HEINRICH HEINE
UNIVERSITÄT DÜSSELDORF

Untersuchungen zur zeitlichen Stabilität und zur Vermeidbarkeit der Wahrheitsillusion

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von
Frank Calio
aus Köln

Düsseldorf, Februar 2019

aus dem Institut für Experimentelle Psychologie
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathematisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent: Prof. Dr. Jochen Musch
Korreferent: Prof. Dr. Axel Buchner

Tag der mündlichen Prüfung: 19.03.2019

Inhaltsverzeichnis

Zusammenfassung

Der Begriff *Wahrheitsillusion* beschreibt das Phänomen, dass Menschen wiederholt präsentierten Aussagen typischerweise einen höheren Wahrheitsgehalt zuschreiben als Aussagen, denen sie nicht zuvor begegnet sind. Zahlreiche Studien zeigen, dass die Wahrheitsillusion unter verschiedenen Rahmenbedingungen zuverlässig auftritt und nicht nur für wahre Aussagen, sondern auch für faktisch falsche Aussagen beobachtet werden kann. Eine der zentralen Fragestellungen der vorliegenden Arbeit ist, ob die Wahrheitsillusion ein individuell stabiles Phänomen ist, ob es also Personen gibt, die zu verschiedenen Zeitpunkten konsistent anfälliger für den Einfluss der Wiederholung auf die beurteilte Glaubwürdigkeit von Aussagen sind als andere Personen. Um diese Fragestellung zu untersuchen, wurden zwei Experimente durchgeführt, in denen die Stärke der Wahrheitsillusion zu verschiedenen Testzeitpunkten erfasst und die Test-Retest-Stabilität der Wahrheitsillusion bestimmt wurde. In beiden Experimenten fiel die Test-Retest-Stabilität für drei verschiedene Operationalisierungen der Stärke der Wahrheitsillusion äußerst gering aus. Dieser Befund hat weitreichende Konsequenzen, da die Stabilität interindividueller Unterschiede eine notwendige Voraussetzung dafür ist, dass replizierbare Zusammenhänge zwischen der Wahrheitsillusion und bestimmten Persönlichkeitsmerkmalen oder kognitiven Traits beobachtet werden können. Da Wiederholung nicht nur die beurteilte Glaubwürdigkeit wahrer Aussagen, sondern auch die Glaubwürdigkeit falscher Informationen zu steigern vermag, kann die Wahrheitsillusion nachteilige gesellschaftliche Konsequenzen haben, etwa wenn Personen wiederholt unzutreffenden Nachrichtenmeldungen oder nachweislich falschen Werbebotschaften ausgesetzt sind. Die vorliegende Arbeit beschäftigt sich daher auch mit der Frage, ob die Wahrheitsillusion durch explizite Warnungen reduziert oder sogar verhindert werden kann und welche kognitiven Prozesse dabei eine Rolle spielen. Ein gängiges Erklärungsmodell der Wahrheitsillusion geht davon aus, dass Wiederholung zu einer leichteren, „flüs-

sigeren“ kognitiven Verarbeitung von Aussagen führt und dass das resultierende metakognitive Gefühl von Verarbeitungsflüssigkeit als Hinweisreiz zur Beurteilung der Wahrheit von Aussagen genutzt wird. Warnungen könnten die Wahrheitsillusion reduzieren, indem sie die Validität der Verarbeitungsflüssigkeit als Hinweisreiz zur Beurteilung der Wahrheit von Aussagen infrage stellen. In Übereinstimmung mit früheren Ergebnissen zeigte sich in einem dritten Experiment eine substanziell reduzierte Wahrheitsillusion für diejenige Hälfte der Versuchsteilnehmer, die über die Wahrheitsillusion aufgeklärt und vor dieser gewarnt worden war. Explizite Warnungen sowie die Aussicht auf eine finanzielle Belohnung für möglichst akkurate Wahrheitsurteile genügten jedoch nicht, um die Wahrheitsillusion vollständig zu eliminieren. Modellbasierte Analysen mithilfe multinomialer Verarbeitungsbäume zeigten darüber hinaus, dass Warnungen vor der Wahrheitsillusion zu einer deutlichen Reduktion der Wahrscheinlichkeit führten, mit der Personen die Validität von wiederholten Aussagen auf der Basis perzipierter Verarbeitungsflüssigkeit beurteilten. Explizite Warnungen wirkten sich jedoch nicht auf den Abruf von Vorwissen sowie das Rateverhalten der Versuchsteilnehmer aus. Zusammengenommen demonstrieren die Ergebnisse, dass die Wahrheitsillusion ein zuverlässig auftretendes und schwierig zu eliminierendes Phänomen ist. Verschwindend geringe Test-Retest-Korrelationen legen jedoch erstmals nahe, dass die Wahrheitsillusion entweder kein individuell stabiles Phänomen ist oder in üblichen Versuchsanordnungen nicht ausreichend reliabel gemessen wird.

Abstract

The finding that people typically ascribe a higher truth value to statements that are presented repeatedly compared to statements that they have heard for the first time is known as the *truth effect*. Previous research has shown that the truth effect occurs under a large variety of conditions and can be observed for both true and false statements. One of the key questions of the present work is whether the truth effect is an individually stable phenomenon, i.e., whether there are people that are consistently more susceptible to the influence of repetition on judged truth at different points in time. To address this question, two experiments were conducted, each of which assessed the strength of the truth effect on two separate occasions. The test-retest stability was found to be low for three different operationalizations of the truth effect's magnitude. This finding is of major importance for the investigation of the truth effect from an individual difference perspective, as replicable correlates between the truth effect and cognitive or personality traits can only be expected if the truth effect is stable on an individual level. Because repetition not only enhances the credibility of true statements, but also boosts the perceived validity of false statements, the truth effect may have detrimental societal consequences, for example if people are repeatedly exposed to factually false news headlines or inaccurate advertising slogans. The present work therefore presents a third experiment that aimed to reduce or even eliminate the truth effect using explicit warnings. Of particular interest was the question how warnings affect the cognitive processes that are assumed to underlie judgments of truth. A common explanation of the truth effect assumes that repetition increases the ease with which statements are processed and that people use the metacognitive experience of processing fluency as a cue for truth. Warnings may reduce the truth effect by questioning the informational value of fluency as a cue. In line with previous findings, the truth effect was substantially reduced for participants that were explicitly informed and warned about the influence of repetition on judged truth. However, explicit warnings and the prospect of a substantial financial incentive in return for accurate judgments of truth were not sufficient to completely eliminate the truth effect. Extending previous research, a model-based analysis using multinomial processing trees showed that warnings led to a marked reduction in the probability with which participants relied on processing fluency when judging a repeated statement's truth. However, warnings did not affect the retrieval of previous knowledge or participants' guessing behavior relative to a control condition. Taken together, the present work demonstrates that the truth effect is a robust phenomenon that is difficult to eliminate. However, surprisingly low test-retest correlations suggest that the truth effect may not be an individually stable phenomenon or cannot be measured reliably in common experimental designs.

1 Einleitung

In einer klassischen Studie über die Verbreitung von kriegsbezogenen Gerüchten beobachteten Allport und Lepkin (1945), dass die von ihnen befragten Personen ein bestimmtes Gerücht eher für wahr hielten, wenn sie diesem nach eigener Aussage zuvor bereits begegnet waren. Allport und Lepkin (1945) folgerten daraus, dass sich die wiederholte Rezeption eines Gerüchtes positiv auf dessen wahrgenommene Glaubwürdigkeit auswirkt. Experimentelle Evidenz für den Einfluss der Wiederholung auf die beurteilte Glaubwürdigkeit einer Aussage erbrachte die Arbeit von Hasher, Goldstein und Toppino (1977). Im Rahmen dieser Studie beurteilten Personen zu mehreren Messzeitpunkten im Abstand von jeweils zwei Wochen die Validität von wahren und falschen Allgemeinwissensaussagen. Ein Teil der Aussagen wurde dabei zu jedem der drei Messzeitpunkte vorgegeben; in jeder Sitzung beurteilten die Versuchsteilnehmer jedoch auch neue Aussagen, die ihnen nicht zuvor präsentiert worden waren. Die Ergebnisse der Studie zeigten, dass die beurteilte Validität wiederholter Aussagen über die Messzeitpunkte hinweg anstieg, während die beurteilte Glaubwürdigkeit neuer Aussagen zwischen den einzelnen Sitzungen ähnlich ausfiel.

Inzwischen liegen zahlreiche weitere Studien vor, die ebenfalls belegen, dass die Wiederholung einer Aussage zu einem Anstieg ihrer beurteilten Glaubwürdigkeit führt. Diese sogenannte *Wahrheitsillusion* (engl. *truth effect*; Schwartz, 1982) gilt heute als robustes Phänomen, das unter verschiedensten Bedingungen auftritt und für unterschiedliche Aussagentypen beobachtet werden kann (für einen Überblick vgl. Dechêne, Stahl, Hansen und Wänke, 2010). So steigerte eine wiederholte Präsentation etwa die beurteilte Glaubwürdigkeit von Allgemeinwissensaussagen (z.B. Bacon, 1979; Hasher et al., 1977; Nadarevic & Erdfelder, 2014), Meinungen (Arkes, Hackett & Boehm, 1989), Gerüchten (DiFonzo, Beckstead, Stupak & Walders, 2016) und Werbebotschaften (z.B. Hawkins & Hoch, 1992; Johar & Roggeveen, 2007; Law, Hawkins & Craik, 1998; Roggeveen & Johar, 2002). Die Wahrheitsillusion tritt dabei nicht nur für wahre, sondern auch für nachweislich falsche Aussagen auf (z.B. Gigerenzer, 1984; Hasher et al., 1977). Pennycook, Cannon und Rand (2018) beobachteten beispielsweise in einer kürzlich durchgeführten Studie, dass wiederholt dargebotene „Fake News“, also faktisch falsche Nachrichtenschlagzeilen, höhere Korrektheitsurteile erhielten als vergleichbare Schlagzeilen, die nicht zuvor präsentiert worden waren. Befunde wie diese legen nahe, dass die Wahrheitsillusion nachteilige gesellschaftliche Konsequenzen haben kann, zum Beispiel wenn durch das Mittel der Wiederholung der Glaube an faktisch falsche Informationen zu gesellschaftlich relevanten Themen, wie etwa dem Impfen, gestärkt wird. In einer Studie von Fazio, Brashier, Payne und Marsh (2015) trat die Wahrheitsillusion darüber hinaus nicht nur für falsche Aussagen auf, für die die Versuchsteilnehmer mutmaßlich kein Vorwissen besaßen, sondern konnte auch für diejenigen falschen Aussagen beobachtet werden, für die die Versuchsteilnehmer in einem späteren Multiple-Choice-Test Wissen nachwiesen. Fazio et al. (2015) interpretierten diese Ergebnisse als Beleg dafür, dass sogar konkretes Wissen nicht immer vor der Wahrheitsillusion zu schützen vermag.

