit waren 2016 Demokratien? | 63% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52838/demokratie | 15.03.19 |
| 2 | Wie viel Prozent der 20-65-Jährigen in der EU waren 2017 erwerbstätig? | 72% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/europa/70587/erwerbstaetigkeit | 15.03.19 |
| 2 | Wie viel Prozent des weltweit entnommenen Frischwassers werden jährlich für den Agrarsektor verbraucht (Stand 2016)? | 70% | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52730/wasserverbrauch | 15.03.19 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 2, 5, 6 | Seit welchem Jahr wird Coca Cola auch in Deutschland abgefüllt? | 1929 | https://www.welt.de/wirtschaft/article13354769/Coca-Cola-die-Brause-die-aus-der-Apotheke-kam.html | 15.03.19 |
| 2, 5, 6 | In welchem Jahr fand der erste Export eines VW-Wagens ins Ausland statt? | 1947 | https://www.volkswagenag.com/presence/konzern/documents/history/deutsch/Heft17_DE.pdf | 15.03.19 |
| 2, 5, 6 | In welchem Jahr trat die Genfer Flüchtlingskonvention (Abkommen der UN über die Rechtsstellung der Flüchtlinge) in Kraft? | 1951 | http://www.bpb.de/apuz/229819/65-jahre-genfer-fluechtlingskonvention | 15.03.19 |
| 2, 5, 6 | In welchem Jahr wurde das US-amerikanische Unternehmen Netflix gegründet? | 1997 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52780/fernsehunterhaltung | 15.03.19 |
| 2, 5, 6 | In welchem Jahr fanden die ersten Telefongespräche von Deutschland in die USA statt? | 1928 | https://www.faz.net/aktuell/gesellschaft/geschichte-heisser-draht-75-jahre-telefonverbindung-in-die-usa-190658.html | 15.03.19 |
| 2, 5, 6 | Wie viele Kernkraftwerke waren 2018 weltweit in Betrieb? | 447 | https://www.nuklearforum.ch/de/fakten-und-wissen/kernkraftwerke-der-welt | 15.03.19 |
| 2, 5, 6 | Wie hoch war der durchschnittliche Literpreis für Benzin in Deutschland im Jahr 1970? | 55 Pfennig | http://www.science-at-home.de/wiki/index.php/Entwicklung_der_Benzinpreise_seit_1948 | 15.03.19 |
| 2, 5, 6 | Auf wie vielen Sprachen stand Netflix Anfang 2016 zur Verfügung? | 20 | http://www.bpb.de/nachschlagen/zahlen-und-fakten/globalisierung/52780/fernsehunterhaltung | 15.03.19 |
| 2, 5 | Wie viele Handelshäfen befinden sich in den Niederlanden? | 1244 | https://www.laenderdaten.info/Europa/Niederlande/index.php | 15.03.19 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 2, 5, 6 | Wie viele Flüge starten und landen täglich am Berliner Flughafen Tegel? (Stand 2014) | 486 | https://www.morgenpost.de/flughafen-BER/article130001053/In-Tegel-starten-und-landen-taeglich-486-Maschinen.html | 15.03.19 |
| 6 | Was ist die Luftlinienentfernung zwischen London und Amsterdam? | 356 km | https://www.luftlinie.org/London-Airport,Hillingdon,Greater-London,England,GBR/Amsterdam-Airport-Gemeente-Haarlemmermeer,NLD | 10.01.20 |
| 7 | In welchem Jahr wurde die erste Karaoke-Maschine von dem japanischen Erfinder Daisuke Inoue entworfen? | 1971 | https://letskaraoke.de/geschichte-der-karaoke-maschinen/ | 24.02.20 |
| 7 | In welchem Jahr erschien die erste Ausgabe des amerikanischen Magazins „Rolling Stone“? | 1967 | https://www.britannica.com/topic/Rolling-Stone | 24.02.20 |
| 7 | In welchem Jahr meldete Emil Berliner das Patent für die Schallplatte an? | 1887 | https://www.plattenspieler-guru.de/plattenspieler-geschichte-daten-und-fakten/ | 24.02.20 |
| 7 | In welchem Jahr wurden die Grammy Awards das erste Mal verliehen? | 1958 | https://www.musicianshalloffame.com/history-of-the-grammy-awards/ | 24.02.20 |
| 7 | In welchem Jahr wurde der Song „Mein kleiner grüner Kaktus“ zum ersten Mal aufgenommen? | 1934 | http://www.comedian-harmonists.net/?page_id=408#Kaktus | 24.02.20 |
| 7 | In welchem Jahr wurde das Museum der „Rock and Roll Hall of Fame“ eröffnet? | 1995 | https://www.rockhall.com/about | 24.02.20 |
| 7 | In welchem Jahr meldete George Beauchamp das Patent für die erste elektrische Gitarre an? | 1932 | https://www.gitarrebass.de/equipment/geschichte-der-e-gitarre/ | 24.02.20 |
| 7 | In welchem Jahr wurde der erste Cassetten Recorder von PHILIPS vorgestellt? | 1967 | https://www.lokalkompass.de/luenen/c-kultur/damals-wars-1967-der-erste-stereo-compact-cassetten-recorder-philips-el-3312-nostalgie-news-uwe-h-sueltz-lokalkompass-luenen_a384733 | 24.02.20 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 7 | In welchem Jahr erfand Fritz Pfleumer das erste Tonband? | 1928 | https://blog.landr.com/de/6-geniale-musikalische-erfindungen/ | 24.02.20 |
| 7 | In welchem Jahr brachte Sony den ersten Walkman heraus? | 1979 | https://blog.landr.com/de/6-geniale-musikalische-erfindungen | 24.02.20 |
| 7 | In welchem Jahr wurde die Mundharmonika von Christian Buschmann erfunden? | 1821 | http://www.harmonicajamz.com/harmonicas-facts-fun/ | 24.02.20 |
| 7 | In welchem Jahr meldete Adolphe Sax das Patent für das Saxophon an? | 1846 | https://www.musik-instrumente-info.de/saxophon/ | 24.02.20 |
| 7 | In welchem Jahr wurde das erste Klavier erfunden? | 1726 | https://www.musik-instrumente-info.de/das-klavier/ | 24.02.20 |