Ein weit verbreiteter Ansatz zur Erklärung der Wahrheitsillusion basiert auf dem Konzept der *Verarbeitungsflüssigkeit* (engl. *processing fluency*; vgl. Unkelbach, 2007). Nach diesem Erklärungsansatz tritt die Wahrheitsillusion auf, weil eine wiederholte Vorgabe zu einer leichteren, „flüssigeren“ Verarbeitung von Aussagen führt und Personen dieses metakognitive Gefühl der Verarbeitungsflüssigkeit typischerweise als Hinweisreiz zur Beurteilung der Wahrheit einer Aussage nutzen (Unkelbach, 2007). Silva, Garcia-Marques und Mello (2016) zufolge wirkt sich die Wiederholung von Aussagen dabei gleich in zweierlei Hinsicht auf die Leichtigkeit ihrer Verarbeitung aus: Zum einen steigert die frühere Begegnung mit einer Aussage die *konzeptuelle* Verarbeitungsflüssigkeit, indem relevante Konzepte im semantischen Netzwerk aktiviert werden und die Bedeutung der Aussage erfasst wird. Zum anderen führt die wiederholte Vorgabe von Aussagen zu einer Steigerung der *perzeptuellen* Verarbeitungsflüssigkeit, weil etwa Wort- und Satzstruktur einer wiederholten Aussage bereits bekannt sind und die erneute Verarbeitung dementsprechend leichter vonstattengeht. Empirische Evidenz für die Rolle der Verarbeitungsflüssigkeit bei der Beurteilung der Wahrheit von Aussagen erbrachten Reber und Schwarz (1999), die in ihrem Experiment allerdings nicht die Auftretenshäufigkeit von Aussagen manipulierten, sondern deren Lesbarkeit. Aussagen, die in einem hohen Kontrastverhältnis zum Hintergrund dargeboten wurden, demzufolge also mutmaßlich leichter zu verarbeiten waren, wurden von den Versuchsteilnehmern als wahrer beurteilt als solche Aussagen, die ein niedriges Kontrastverhältnis zum Hintergrund aufwiesen. Wiederholung ist also lediglich eines von vielen Mitteln zur Beeinflussung der empfundenen Verarbeitungsflüssigkeit (einen Überblick geben Alter & Oppenheimer, 2009).

Obwohl die Wahrheitsillusion seit über 40 Jahren Gegenstand empirischer Forschung ist, bestehen zahlreiche offene Fragen. So wurde bereits wiederholt darauf hingewiesen, dass interindividuelle Unterschiede in der Stärke der Wahrheitsillusion und deren Korrelate unzureichend erforscht sind (Arkes, Boehm & Xu, 1991; Dechêne et al., 2010). Sind manche Personen also empfänglicher für die Wahrheitsillusion als andere? Und wenn ja: Was kennzeichnet diejenigen Personen, die besonders anfällig (oder aber wenig anfällig) für die Wahrheitsillusion sind? Da die zeitliche Stabilität eines Phänomens eine notwendige Voraussetzung dafür ist, dass replizierbare Korrelationen zwischen dem Phänomen und Persönlichkeitsmerkmalen oder kognitiven Traits beobachtet werden können (vgl. Michalkiewicz & Erdfelder, 2016), war ein Ziel der vorliegenden Arbeit die Bestimmung der Test-Retest-Stabilität der Wahrheitsillusion. Evidenz für die Stabilität der Wahrheitsillusion läge beispielsweise dann vor, wenn diejenigen Personen, die zu einem ersten Testzeitpunkt einer überdurchschnittlich bzw. unterdurchschnittlich starken Wahrheitsillusion unterliegen, tendenziell dieselben Personen sind, die auch zu einem zweiten Testzeitpunkt eine überdurchschnittlich bzw. unterdurchschnittlich starke Wahrheitsillusion aufweisen.

Neben dieser differentialpsychologischen Fragestellung beschäftigt sich die vorliegende Arbeit auch mit der Frage, ob und wie die Wahrheitsillusion reduziert oder gar verhindert werden kann. Diese Fragestellung ist von hoher praktischer Relevanz, da die Wahrheitsillusion nachteilige gesellschaftliche Konsequenzen haben kann, zum Beispiel wenn Personen im Alltag wiederholt unzutreffenden Nachrichtenmeldungen, Gerüchten oder nachweislich falschen Werbebotschaften ausgesetzt sind. Die vorliegende Arbeit prüft – ausgehend von einer Studie von Nadarevic und Aßfalg (2017) –, ob sich die Wahrheitsillusion eliminieren lässt, wenn explizit vor ihr gewarnt wird und Personen mit einem finanziellen Anreiz motiviert werden, möglichst akkurate und unverzerrte Wahrheitsurteile abzugeben. Von zentralem Interesse ist dabei die Frage, wie sich Warnungen auf diejenigen kognitiven Prozesse auswirken, die bei der Beurteilung der Wahrheit von Aussagen mutmaßlich eine Rolle spielen. Zur Beantwortung dieser Frage werden die in einem entsprechenden Experiment gewonnenen Daten modellbasiert mithilfe eines multinomialen Verarbeitungsbaummodells (vgl. Batchelder & Riefer, 1990; Erdfelder, Cüpper & Auer, 2006) ausgewertet.

2 Experimente 1 & 2

Bereits Arkes et al. (1991) und später Dechêne et al. (2010) wiesen darauf hin, dass interindividuelle Unterschiede in der Wahrheitsillusion und deren Korrelate unzureichend erforscht sind. Dabei wäre durchaus denkbar, dass Menschen unterschiedlich anfällig für die Wahrheitsillusion sind. So vermuteten etwa Dechêne et al. (2010), dass Personen, die bevorzugt intuitiv urteilen und entscheiden, besonders empfänglich für den Einfluss von Wiederholung auf die beurteilte Glaubwürdigkeit von Aussagen sind. Diese Annahme basiert auf der Vorstellung, dass Menschen, die sich häufig auf ihre Intuition verlassen, tendenziell sensibler auf Veränderungen der wahrgenommenen Leichtigkeit der Verarbeitung von Stimuli reagieren und metakognitiven Hinweisreizen wie der Verarbeitungsflüssigkeit bei der Urteilsfindung ein stärkeres Gewicht beimessen. Umgekehrt ergibt sich aus diesen Überlegungen, dass Menschen mit einem ausgeprägten rational-analytischen Denkstil oder Personen mit einer hohen Ausprägung im Persönlichkeitsmerkmal Skeptizismus eine weniger stark ausgeprägte Wahrheitsillusion aufweisen sollten.

Zu den vergleichsweise wenigen Studien, die den Zusammenhang zwischen der Wahrheitsillusion und bestimmten kognitiven Traits oder Persönlichkeitsmerkmalen untersucht haben, zählen die Arbeiten von Arkes et al. (1991) und Boehm (1994). In beiden Studien fand sich kein Unterschied in der Stärke der Wahrheitsillusion zwischen Personen mit einer niedrigen und Personen mit einer hohen Ausprägung im Persönlichkeitsmerkmal *Need for Cognition* (Cacioppo & Petty, 1982). Dieses beschreibt das Ausmaß, in dem Menschen „Engagement und Freude bei Denkaufgaben“ (Bless, Wänke, Bohner, Fellhauer & Schwarz, 1994, S. 147) zeigen. Zwei weitere Arbeiten überprüften die Hypothese, dass Personen mit einer hohen Ausprägung im Persönlichkeitsmerkmal *Skeptizismus* weniger anfällig für die Wahrheitsillusion sind als Personen mit einer niedrigen Ausprägung. Während die Studie von Kim (2001) keine klaren Ergebnisse im Hinblick auf die Fragestellung erbrachte, da der Zusammenhang zwischen Skeptizismus und der Wahrheitsillusion in Abhängigkeit vom verwendeten Aussagenmaterial unterschiedlich ausfiel, berichteten DiFonzo et al. (2016) von einem hypothesenkonformen, jedoch schwachen Effekt von Skeptizismus auf die Stärke der Wahrheitsillusion. In einer weiteren Studie nahmen Sundar, Kardes und Wright (2015) an, dass Personen mit einer hohen Ausprägung im Persönlichkeitsmerkmal *Need for Affect* (Maio & Esses, 2001) anfälliger für die Wahrheitsillusion sein würden als Personen mit einer geringen Ausprägung. Dieser Hypothese lag die Überlegung zugrunde, dass Personen mit einem hohen Bedürfnis nach emotionalem Erleben und Verhalten tendenziell sensibler auf Gefühle der Verarbeitungsflüssigkeit reagieren und dass diese Gefühle für sie eher entscheidungsrelevant sind als für Personen mit einem gering ausgeprägten Bedürfnis. Tatsächlich beobachteten Sundar et al. (2015) nur für Personen mit einer hohen Ausprägung im Persönlichkeitsmerkmal Need for Affect eine Wahrheitsillusion. Neben diesen Befunden fanden sich in einer aktuellen Studie von De keersmaecker, Roets, Pennycook und Rand (2018) keine Belege für einen Einfluss der kognitiven Leistungsfähigkeit von Personen auf die Stärke der Wahrheitsillusion. Individuelle Unterschiede im *Cognitive Reflection Test* (Frederick, 2005), der die Neigung von Personen erfasst, sich aufdrängende, intuitive Antworten zu wählen und diese nicht ausreichend zu hinterfragen bzw. zu korrigieren, moderierten die Wahrheitsillusion ebenfalls nicht. De keersmaecker et al. (2018) untersuchten auch einen möglichen Zusammenhang zwischen dem *Bedürfnis nach kognitiver Geschlossenheit* (Webster & Kruglanski, 1994) und der Stärke der Wahrheitsillusion. Dabei bezeichnet das Bedürfnis nach kognitiver Geschlossenheit den „Wunsch eines Individuums nach einer sicheren, eindeutigen Antwort auf eine Frage oder eine Problemstellung und eine Abneigung gegenüber Ambiguität“ (Kruglanski & Webster, 1996, S. 264, eigene Übersetzung). Da das Bedürfnis nach kognitiver Geschlossenheit auch mit einer eingeschränkten Informationsverarbeitung in Verbindung gebracht wird (vgl. Schlink & Walther, 2007), nahmen De keersmaecker et al. (2018) an, dass sich Personen mit einem hohen Bedürfnis nach kognitiver Geschlossenheit bei der Beurteilung des Wahrheitsgehaltes von Aussagen verstärkt auf metakognitive Gefühle der Verarbeitungsflüssigkeit verlassen. Belege für diese Annahme fanden sie in ihren Daten jedoch nicht. Zuletzt untersuchten De keersmaecker et al. (2018) auch den Einfluss rationaler bzw. intuitiver Denkstile auf die Wahrheitsillusion, indem sie ihren Probanden das *Rational-Experiential Inventory* (REI) von Pacini und Epstein (1999) vorgaben. Individuelle Unterschiede in der Tendenz zu systematischer bzw. analytisch-rationaler Verarbeitung moderierten die Stärke der Wahrheitsillusion nicht. Personen mit einem stark ausgeprägten experientiell-intuitiven Denkstil zeigten zwar in einem ersten Experiment von De keersmaecker et al. (2018) eine stärkere Wahrheitsillusion als Personen mit einer schwachen Ausprägung, dieses Ergebnis ließ sich jedoch in einem zweiten Experiment nicht replizieren. In einer Meta-Analyse von Dechêne et al. (2010) fand sich darüber hinaus kein moderierender Einfluss des Alters von Personen auf die Stärke der Wahrheitsillusion.

Neben diesen Befunden wurden auch im Bereich der Neuropsychologie erste Versuche unternommen, personenbezogene Einflüsse auf die Stärke der Wahrheitsillusion zu identifizieren. So suchte Ladowsky-Brooks (2010) in einer Stichprobe von Personen mit einem Schädel-Hirn-Trauma nach Zusammenhängen zwischen der Stärke der Wahrheitsillusion und der Leistung in diversen neuropsychologischen Testverfahren, darunter Gedächtnistests und Tests zur Beurteilung exekutiver Funktionen. Dabei zeigten sich über zwei verschiedene Operationalisierungen der Wahrheitsillusion hinweg betrachtet zwar vereinzelte signifikante Zusammenhänge (z.B. zu einem Untertest der Wechsler Memory Scale), jedoch insgesamt keine konsistenten Korrelationsmuster. In einer online durchgeführten Studie untersuchten Moritz et al. (2012) die Wahrheitsillusion in einer Gruppe gesunder Probanden sowie einer Gruppe von Personen, die zuvor in Online-Selbsthilfeforen zum Thema Schizophrenie rekrutiert worden waren. Dabei zeigte sich für beide Personengruppen eine positive Korrelation zwischen der Stärke der Wahrheitsillusion für emotionale, wahnbezogene Aussagen (z.B. “The German federal police uses approximately 3000 cameras for the purpose of video-based face-detection.”, S. 1058) und dem Ausprägungsgrad an Positivsymptomatik, der mithilfe eines Fragebogens erfasst worden war. Die Stärke der Wahrheitsillusion für neutrales Aussagenmaterial korrelierte hingegen nicht mit einer schizophrenieassoziierten Symptomatik.

Der Überblick über die bis dato durchgeführten Studien verdeutlicht, dass die Suche nach Korrelaten interindividueller Unterschiede in der Stärke der Wahrheitsillusion bislang wenig erfolgreich verlief. So konnten für einige ausgewählte Persönlichkeitsmerkmale und kognitive Variablen, für die ein Zusammenhang mit der Wahrheitsillusion aus theoretischer Perspektive plausibel erschien, keine bedeutsamen Zusammenhänge beobachtet werden (Arkes et al., 1991; Boehm, 1994; De keersmaecker et al., 2018). Von besonderem Interesse sind dabei die Befunde von De keersmaecker et al. (2018), die zeigten, dass sich der in einer ersten Studie identifizierte Zusammenhang zwischen einem experientiell-intuitiven Denkstil und der Stärke der Wahrheitsillusion in einem zweiten, präregistrierten und teststärkeren Experiment als nicht replizierbar erwies. Eine mögliche Erklärung für diese widersprüchlichen Befunde findet sich bei Michalkiewicz und Erdfelder (2016), die sich mit interindividuellen Unterschieden in der Nutzung der *Rekognitionsheuristik* befassten. Michalkiewicz und Erdfelder (2016) argumentieren, dass replizierbare Korrelate zu einem bestimmten Phänomen nur dann zu erwarten sind, wenn interindividuelle Unterschiede in der Stärke des Phänomens zeitlich stabil auftreten. Die Frage nach der zeitlichen Stabilität eines Phänomens kann durch die Bestimmung seiner Test-Retest-Stabilität beantwortet werden. Individuelle Unterschiede in der Wahrheitsillusion wären etwa dann zeitlich stabil, wenn diejenigen Personen, die zu einem ersten Testzeitpunkt einer überdurchschnittlich bzw. unterdurchschnittlich starken Wahrheitsillusion unterliegen, tendenziell auch zu einem zweiten Testzeitpunkt eine überdurchschnittlich bzw. unterdurchschnittlich starke Wahrheitsillusion aufweisen.

In der vorliegenden Arbeit wurde die Test-Retest-Stabilität der Wahrheitsillusion in zwei Experimenten bestimmt. In jedem Experiment durchliefen die Versuchsteilnehmer wiederholt ein klassisches Wahrheitsillusionsparadigma, wurden also mehrfach um die Beurteilung der Korrektheit von neuen und wiederholten Aussagen gebeten. Die in den beiden Studien verwendeten wahren und falschen Aussagen entstammten einem Pool von insgesamt 404 Aussagen aus unterschiedlichen Themengebieten (z.B. „Der Zwergplanet Pluto benötigt für eine Sonnenumrundung mehr als 200 Jahre.“). Diese Aussagen waren zuvor unter Zuhilfenahme verschiedener Quellen (z.B. Quiz-Bücher) erstellt und in einer Voruntersuchung von mindestens 23 Personen hinsichtlich ihrer wahrgenommenen Korrektheit beurteilt worden. Für die beiden nachfolgend beschriebenen Experimente wurden Aussagen ausgewählt, deren tatsächlicher Wahrheitsstatus den Versuchsteilnehmern in der Voruntersuchung unbekannt war. Eine solche Materialauswahl ist notwendig, da beispielsweise keine Wahrheitsillusion für Aussagen auftreten kann, die von vornherein von nahezu allen Personen als wahr beurteilt werden (vgl. Fazio et al., 2015; Pennycook et al., 2018).

Bei der ersten Studie handelte es sich um ein internetgestützt durchgeführtes Experiment, bei dem die Daten von 166 freiwilligen Versuchsteilnehmerinnen und Versuchsteilnehmern ausgewertet wurden. Das Experiment setzte sich aus zwei Sitzungen zusammen, zwischen denen ein etwa einwöchiges Zeitintervall lag. Jede Sitzung bestand aus drei Phasen: In einer ersten Phase ordneten die Versuchsteilnehmer zunächst eine Menge von wahren und falschen Aussagen verschiedenen Themengebieten (z.B. Erdkunde, Tiere & Pflanzen, Unterhaltung) zu. Anschließend folgte eine zehnminütige, nonverbale Distraktoraufgabe. In einer dritten und letzten Phase wurden den Versuchsteilnehmern dann sowohl neue Aussagen als auch solche Statements präsentiert, die bereits während der ersten Phase der jeweiligen Sitzung gezeigt worden waren. Die Aufgabe der Versuchsteilnehmer bestand darin, die Wahrheit der Aussagen auf einer sechsstufigen Likert-Skala von *definitiv falsch* (1) bis *definitiv wahr* (6) zu beurteilen. Die Stärke der Wahrheitsillusion wurde für jeden

Versuchsteilnehmer in jeder der beiden Sitzungen individuell ermittelt, indem die vergebenen Validitätsurteile für eine bestimmte Menge neuer Aussagen mit den Validitätsurteilen für eine andere Menge wiederholter Aussagen verglichen wurden. Die Bestimmung der Wahrheitsillusion erfolgte somit über das sogenannte *between-items criterion* (Dechêne et al., 2010), also über den Vergleich von Validitätsurteilen für zwei *unterschiedliche* Aussagenmengen.