| 7 | In welchem Jahr kam das erste Schlagzeug in den Handel? | 1918 | https://www.musik-instrumente-info.de/schlagzeug/ | 24.02.20 |
| 7 | In welchem Jahr wurde der ECHO das erste Mal verliehen? | 1992 | https://echopop-archiv.de/uber-den-echo/ | 24.02.20 |
| 7 | In welchem Jahr wurde die erste Hörfunk-Sendung der Welt ausgestrahlt? | 1906 | https://www.helpster.de/wer-war-der-erfinder-des-radios_227872 | 24.02.20 |
| 7 | In welchem Jahr erfand Edison den ersten Phonographen? | 1877 | https://phonographmania.weebly.com/facts.html | 24.02.20 |
| 7 | In welchem Jahr nahm MTV seinen Sendebetrieb auf? | 1981 | https://www.britannica.com/topic/MTV | 24.02.20 |
| 7 | In welchem Jahr eröffnete in Philadelphia das erste Tonstudio der Welt? | 1897 | https://www.tonmenschen.de/tonstudio/ | 24.02.20 |
| 7 | In welchem Jahr wurden die Ohropax der Öffentlichkeit das erste Mal vorgestellt? | 1907 | https://www.bpb.de/geschichte/zeitgeschichte/sound-des-jahrhunderts/209560/antiphon-und-ohropax | 24.02.20 |
| 7 | In welchem Jahr wurde das Mikrofon erfunden? | 1877 | https://www.mikrofonwelt.de/mikrofon-erfindung/ | 24.02.20 |
| 7 | In welchem Jahr begann erstmals die kommerzielle Vermarktung von Kopfhörern? | 1910 | https://www.connect.de/ratgeber/kopfhoerer-geschichte-historie-1931649.html | 24.02.20 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 7 | In welchem Jahr wurde die erste Langspielplatte (Vinyl) vorgestellt? | 1948 | https://www.planet-wissen.de/kultur/musik/geschichte_der_tontraeger/pwiedieschallplatte100.html | 24.02.20 |
| 7 | In welchem Jahr wurden die ersten Sterne des Hollywood Walk of Fame enthüllt? | 1958 | http://www.walkoffame.com/pages/history | 24.02.20 |
| 7 | In welchem Jahr wurde der erste Automat erfunden, der durch einen Münzeinwurf Musik abspielen konnte? | 1889 | https://www.was-war-wann.de/geschichte/jukebox.html | 24.02.20 |
| 7 | In welchem Jahr erschien in Deutschland die erste Hitparade? | 1953 | https://www.swr.de/swr1/bw/hitparade/Entstehung-der-Hitparaden-Die-Geschichte-der-Musikcharts,article-swr-9084~_detailPage-2_-f1ad0353fb529fe17479d652df37069709a632a1.html | 24.02.20 |
| 7 | In welchem Jahr wurde die erste Hitparade in dem Billboard Magazine veröffentlicht? | 1936 | https://www.swr.de/swr1/bw/hitparade/Entstehung-der-Hitparaden-Die-Geschichte-der-Musikcharts,article-swr-9084~_detailPage-4_-842110cd7a51e00732d796bf23f7b4a09ca4e942.html | 24.02.20 |
| 7 | In welchem Jahr erschien die erste Ausgabe des Billboard Magazine in den USA? | 1894 | http://events-in-music.com/fascinating-history-billboard-magazine/ | 24.02.20 |
| 7 | In welchem Jahr entstand die Melodie von „Happy Birthday to you“? | 1893 | http://events-in-music.com/who-wrote-happy-birthday-to-you/ | 24.02.20 |
| 7 | In welchem Jahr fand der erste Wiener Opernball statt? | 1935 | https://www.wiener-staatsoper.at/opernball/geschichte | 24.02.20 |
| 7 | In welchem Jahr wurde das erste Musical überhaupt produziert? | 1866 | http://www.sweet-infernal-noise.de/musical.php | 24.02.20 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 7 | In welchem Jahr wurde die Trompete, wie wir sie heute kennen, erfunden? | 1832 | http://www.werhatxerfunden.com/wer-hat-die-trompete-erfunden/ | 24.02.20 |
| 7 | In welchem Jahr wurde der erste Synthesizer der Welt erfunden? | 1964 | https://www.musikunterricht.de/synthesizer | 24.02.20 |
| 7 | In welchem Jahr wurde die Stimmgabel erfunden? | 1711 | https://www.amton-klang.de/stimmgabeln/ | 24.02.20 |
| 7 | Wann war die Uraufführung von „Stille Nacht, Heilige Nacht“? | 1818 | https://www.theology.de/kirche/kirchenjahr/weihnachtslieder.php | 24.02.20 |
| 7 | In welchem Jahr entstand das Weihnachtslied „O Tannenbaum“? | 1820 | https://www.theology.de/kirche/kirchenjahr/weihnachtslieder.php | 24.02.20 |
| 7 | In welchem Jahr wurde das Orchester der Berliner Philharmoniker gegründet? | 1882 | https://www.berliner-philharmoniker.de/geschichte/anfang/#event-grundung-eines-neuen-orchesters | 24.02.20 |
| 7 | In welchem Jahr wurde in Leipzig die älteste Musikschule Deutschlands gegründet? | 1843 | https://www.hmt-leipzig.de/home/hochschule/geschichte_hmt | 24.02.20 |
| 7 | In welchem Jahr wurde das Patent für die elektromechanische Hammond-Orgel angemeldet? | 1934 | https://mysoundbook.eu/laurens-hammond-der-erfinder-der-hammond-orgel/ | 24.02.20 |
| 7 | In welchem Jahr wurde das Patent für das Akkordeon angemeldet? | 1829 | http://www.akkordeon.com/index/his/de_his_inv_dev.shtml | 24.02.20 |
| 7 | In welchem Jahr wurde der Radiosender BBC gegründet? | 1922 | https://www.welt.de/welt_print/article1302985/BBC-Geschichte.html | 24.02.20 |
| 7 | In welchem Jahr eröffnete die erste Diskothek Deutschlands? | 1959 | https://www.faz.net/aktuell/gesellschaft/jugend-schreibt/jugend-schreibt/die-erste-diskothek-made-in-germany-11489624.html | 24.02.20 |
| 7 | In welchem Jahr fand das Festival „Rock am Ring“ das erste Mal statt? | 1985 | https://de.wikipedia.org/wiki/Rock_am_Ring | 24.02.20 |