Die Test-Retest-Stabilität wurde für drei verschiedene Operationalisierungen der Stärke der Wahrheitsillusion ermittelt. Durch dieses Vorgehen sollte überprüft werden, ob die Test-Retest-Stabilität der Wahrheitsillusion für verschiedene Operationalisierungen unterschiedlich hoch ausfallen würde. Ein erster, intuitiv nachvollziehbarer Index für die Stärke der Wahrheitsillusion wurde berechnet, indem die Validitätsurteile für wiederholt präsentierte sowie für neue Aussagen pro Versuchsteilnehmer zunächst separat gemittelt und diese Mittelwerte in einem zweiten Schritt voneinander subtrahiert wurden. Die Test-Retest-Stabilität der Wahrheitsillusion ließ sich dann über die Korrelation zwischen den Differenzwerten aus beiden Experimentalsitzungen bestimmen.

Ein zweiter Index für die Stärke der Wahrheitsillusion wurde auf der Basis signalentdeckungstheoretischer Bias-Parameter gebildet (Stanislaw & Todorov, 1999). Eine signalentdeckungstheoretische Auswertung kam bereits in früheren Arbeiten zur Wahrheitsillusion zum Einsatz (Nadarevic, 2010; Unkelbach, 2007). Die „wahr“-Antwort eines Versuchsteilnehmers bei der Beurteilung der Validität von Aussagen entspricht dabei einer „Signal vorhanden“-Antwort in einem klassischen signalentdeckungstheoretischen Kontext. Bei der Untersuchung der Wahrheitsillusion steht die Berechnung eines Maßes für die Antworttendenz im Vordergrund. Eine Wahrheitsillusion liegt dann vor, wenn die Antworttendenz für wiederholte Aussagen liberaler ausfällt als für neue Aussagen, wenn ein Versuchsteilnehmer also bei wiederholten Aussagen eher zu einer „wahr“-Antwort tendiert als bei nicht zuvor präsentierten Aussagen (Unkelbach, 2007). In beiden Sitzungen von Experiment 1 wurde für jeden Versuchsteilnehmer jeweils ein individueller Index für die Stärke der Wahrheitsillusion berechnet, indem die Antworttendenz für neue Aussagen (c_{neu}) von der Antworttendenz für wiederholt präsentierte Aussagen ($c_{\text{wiederholt}}$) subtrahiert wurde. Die Indizes beider Experimentalsitzungen (Δc_{t1}, Δc_{t2}) wurden anschließend miteinander korreliert, um die Stabilität interindividueller Unterschiede in der Wahrheitsillusion zu ermitteln.

Die Verwendung einfacher Differenzscores war lange Zeit umstritten, unter anderem weil sie als potenziell unreliabel galten (für einen Überblick vgl. Zumbo, 1999); tatsächlich ist diese Kritik nur unter bestimmten Umständen gerechtfertigt (vgl. z.B. Zimmerman & Williams, 1982). Neben einfachen Differenzscores wurde deshalb zur Überprüfung der Verallgemeinerbarkeit der Befunde über unterschiedliche Operationalisierungen hinweg ein zusätzliches Maß für individuelle Unterschiede in der Stärke der Wahrheitsillusion gebildet, welches auf Regressionsresiduen bzw. Residualscores basiert. Die Reliabilität dieser Scores kann in bestimmten Situationen höher ausfallen als die Reliabilität einfacher Differenzen (Williams & Zimmerman, 1983; Zumbo, 1992). Dies zeigte sich zum Beispiel in den Arbeiten von Caruso (2004) sowie Williams, Zimmerman und Mazzagatti (1987). Zur Schätzung der Test-Retest-Stabilität der Wahrheitsillusion auf der Basis von Residualscores wurden in beiden Experimentalsitzungen mithilfe einer linearen Regression zunächst die Validitätsurteile für wiederholt präsentierte Aussagen anhand der Validitätsurteile für neue Aussagen vorhergesagt. Für jeden Versuchsteilnehmer konnte so in jeder der beiden Experimentalsitzungen ein Regressionsresiduum abgeleitet werden. Diese Regressionsresiduen können als Indikator für interindividuelle Unterschiede in der Stärke der Wahrheitsillusion betrachtet werden, weil sie abbilden, inwieweit sich die mittleren Validitätsurteile für wiederholte Aussagen von denjenigen Werten unterscheiden, die auf der Basis der mittleren Validitätsurteile für neue Aussagen zu erwarten gewesen wären (zur Interpretation von Residualscores siehe Cronbach &

Furby, 1970; vgl. aber auch Willett, 1988). So bedeutet beispielsweise ein positives Residuum, dass ein Versuchsteilnehmer wiederholte Aussagen als wahrer beurteilt hat als auf der Basis seiner Validitätsurteile für neue Aussagen zu erwarten gewesen wäre. Zur Bestimmung der Test-Retest-Stabilität der Wahrheitsillusion wurde die Korrelation zwischen den Regressionsresiduen beider Experimentalsitzungen berechnet.

Die Ergebnisse von Experiment 1 zeigten, dass in beiden Experimentalsitzungen durchschnittlich, d.h. über alle Probanden hinweg betrachtet, eine Wahrheitsillusion aufgetreten war. Wiederholt präsentierte Aussagen wurden sowohl in Sitzung 1 als auch in Sitzung 2 als signifikant wahrer beurteilt als neue, nicht zuvor präsentierte Aussagen. Wurde die Ausprägung der Wahrheitsillusion für jeden Versuchsteilnehmer individuell über einfache Mittelwertsdifferenzen bestimmt, zeigte sich jedoch, dass nicht alle Versuchsteilnehmer einer Wahrheitsillusion unterlagen: So vergaben in der ersten Experimentalsitzung insgesamt 64% der Versuchsteilnehmer höhere Validitätsurteile für wiederholt präsentierte als für neue Aussagen; in der zweiten Experimentalsitzung war dies bei 63% der Versuchsteilnehmer der Fall. Die Korrelation zwischen den einfachen Mittelwertsdifferenzen in beiden Experimentalsitzungen fiel äußerst gering aus und unterschied sich nicht signifikant von null, $r = .04$, $p = .579$. Die individuelle Ausprägung der Wahrheitsillusion in Sitzung 2 ließ sich somit nicht anhand der individuellen Ausprägung der Wahrheitsillusion in Sitzung 1 vorhersagen. Ein nahezu identisches Ergebnis zeigte sich bei der Bestimmung der Test-Retest-Stabilität auf der Basis der signalentdeckungstheoretischen Antworttendenz-Parameter, $r = .05$, $p = .547$. Auch die Korrelation zwischen den in beiden Experimentalsitzungen ermittelten Residualscores fiel mit $r = .16$ niedrig aus, war jedoch signifikant von null verschieden, $p = .034$, und signifikant größer als die Test-Retest-Stabilität, die auf Basis der einfachen Mittelwertsdifferenzen ermittelt worden war ($r = .04$), $z = -3.14$, $p = .002$ (Diedenhofen & Musch, 2015, unter Anwendung der Formeln von Raghunathan, Rosenthal & Rubin, 1996).

Insgesamt erbrachte Experiment 1 keine Belege für eine nennenswerte Test-Retest-Stabilität der Wahrheitsillusion. In der beschriebenen Studie wurde die Wahrheitsillusion dabei in beiden Sitzungen über den Vergleich von Validitätsurteilen für zwei *verschiedene* Aussagenmengen erfasst, von denen die eine Aussagenmenge wiederholt, die andere jedoch stets nur einmal präsentiert worden war. Alternativ zur Bestimmung der Wahrheitsillusion über dieses sogenannte *between-items criterion* wurde der Effekt der Wiederholung auf die beurteilte Glaubwürdigkeit von Aussagen in bisherigen Forschungsarbeiten oft auch über den Vergleich von Validitätsurteilen für *eine* bestimmte, zu mehreren Gelegenheiten vorgegebene Aussagenmenge erfasst. Dieser Vergleich wendet ein *within-items criterion* an (vgl. Dechêne et al., 2010). Ziel von Experiment 2 war die Überprüfung der Replizierbarkeit der Befunde von Experiment 1 für diese alternative Methode zur Erfassung der Wahrheitsillusion.

Experiment 2 fand im Vorlesungskontext mit einer studentischen Stichprobe statt. Ausgewertet wurden die Daten von 116 Versuchsteilnehmerinnen und Versuchsteilnehmern, die an insgesamt *drei* verschiedenen Experimentalsitzungen im Abstand von acht bzw. sieben Tagen teilgenommen hatten. In jeder Sitzung beurteilten die Versuchsteilnehmer 40 kritische und 8 Filler-Aussagen auf einer sechsstufigen Likert-Skala von *definitiv falsch* (1) bis *definitiv wahr* (6). Die Hälfte der kritischen Aussagen, die in Sitzung 1 präsentiert worden war, wurde in Sitzung 2 erneut dargeboten, gemeinsam mit 20 neuen Aussagen. Diejenigen Aussagen, die erstmals in Sitzung 2 vorgegeben worden waren, wurden auch in Sitzung 3 präsentiert, wiederum gemeinsam mit 20 Aussagen, die nicht zuvor gezeigt worden waren. Durch diese Versuchsanordnung ließ sich die Stärke der Wahrheitsillusion sowohl über den Vergleich von Validitätsurteilen für diejenige Aussagenmenge bestimmen, die in Sitzung 1 sowie in Sitzung 2 präsentiert worden war, als auch über den Vergleich von Validitätsurteilen für diejenige Aussagenmenge, die erstmals in Sitzung 2 und später in Sitzung 3 dargeboten wurde.

Die Ergebnisse von Experiment 2 zeigten, dass Aussagen während ihrer zweiten Präsentation als signifikant wahrer beurteilt wurden als bei ihrer ersten Darbietung. Dies galt sowohl für diejenigen Aussagen, die in Sitzung 1 und in Sitzung 2 präsentiert worden waren, als auch für diejenigen Aussagen, die erstmalig in Sitzung 2 und später erneut in Sitzung 3 auftraten. Wie auch in Experiment 1, wurde die Test-Retest-Stabilität der Wahrheitsillusion auf der Basis dreier verschiedener Indizes bestimmt. Zunächst wurden für jeden Versuchsteilnehmer einfache Differenzscores berechnet, indem die für die wiederholt präsentierte Aussagenmenge vergebenen mittleren Validitätsurteile in Sitzung 1 und 2 voneinander subtrahiert wurden. Gleiches geschah mit den mittleren Validitätsurteilen für diejenige Aussagenmenge, die die Versuchsteilnehmer in Sitzung 2 und in Sitzung 3 beurteilt hatten. Beide Differenzen fielen für die überwiegende Mehrheit der Versuchsteilnehmer positiv aus (57% und 69%). Wiederholt vorgegebene Aussagen wurden also während ihrer zweiten Darbietung von der Mehrheit der Versuchsteilnehmer, nicht jedoch von allen Versuchsteilnehmern, als wahrer eingeschätzt als bei ihrer ersten Darbietung. Die Test-Retest-Stabilität der Wahrheitsillusion wurde ermittelt, indem die für die Versuchsteilnehmer vorliegenden Differenzwertpaare miteinander korreliert wurden. Diese Korrelation fiel gering aus und unterschied sich nicht signifikant von null, $r = .12$, $p = .191$. Ein ähnliches Ergebnismuster zeigte sich, wenn diejenigen Indizes für die Stärke der Wahrheitsillusion miteinander korreliert wurden, die auf der Basis von signalentdeckungstheoretischen Antworttendenz-Parametern berechnet worden waren, $r = .06$, $p = .499$. Wie in Experiment 1 wurde die Test-Retest-Stabilität der Wahrheitsillusion darüber hinaus auch auf der Basis von Residualscores bestimmt. Die entsprechende Korrelation fiel ebenfalls niedrig aus, $r = .27$, war jedoch signifikant größer als null, $p = .003$, und zudem signifikant größer als die Korrelation, die für die einfachen Differenzscores ermittelt worden war ($r = .12$), $z = -2.70$, $p = .007$.

Zusammengenommen fanden sich weder in Experiment 1 noch in Experiment 2 Belege für eine nennenswerte zeitliche Stabilität der Wahrheitsillusion. Die

Test-Retest-Stabilität fiel in beiden Experimenten gering aus, unabhängig davon, ob die Wahrheitsillusion über das between-items criterion oder das within-items criterion (Dechêne et al., 2010) erfasst wurde. Ähnliche Ergebnisse traten zudem für drei verschiedene Operationalisierungen der Stärke der Wahrheitsillusion auf, obwohl die Test-Retest-Stabilität etwas höher ausfiel, wenn Residualscores als Maß für interindividuelle Unterschiede in der Stärke der Wahrheitsillusion herangezogen wurden. Auch in diesem Fall war die Test-Retest-Stabilität der Wahrheitsillusion jedoch sehr gering und blieb deutlich hinter Schätzern der zeitlichen Stabilität anderer Phänomene aus dem Bereich der Urteils- und Entscheidungsforschung zurück (z.B. Kantner & Lindsay, 2012; Kirby, 2009; Michalkiewicz & Erdfelder, 2016).

3 Experiment 3

Neben der zeitlichen Stabilität interindividueller Unterschiede in der Stärke der Wahrheitsillusion beschäftigt sich die vorliegende Arbeit auch mit der Frage, ob und wie die Wahrheitsillusion reduziert oder gar verhindert werden kann und welche kognitiven Prozesse dabei eine Rolle spielen. Diese Frage ist von hoher praktischer Relevanz, da zahlreiche Studien zeigen konnten, dass eine wiederholte Darbietung nicht nur die beurteilte Glaubwürdigkeit von wahren, sondern auch die beurteilte Glaubwürdigkeit von nachweislich falschen Aussagen ansteigen lässt (vgl. z.B. Gigerenzer, 1984; Hasher et al., 1977; Pennycook et al., 2018). Wichtig ist dabei, dass die Wahrheitsillusion keineswegs nur unter Laborbedingungen auftritt, sondern auch dann beobachtet werden kann, wenn Personen in ihrem natürlichen Umfeld bestimmten Aussagen wiederholt ausgesetzt sind (Boehm, 1994) oder Zufallsstichproben aus der Allgemeinbevölkerung ein klassisches Wahrheitsillusionsparadigma durchlaufen (Gigerenzer, 1984). Da Wiederholung auch die beurteilte Glaubwürdigkeit von Meinungsaussagen (z.B. Arkes et al., 1989) und Werbebotschaften (z.B. Hawkins & Hoch, 1992; Johar & Roggeveen, 2007; Law et al., 1998; Roggeveen & Johar, 2002) zu steigern vermag, ist anzunehmen, dass die Wahrheitsillusion für politische Absichten oder Werbezwecke genutzt werden kann und somit unmittelbaren Einfluss auf die subjektiv empfundene Richtigkeit von Informationen im Alltag hat.

Bis heute erbrachten nur wenige Studien Hinweise darauf, wie der glaubwürdigkeitssteigernde Effekt der Wiederholung auf die beurteilte Wahrheit von Aussagen vermieden oder zumindest reduziert werden kann. Einzelne Arbeiten untersuchten etwa, wie sich direktes Feedback zum Wahrheitsgehalt der präsentierten Aussagen auf die Wahrheitsillusion auswirkt (z.B. Brown & Nix, 1996; Mutter, Lindsey & Pliske, 1995; Skurnik, Yoon, Park & Schwarz, 2005). So zeigte eine Studie von Brown und Nix (1996, Exp. 1), dass wiederholte falsche Aussagen als signifikant weniger wahr beurteilt wurden als neue falsche Aussagen, wenn die Versuchsteilnehmer bei der ersten Begegnung mit diesen Aussagen eine Woche zuvor über den tatsächlichen Wahrheitsstatus der Aussagen informiert worden waren; betrug das Retentionsintervall jedoch einen Monat, kehrte sich das Ergebnismuster um und es trat eine klassische Wahrheitsillusion auf. Befunde wie diese legen nahe, dass Feedback die Wahrheitsillusion lediglich kurzzeitig verhindern kann, der glaubwürdigkeitssteigernde Effekt der Wiederholung aber sogar dann noch wirkt, wenn die Erinnerung an das ursprüngliche Feedback bereits verblasst ist (Brown & Nix, 1996; Skurnik et al., 2005; vgl. aber auch Swire, Ecker & Lewandowsky, 2017). In einer anderen Studie untersuchte Boehm (1994) die Wirksamkeit einer „Accountability"-Manipulation, indem er der Hälfte der Versuchsteilnehmer in einem klassischen Wahrheitsillusionsexperiment mitteilte, dass sie ihre Antworten später vor einer Gruppe von Peers würden rechtfertigen müssen. Boehm (1994) nahm an, dass eine solche Manipulation zu einer tieferen Verarbeitung der Informationen führen würde, was die Wahrheitsillusion reduzieren oder eliminieren sollte. Entgegen seiner Erwartungen konnte Boehm (1994) jedoch keinen Effekt der experimentellen Manipulation nachweisen.

Explizite Warnungen vor dem Einfluss der Wiederholung auf die beurteilte Glaubwürdigkeit von Aussagen kommen als weitere Strategie zur Abschwächung der Wahrheitsillusion infrage. Warnungen wurden bereits in zahlreichen Studien zur Bekämpfung von kognitiven Verzerrungen eingesetzt, allerdings mit unterschiedlichem Erfolg. In verschiedenen Studien konnte beispielsweise der *Rückschaufehler* (engl. *hindsight bias*), also die Tendenz eines Menschen, die Vorhersagbarkeit eines inzwischen eingetretenen Ereignisses zu überschätzen, durch Warnungen nicht bedeutsam reduziert werden (z.B. Bernstein, Wilson, Pernat & Meilleur, 2012; Davies, 1993; Harley, Carlsen & Loftus, 2004; Pohl & Hell, 1996; Sharpe & Adair, 1993). Ein ähnliches Bild zeigte sich für den *Outcome Bias*, der die Tendenz von Personen beschreibt, bei der Beurteilung der Qualität von bereits getroffenen Entscheidungen auch das Ergebnis der Entscheidung zu berücksichtigen, obwohl dieses zum Zeitpunkt der Entscheidung selbstverständlich nicht bekannt war. Einfache Warnungen, bei denen die Versuchsteilnehmer instruiert wurden, den Outcome Bias zu vermeiden, genügten in den Arbeiten von Clarkson, Emby und Watt (2002) sowie Grenier, Peecher und Piercey (2007) nicht, um den Bias zu verhindern. Andererseits wurden explizite Warnungen beispielsweise bei der Bekämpfung des *Fehlinformationseffekts* (engl. *misinformation effect*) durchaus mit Erfolg eingesetzt. Als Fehlinformationseffekt wird das Phänomen bezeichnet, dass die Erinnerungen einer Person an ein Ereignis verzerrt werden, wenn diese Person nachträglich neuen und irreführenden Informationen über das Ereignis ausgesetzt wird (vgl. Loftus, 2005; Loftus, Miller & Burns, 1978). Eine Meta-Analyse von Blank und Launay (2014) zeigte, dass bestimmte Warnungen, sogenannte „enlightenment procedures“, besonders wirksam bei der Bekämpfung des Fehlinformationseffekts sind und diesen sogar gänzlich eliminieren können (vgl. z.B. Oeberst & Blank, 2012). Bei dieser besonderen Form der Warnung werden die Versuchsteilnehmer nicht nur auf die Existenz von Falschinformation hingewiesen, sondern auch über die Logik und wissenschaftliche Motivation der experimentellen Manipulation aufgeklärt (Blank & Launay, 2014).