**Tabelle Anhang A** (fortgesetzt)

| Studie | Aufgabe | Wahrer Wert | Quelle | Abgerufen |
|---|---|---|---|---|
| 7 | In welchem Jahr begann der Bau des Sydney Opera House? | 1959 | https://de.wikipedia.org/wiki/Sydney_Opera_House | 24.02.20 |
| 7 | In welchem Jahr wurde die Oper „Don Giovanni“ von Wolfgang Amadeus Mozart uraufgeführt? | 1787 | https://de.wikipedia.org/wiki/Don_Giovanni | 24.02.20 |
| 7 | In welchem Jahr wurde der berühmte Song „My Way“ von Frank Sinatra veröffentlicht? | 1969 | https://www.discogs.com/Frank-Sinatra-My-Way/master/144074 | 24.02.20 |
| 7 | In welchem Jahr wurde Aretha Franklin als erste Frau in die „Rock And Roll Hall Of Fame“ aufgenommen? | 1987 | https://www.songfacts.com/facts/aretha-franklin | 24.02.20 |
| 7 | In welchem Jahr wurde das Audio-Format „MP3“ entwickelt? | 1982 | https://de.wikipedia.org/wiki/MP3 | 24.02.20 |
| 7 | In welchem Jahr wurde die renommierte „Juilliard School“ an der Fifth Avenue in New York City gegründet? | 1905 | https://de.wikipedia.org/wiki/Juilliard_School | 24.02.20 |
| 7 | In welchem Jahr erschien das Album „Kind of Blue“ von Miles Davis? | 1959 | https://de.wikipedia.org/wiki/Kind_of_Blue | 24.02.20 |