Bisherige Studien über die Wirksamkeit von Warnungen zur Vermeidung kognitiver Verzerrungen zeichnen also ein sehr uneinheitliches Bild. Dies ist nicht überraschend, wenn die Wirksamkeit von Warnungen – wie von Wilson und Brekke (1994) vermutet – von verschiedenen Faktoren abhängt, wie zum Beispiel von der Motivation der gewarnten Personen, verzerrende Einflüsse zu korrigieren oder der Frage, ob Personen überhaupt dazu in der Lage sind, die relevanten kognitiven Prozesse zu beeinflussen. Letztgenannter Punkt ist von unmittelbarer Relevanz für die Wahrheitsillusion. So nimmt beispielsweise Schwarz (2004) an, dass die Nutzung von Verarbeitungsflüssigkeit als Hinweisreiz bei der Urteilsfindung ein weitgehend automatischer Prozess ist. Verarbeitungsflüssigkeit kann jedoch als Hinweisreiz ungenutzt bleiben, wenn ein Urteiler erkennt, dass die Verarbeitungsflüssigkeit auf eine Quelle zurückgeht, die für das eigentliche Urteil irrelevant ist (Alter & Oppenheimer, 2009; Schwarz, 2004). In einem Experiment von McGlone und Tofighbakhsh (2000) wurden unbekannte Aphorismen in Reimform (z.B. “What sobriety conceals, alcohol reveals”) von den Versuchsteilnehmern als zutreffender beurteilt als semantisch äquivalente Aphorismen, die sich nicht reimten (z.B. “What sobriety conceals, alcohol unmasks”), mutmaßlich weil die Reimform zu einer gesteigerten Verarbeitungsflüssigkeit führte. Wurden die Versuchsteilnehmer jedoch instruiert, die Validität der Aphorismen ausschließlich auf Basis des Aussageninhalts zu beurteilen und sich dabei nicht von deren „poetischer Qualität“ beeinflussen zu lassen, zeigte sich kein Unterschied zwischen den Validitätsurteilen für beide Arten von Aphorismen. McGlone und Tofighbakhsh (2000) interpretierten dies als Beleg dafür, dass entsprechend instruierte bzw. gewarnte Versuchsteilnehmer die erhöhte Verarbeitungsflüssigkeit von Aphorismen in Reimform nicht länger auf deren Validität fehlattribuierten.

Angesichts dieser Befunde ist denkbar, dass auch die wiederholungsbasierte Wahrheitsillusion durch explizite Warnungen reduziert oder sogar vollständig eliminiert werden kann. In zwei Experimenten informierten Nadarevic und Aßfalg (2017) die Hälfte ihrer Versuchsteilnehmer detailliert über die Wahrheitsillusion und forderten sie dazu auf, diese zu verhindern. Tatsächlich fiel die Wahrheitsillusion nach einer solchen Warnung weniger als halb so groß aus wie in einer Gruppe von Personen, die keine Warnung erhalten hatte; in keinem der beiden Experimente konnte die Wahrheitsillusion jedoch gänzlich eliminiert werden. Nadarevic und Aßfalg (2017) interpretierten diese Ergebnisse als Beleg dafür, dass Menschen ein gewisses Ausmaß an Kontrolle über diejenigen Attributionsprozesse besitzen, die für die Wahrheitsillusion verantwortlich gemacht werden: “If fluency is involved in the truth effect, this implies that participants have control over their fluency-truth attributions” (S. 824).

Um zu prüfen, wie sich Warnungen vor der Wahrheitsillusion auf diejenigen kognitiven Prozesse auswirken, die bei der Beurteilung der Wahrheit von Aussagen mutmaßlich eine Rolle spielen, wurde in Experiment 3 der vorliegenden Arbeit ein klassisches Wahrheitsillusionsparadigma eingesetzt, bei dem die Hälfte der Versuchsteilnehmer vor der Wahrheitsillusion gewarnt wurde. Abweichend von den Experimenten von Nadarevic und Aßfalg (2017) erfolgte die Auswertung der Daten jedoch mithilfe eines multinomialen Verarbeitungsbaummodells (Batchelder & Riefer, 1990; Erdfelder et al., 2006). Derartige Modelle wurden bereits in zwei früheren Studien zur Wahrheitsillusion eingesetzt (Fazio et al., 2015; Unkelbach & Stahl, 2009), weil sie einen Blick auf latente kognitive Prozesse erlauben, die dem beobachtbaren Verhalten zugrunde liegen. Darüber hinaus untersuchte Experiment 3 eine weitere Frage, die in der Studie von Nadarevic und Aßfalg (2017) unbeantwortet geblieben war. Nach Wilson und Brekke (1994; vgl. auch Wilson, Centerbar & Brekke, 2002) kann die Wirksamkeit von Debiasing-Manipulationen wesentlich von motivationalen Faktoren abhängen. Angesichts der Tatsache, dass Warnungen die Wahrheitsillusion in zwei Experimenten zwar reduzieren, jedoch nicht eliminieren konnten, diskutierten Nadarevic und Aßfalg (2017) die Frage, ob ihre Versuchsteilnehmer möglicherweise nicht hinreichend motiviert waren, ihre Wahrheitsurteile in einem ausreichenden Maße zu kontrollieren und ggf. anzupassen. Zur Klärung dieser Frage wurde in Experiment 3 erstmals ein finanzieller Anreiz eingesetzt, der die Versuchsteilnehmer dazu animieren sollte, möglichst akkurate und unverzerrte Wahrheitsurteile abzugeben. Da Nadarevic und Aßfalg (2017) in ihrer Studie auch darüber spekulierten, ob sich die Wahrheitsillusion mithilfe einer Warnung eher für Aussagen vermeiden lässt, hinsichtlich derer die Versuchsteilnehmer über relevantes Vorwissen verfügen, wurden in der hier vorgestellten Studie neben maximal schweren auch vergleichsweise leichte Aussagen eingesetzt.

Ausgewertet wurden die Daten von 167 Versuchsteilnehmerinnen und Versuchsteilnehmern, die auf dem Campus der Heinrich-Heine-Universität Düsseldorf rekrutiert worden waren. Die computergestützte Laborstudie gliederte sich in insgesamt drei Abschnitte: In einer ersten Phase ordneten die Versuchsteilnehmer eine

Menge von wahren und falschen Aussagen unterschiedlichen Themengebieten (z.B. Erdkunde, Tiere & Pflanzen, Sport & Unterhaltung) zu. Im Anschluss daran beurteilten alle Teilnehmer in einer zweiten Phase die Wahrheit von Aussagen, von denen ein Teil bereits in der ersten Phase, ein anderer Teil zuvor jedoch nicht präsentiert worden war. Jede Aussage sollte dabei entweder als „wahr" oder als „falsch" klassifiziert werden. Unmittelbar vor dieser zweiten Phase wurde die Hälfte der Versuchsteilnehmer über den Zweck der Untersuchung und die Wahrheitsillusion informiert und darüber hinaus gebeten, die Wahrheitsillusion zu verhindern. Um die Motivation zu erhöhen, sich bei der Beurteilung der Wahrheit von Aussagen ausschließlich auf den Wahrheitsstatus der jeweiligen Aussage zu konzentrieren und sich von einer etwaigen wiederholten Präsentation nicht beeinflussen zu lassen, wurde ein Geldpreis von 30 € für diejenigen zehn Versuchsteilnehmer in Aussicht gestellt, die die meisten Aussagen korrekt klassifizieren würden. Eine abschließende dritte Experimentalphase sollte lediglich dabei helfen einzuschätzen, inwieweit die Versuchsteilnehmer bei der Beurteilung der Wahrheit von Aussagen auf Vorwissen zurückgreifen konnten. Hierzu wurde jede einzelne im Experiment gezeigte Aussage (z.B. „Marie Antoinette wurde in Österreich geboren.") zu einer Frage umformuliert („Wo wurde Marie Antoinette geboren?"), die den Versuchsteilnehmern dann jeweils mit der richtigen Lösung („Österreich"), einem Distraktor („Ungarn") sowie einer „ich weiß nicht"-Antwortalternative vorgegeben wurde.

Bei einer varianzanalytischen Auswertung der Daten, bei der der Prozentsatz an „wahr"-Urteilen (*proportion of „true"-judgments*, PTJ) als abhängige Variable diente, fiel die Wahrheitsillusion in der Warnungsgruppe erwartungsgemäß signifikant kleiner aus als in der Kontrollgruppe ($d_z = 0.27$ vs. $d_z = 0.57$). Die Wahrheitsillusion war jedoch auch in der Gruppe der gewarnten Versuchsteilnehmer noch statistisch nachweisbar. Wie bei Nadarevic und Aßfalg (2017) konnte die Wahrheitsillusion also auch in der vorliegenden Arbeit mithilfe von Warnungen zwar reduziert, jedoch nicht vollständig eliminiert werden. Erwähnenswert ist darüber hinaus, dass die Reduktion der Wahrheitsillusion für die verschiedenen Aussagetypen (wahr & leicht, wahr & schwer, falsch & leicht, falsch & schwer) unterschiedlich stark ausfiel. So konnte die Wahrheitsillusion für leichte, falsche Aussagen deskriptiv am stärksten reduziert werden, während keine nennenswerte Reduktion für schwere, falsche Aussagen zu beobachten war. Bei den wahren Aussagen hingegen fiel die Reduktion der Wahrheitsillusion für schwere Aussagen deskriptiv etwas stärker aus als für leichte Aussagen.

Um die Wirkung der Warnung auf Prozessebene zu untersuchen, wurden die Daten mithilfe eines multinomialen Verarbeitungsbaummodells analysiert. Verwendet wurde dazu eine sparsamere Variante des ursprünglich von Fazio et al. (2015) vorgeschlagenen *„fluency-conditional"*-Modells. Die eingesetzte Modellvariante unterscheidet drei übergeordnete, latente Parameter. Dem Modell zufolge verlassen sich Versuchsteilnehmer bei der Beurteilung der Wahrheit von Aussagen entweder (1) auf das metakognitive Gefühl der Verarbeitungsflüssigkeit, (2) auf ihr Vorwissen, oder (3) raten einfach, ob eine Aussage wahr oder falsch ist. Wird eine Aussage wiederholt präsentiert, ist es möglich, dass diese mit der Wahrscheinlichkeit *f* als „wahr" beurteilt wird, weil die Verarbeitung dieser Aussage aufgrund ihrer wiederholten Darbietung besonders leicht bzw. flüssig abläuft. Ist keine Verarbeitungsflüssigkeit gegeben oder wird diese nicht genutzt, greift ein Versuchsteilnehmer bei der Beurteilung der Aussage mit der Wahrscheinlichkeit *k* auf Vorwissen zurück, was zu einer korrekten Beurteilung des Wahrheitsstatus der Aussage führt. Liegt weder Verarbeitungsflüssigkeit noch Wissen vor, wird ein Versuchsteilnehmer raten, ob die präsentierte Aussage wahr oder falsch ist. Der Versuchsteilnehmer entscheidet sich dann mit der Wahrscheinlichkeit *g* für eine „wahr"-Antwort und mit der Gegenwahrscheinlichkeit (1 - *g*) für eine „falsch"-Antwort. Soll der Wahrheitsstatus einer erstmalig präsentierten Aussage beurteilt werden, kann keine Verarbeitungsflüssigkeit aufgrund einer früheren Präsentation vorliegen. Aus diesem Grund entfällt die entsprechende Gabelung im Verarbeitungsbaummodell. Wie auch bei der Beurteilung einer wiederholten Aussage wird jedoch angenommen, dass ein Versuchsteilnehmer den tatsächlichen Wahrheitsstatus einer erstmalig präsentierten Aussage mit der Wahrscheinlichkeit *k* auf der Basis seines Vorwissens korrekt beurteilen kann. Ist kein relevantes Vorwissen vorhanden, rät der Versuchsteilnehmer und entscheidet sich dabei mit der Wahrscheinlichkeit *g* für eine „wahr"-Antwort und mit der Gegenwahrscheinlichkeit (1 - *g*) für eine „falsch"-Antwort.

Bei der Auswertung der Daten wurde der Wissensparameter *k* separat für leichte und schwere Aussagen geschätzt. Sämtliche Parameter im Modell wurden darüber hinaus getrennt für die Kontroll- und die Warnungsgruppe ermittelt. Das verwendete Modell vermochte die Daten gut zu beschreiben, $G^2(8) = 10.78$, $p = .215$. Weitere Modelltests ergaben, dass sich die Warnung weder signifikant auf das Rateverhalten der Versuchsteilnehmer, noch auf die Wahrscheinlichkeit des Abrufs von Vorwissen auswirkte. Stattdessen resultierte die Warnung in einer signifikanten Reduktion der Wahrscheinlichkeit, mit der die Versuchsteilnehmer ihre Urteile auf der Basis von Verarbeitungsleichtigkeit trafen. Dieses Ergebnismuster ist kompatibel mit der Überlegung, dass die Flüssigkeit der Verarbeitung als Informationsquelle bei der Beurteilung der Wahrheit von Aussagen ungenutzt bleiben kann, wenn ein Urteiler erkennt, dass die Verarbeitungsflüssigkeit auf eine Quelle zurückgeht, die für das eigentliche Urteil irrelevant ist (Alter & Oppenheimer, 2009; Schwarz, 2004).

4 Diskussion

Menschen schreiben wiederholt präsentierten Aussagen typischerweise einen höheren Wahrheitsgehalt zu als Aussagen, denen sie nicht zuvor begegnet sind. Diese sogenannte Wahrheitsillusion gilt als robustes Phänomen, das zum Beispiel für Allgemeinwissensaussagen, Nachrichtenschlagzeilen, Werbebotschaften und Meinungsaussagen beobachtet werden kann und in verschiedenen Situationen zuverlässig auftritt (Dechêne et al., 2010). Auch in der vorliegenden Arbeit wird die Robustheit der Wahrheitsillusion in gleich mehrerlei Hinsicht belegt: So profitierte die Glaubwürdigkeit von Aussagen nicht nur in einer heterogenen Stichprobe von Freiwilligen während eines Online-Experimentes von einer wiederholten Darbietung (Experiment 1), sondern auch in zwei Experimenten mit studentischen Versuchsteilnehmern im Vorlesungs- (Experiment 2) bzw. Laborkontext (Experiment 3). Sogar eine explizite Warnung vor der Wahrheitsillusion und die Aussicht auf eine finanzielle Belohnung für möglichst unverfälschte, akkurate Wahrheitsurteile konnten die Wahrheitsillusion zwar reduzieren, jedoch nicht vollständig verhindern (Experiment 3).

Neben diesen Belegen für die Robustheit der Wahrheitsillusion wurde in der vorliegenden Arbeit erstmals ihre Test-Retest-Stabilität bestimmt. Diese fiel in zwei Experimenten (Experiment 1 & 2) für verschiedene Operationalisierungen des Phänomens ausgesprochen gering aus. Versuchsteilnehmer, die zu einem ersten Testzeitpunkt eine über- bzw. unterdurchschnittlich starke Wahrheitsillusion aufwiesen, gehörten also zu einem späteren Testzeitpunkt nicht zwangsläufig erneut zu denjenigen Personen, für die eine über- bzw. unterdurchschnittlich starke Wahrheitsillusion beobachtet werden konnte. Die in der vorliegenden Arbeit ermittelten, sehr geringen Test-Retest-Korrelationen legen nahe, dass es sich bei der Wahrheitsillusion entweder nicht um ein individuell stabiles Phänomen handelt, oder aber dass die Reliabilität der Messung der Wahrheitsillusion in einer typischen Versuchsanordnung vergleichsweise gering ausfällt (vgl. Bortz & Döring, 2006; Crocker & Algina, 1986). Beide Möglichkeiten stehen dem in früheren Studien ausgerufenen Ziel entgegen, interindividuelle Unterschiede in der Stärke der Wahrheitsillusion bzw. deren Korrelate näher zu erforschen (Arkes et al., 1991; Dechêne et al., 2010). Angesichts der sehr geringen Test-Retest-Stabilität der Wahrheitsillusion ist deshalb zu befürchten, dass sich gelegentlich beobachtete Zusammenhänge zwischen der Stärke der Wahrheitsillusion und kognitiven Traits oder Persönlichkeitsmerkmalen in zukünftigen Experimenten als nicht replizierbar erweisen könnten (vgl. Michalkiewicz & Erdfelder, 2016). Ein solcher Fall wurde kürzlich in der Studie von De keersmaecker et al. (2018) beschrieben. Personen mit einem stark ausgeprägten experientiell-intuitiven Denkstil zeigten zwar in einem ersten Experiment eine stärkere Wahrheitsillusion als Personen mit einer geringen Ausprägung, dieses Ergebnis ließ sich jedoch in einem zweiten, teststärkeren Experiment nicht replizieren. Derartige Befunde sind angesichts der in dieser Arbeit vorgestellten Ergebnisse nicht überraschend.

Die Ergebnisse von Experiment 1 und 2 betonen die Wichtigkeit einer getrennten Betrachtung der *Robustheit* und der *Reliabilität* eines Phänomens sowie die Notwendigkeit einer trennscharfen Verwendung der jeweiligen Begrifflichkeiten (vgl. auch Hedge, Powell & Sumner, 2018; Pohl, 1999). So wiesen Hedge et al. (2018) jüngst darauf hin, dass ein Effekt keinesfalls reliabel messbar sein muss, nur weil er zuverlässig auftritt und dementsprechend als „robust“ gelten darf. Hedge et al. (2018) argumentierten sogar, dass robuste experimentelle Effekte typischerweise mit einer geringen Zwischensubjektvarianz einhergehen, eine geringe Zwischensubjektvarianz aber unweigerlich zu einer geringen Reliabilität führt und dementsprechend die Untersuchung eines Paradigmas oder Effekts aus einer differentialpsychologischen Sicht erschwert. Dass eine mangelhafte Reliabilität mit weiteren Problemen verbunden ist und sogar zu irreführenden Interpretationen von Forschungsbefunden führen kann, haben Studien in anderen Gebieten bereits wiederholt demonstriert. So wurde beispielsweise gezeigt, dass beobachtete einfache Dissoziationen zwischen impliziten und expliziten Gedächtnismaßen nicht unbedingt funktionelle Dissoziationen der mutmaßlich zugrundeliegenden Gedächtnissysteme widerspiegeln, sondern möglicherweise (auch) auf die unterschiedlich hohe Reliabilität der eingesetzten Maße zurückgeführt werden können (vgl. z.B. Buchner & Brandt, 2003; Buchner & Wippich, 2000; Meier & Perrig, 2000).