## Anhang B: Übersicht zum Weisheitsbewusstsein mit Ausschluss extremer Fehler

### Tabelle Anhang B

*Übersicht von DJR, TDR, & CDR für alle Experimentalbedingungen und Aufgabenkategorien mit Ausschluss extremer Fehler (10%).*

| Stichprobe/Aufgabenkategorie | | DJR | | TDR | | CDR | |
|---|---|---|---|---|---|---|---|
| | *N* | *M* | *SD* | *M* (vs. .50) | *SD* | *M* (vs. .50) | *SD* |
| Studie 1 | | | | | | | |
| Historisch (Globalisierung) | 100 | .79 | .20 | .55* | .22 | .56** | .22 |
| Unten begrenzt | 100 | .85 | .18 | .58*** | .22 | .67*** | .21 |
| Prozent | 99 | .76 | .21 | .47 | .19 | .53 | .23 |
| Gesamt[a] | 100 | .80 | .16 | .54††† | .10 | .60*** | .11 |
| Studie 2 (fremde Urteile) | | | | | | | |
| Historisch (Globalisierung) | 75 | .79 | .20 | .66*** | .16 | .64*** | .16 |
| Unten begrenzt | 75 | .84 | .15 | .68*** | .14 | .69*** | .18 |
| Prozent | 75 | .83 | .15 | .57*** | .18 | .64*** | .17 |
| Gesamt[a] | 75 | .82 | .14 | .64††† | .09 | .66*** | .09 |
| Studie 2 (eigene Urteile) | | | | | | | |
| Historisch (Globalisierung) | 73 | .74 | .22 | .60*** | .17 | .60*** | .20 |
| Unten begrenzt | 74 | .79 | .20 | .63*** | .16 | .71*** | .18 |
| Prozent | 74 | .71 | .24 | .53 | .16 | .56** | .19 |
| Gesamt[a] | 74 | .75 | .19 | .59††† | .10 | .63*** | .12 |
| Studie 3 | | | | | | | |
| Kalorien | 289 | .67 | .22 | .56*** | .21 | .59*** | .22 |
| Studie 4 (EG + KI) | | | | | | | |
| Alter | 167 | .65 | .27 | .50 | .19 | .58*** | .20 |
| Studie 4 (EG) | | | | | | | |
| Alter[a] | 174 | .70 | .27 | .49 | .18 | .52 | .18 |
| Studie 5 (EG) | | | | | | | |
| Historisch (Globalisierung) | 140 | .62 | .24 | .58*** | .22 | .57*** | .20 |
| Unten begrenzt | 139 | .68 | .23 | .59*** | .20 | .62*** | .19 |
| Gesamt[a] | 140 | .65 | .21 | .58††† | .15 | .60*** | .15 |
| Studie 6 | | | | | | | |
| Historisch (Globalisierung) | 138 | .68 | .25 | .58*** | .19 | .60*** | .19 |
| Unten begrenzt | 138 | .71 | .24 | .58*** | .17 | .65*** | .19 |
| Gesamt[a] | 138 | .69 | .23 | .58††† | .13 | .62*** | .13 |
| Studie 7 | | | | | | | |
| Historisch (Musik) | 73 | .58 | .27 | .57*** | .15 | .59*** | .14 |
| Studie 8 (R+) | | | | | | | |
| Kalorien[a] | 86 | .67 | .29 | .61††† | .26 | .63*** | .25 |
| Studie 8 (R-) | | | | | | | |
| Kalorien[a] | 86 | .55 | .28 | .54 | .25 | .54 | .26 |

*Anmerkung*. Version von Tabelle 2 (siehe Abschnitt 4) mit Ausschluss von je 10% der Beobachtungen pro Stimulus mit extremen absoluten Fehlern im Initialurteil (Phase 1). DJR = Directional Judgment Rate. TDR = Truth Detection Rate. CDR = Crowd Detection Rate. EG =

Experimentalgruppe. EG + KI = Experimentalgruppe + Abfrage Konfidenzintervall. R+ = Bedingung mit Ratschlägen. R- = Bedingung ohne Ratschläge.

[a] Einseitiger Test präregistriert (zweiseitige Tests in allen anderen Fällen).

† $p < .05$, einseitig. †† $p < .01$, einseitig. ††† $p < .001$, einseitig. * $p < .05$, zweiseitig. ** $p < .01$, zweiseitig. *** $p < .001$, zweiseitig.

**Anhang C: Digitaler Anhang**

Der digitale Anhang (Daten, Auswertungsskript, Material) ist Online unter https://osf.io/y23fs/ zu finden. Siehe Abschnitt 3 für weitere Details.