Die Frage nach der Reliabilität der Messungen ist in bisherigen Studien zur Wahrheitsillusion vollständig vernachlässigt worden. In der vorliegenden Arbeit wurde deshalb erstmals die Test-Retest-Stabilität der Wahrheitsillusion für verschiedene Operationalisierungen des Phänomens untersucht. Zukünftige Studien sollten sich der Frage widmen, ob und wie die Wahrheitsillusion auf individueller Ebene reliabel gemessen werden kann. Neben den in der vorliegenden Arbeit eingesetzten Operationalisierungen der Wahrheitsillusion sollte zukünftig auch die Reliabilität bzw. Stabilität für alternative Operationalisierungsmöglichkeiten untersucht werden. Dabei könnte sich der Einsatz von hierarchischen multinomialen Verarbeitungsbaummodellen (vgl. z.B. Smith & Batchelder, 2010) als nützlich erweisen. Abweichend von klassischen multinomialen Verarbeitungsbaummodellen, die die Auftretenswahrscheinlichkeit verschiedener latenter kognitiver Prozesse typischerweise auf Gruppenebene schätzen, sind hierarchische Verarbeitungsbaummodelle explizit für die Untersuchung interindividueller Unterschiede vorgesehen, da sie separate Parameterschätzungen für einzelne Personen liefern. Hierarchische Verarbeitungsbaummodelle wurden bereits erfolgreich bei der Bestimmung der Stabilität der Rekognitionsheuristiknutzung eingesetzt (Michalkiewicz & Erdfelder, 2016), verlangen jedoch – wie klassische multinomiale Verarbeitungsbaummodelle auch – das Vorliegen von kategorialen Daten.

Neben der Stabilität der Wahrheitsillusion beschäftigte sich die vorliegende Arbeit auch mit den Effekten von Warnungen auf die Wahrheitsillusion sowie auf diejenigen kognitiven Prozesse, die bei der Beurteilung der Wahrheit von Aussagen mutmaßlich eine Rolle spielen. Die vorliegende Arbeit reiht sich damit in die bislang noch kleine Zahl von Studien ein, die sich mit der Vermeidung bzw. Verhinderung der Wahrheitsillusion befasst haben. Derartige Studien sind von großer Bedeutung, da die Wahrheitsillusion nicht nur unter Laborbedingungen auftritt, sondern auch dann beobachtet werden kann, wenn Personen in ihrem natürlichen Umfeld bestimmten Aussagen wiederholt ausgesetzt sind (Boehm, 1994) oder Zufallsstichproben aus der Allgemeinbevölkerung ein klassisches Wahrheitsillusionsparadigma durchlaufen (Gigerenzer, 1984). Die Wahrheitsillusion kann also auch im Alltag nachteilige Konsequenzen haben, etwa wenn Personen in sozialen Netzwerken oder auf Nachrichtenportalen wiederholt falschen Informationen ausgesetzt sind (vgl. Pennycook et al., 2018). Bisherige Studien zeigten, dass die Wahrheitsillusion vergleichsweise widerstandsfähig gegenüber gezielten Versuchen ist, sie zu eliminieren. Eine „Accountability“-Manipulation etwa konnte die Wahrheitsillusion nicht bedeutsam reduzieren (Boehm, 1994); direktes Feedback zum Wahrheitsstatus der präsentierten Aussagen eignete sich zwar als kurzfristig wirksame, jedoch nicht unbedingt als langfristig erfolgreiche Strategie zur Bekämpfung der Wahrheitsillusion (Brown & Nix, 1996; Skurnik et al., 2005; vgl. aber Swire et al., 2017).

In einer kürzlich durchgeführten, weiteren Debiasing-Studie konnten Nadarevic und Aßfalg (2017) die Wahrheitsillusion durch eine explizite Warnung signifikant reduzieren. Sie folgerten daraus, dass Menschen ein gewisses Maß an Kontrolle über die Attribution von metakognitiven Gefühlen der Verarbeitungsflüssigkeit auf die wahrgenommene Glaubwürdigkeit einer präsentierten Aussage ausüben können. Die vorliegende Arbeit untersuchte erstmals mithilfe eines multinomialen Modells, wie sich eine Warnung vor der Wahrheitsillusion auf die Nutzung von Verarbeitungsflüssigkeit, den Abruf von Vorwissen, sowie das Rateverhalten von Versuchsteilnehmern auswirkt. Dabei zeigte sich kein Einfluss der Manipulation auf die beiden letztgenannten Prozesse. Die Warnung reduzierte jedoch signifikant die Wahrscheinlichkeit, mit der sich die Versuchsteilnehmer bei der

Beurteilung der Wahrheit einer wiederholten Aussage auf metakognitive Gefühle der Verarbeitungsflüssigkeit verließen. Dieser Befund ist vereinbar mit der Annahme, dass Verarbeitungsflüssigkeit als Hinweisreiz zur Beurteilung der Wahrheit einer Aussage unter bestimmten Umständen ungenutzt bleiben kann, etwa wenn ihr Informationswert infrage gestellt wird (Alter & Oppenheimer, 2009; Schwarz, 2004). Es ist jedoch zu beachten, dass die in der vorliegenden Arbeit eingesetzte Warnung zwar zu einer Reduktion der Wahrheitsillusion führte, diese jedoch nicht vollständig eliminieren konnte. Dies repliziert den Befund von Nadarevic und Aßfalg (2017). Da den Versuchsteilnehmern in Experiment 3 der vorliegenden Arbeit jedoch erstmals – und damit abweichend von Nadarevic und Aßfalg (2017) – eine finanzielle Belohnung für möglichst akkurate Wahrheitsurteile in Aussicht gestellt worden war, die Teilnehmer also mutmaßlich ausreichend motiviert waren, liegt die Schlussfolgerung nahe, dass eine unzureichende Motivation offenbar nicht dafür verantwortlich war, dass die Wahrheitsillusion in den Experimenten von Nadarevic und Aßfalg (2017) trotz Warnungen nicht vollständig eliminiert werden konnte.

Angesichts der Ergebnisse von Nadarevic und Aßfalg (2017) sowie der Befunde der vorliegenden Arbeit erscheinen explizite Warnungen als eine potenziell nützliche Strategie, um die Stärke der Wahrheitsillusion wenigstens zu reduzieren. Weitere Studien sollten sich jedoch mit der Frage beschäftigen, ob Warnungen auch dann einen Einfluss auf die Wahrheitsillusion ausüben können, wenn die erste Begegnung mit später wiederholten Aussagen länger zurückliegt. Erste Hinweise darauf, dass dies möglicherweise der Fall ist, erbrachte die Studie von Nadarevic und Aßfalg (2017); weitere empirische Arbeiten zu dieser Frage sind jedoch vonnöten.

Die Tatsache, dass es in bisherigen Experimenten bislang nicht gelungen ist, die Wahrheitsillusion durch Warnungen vollständig zu eliminieren, belegt die verblüffende Robustheit und Widerstandsfähigkeit der Wahrheitsillusion. Für die Praxis und insbesondere für den Kampf gegen „Fake News“ (Lazer et al., 2018) bedeutet dies, dass es vielversprechender sein könnte, die Vorzüge der Wahrheitsillusion zu nutzen anstatt – möglicherweise vergebens – zu versuchen, den Einfluss der Wiederholung auf die beurteilte Glaubwürdigkeit von falschen Aussagen zu unterbinden. Um Falschnachrichten zu entkräften, ist es daher möglicherweise besser, Richtigstellungen wiederholt und auf möglichst leicht verarbeitbare Weise zu präsentieren (Cook & Lewandowsky, 2011) und die zu entkräftende Falschnachricht dabei nicht häufiger zu wiederholen als unbedingt nötig (Ecker, Hogan & Lewandowsky, 2017; Swire et al., 2017).

Literaturverzeichnis

Allport, F. H. & Lepkin, M. (1945). Wartime rumors of waste and special privilege: Why some people believe them. *Journal of Abnormal and Social Psychology, 40*, 3–36. doi:10.1037/h0058110

Alter, A. L. & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review, 13*, 219–235. doi:10.1177/1088868309341564

Arkes, H. R., Boehm, L. E. & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology, 27*, 576–605. doi:10.1016/0022-1031(91)90026-3

Arkes, H. R., Hackett, C. & Boehm, L. (1989). The generality of the relation between familiarity and judged validity. *Journal of Behavioral Decision Making, 2*, 81–94. doi:10.1002/bdm.3960020203

Bacon, F. T. (1979). Credibility of repeated statements: Memory for trivia. *Journal of Experimental Psychology: Human Learning and Memory, 5*, 241–252. doi:10.1037/0278-7393.5.3.241

Batchelder, W. H. & Riefer, D. M. (1990). Multinomial processing models of source monitoring. *Psychological Review, 97*, 548–564. doi:10.1037/0033-295X.97.4.548

Bernstein, D. M., Wilson, A. M., Pernat, N. L. M. & Meilleur, L. R. (2012). Auditory hindsight bias. *Psychonomic Bulletin & Review, 19*, 588–593. doi:10.3758/s13423-012-0268-0

Blank, H. & Launay, C. (2014). How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *Journal of Applied Research in Memory and Cognition, 3*, 77–88. doi:10.1016/j.jarmac.2014.03.005

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F. & Schwarz, N. (1994). Need for Cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben. *Zeitschrift für Sozialpsychologie, 25*, 147–154.

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin, 20*, 285–293. doi:10.1177/0146167294203006

Bortz, J. & Döring, N. (2006). *Forschungsmethoden und Evaluation für Human- und Sozialwissenschaftler* (4., überarbeitete Aufl.). Heidelberg: Springer.

Brown, A. S. & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1088–1100. doi:10.1037/0278-7393.22.5.1088

Buchner, A. & Brandt, M. (2003). Further evidence for systematic reliability differences between explicit and implicit memory tests. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *56*, 193–209. doi:10.1080/02724980244000260

Buchner, A. & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*, 227–259. doi:10.1006/cogp.1999.0731

Cacioppo, J. T. & Petty, R. E. (1982). The need for cognition. *Journal of Personality and Social Psychology*, *42*, 116–131. doi:10.1037//0022-3514.42.1.116

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment*, *20*, 166–171. doi:10.1027/1015-5759.20.3.166

Clarkson, P. M., Emby, C. & Watt, V. W.-S. (2002). Debiasing the outcome effect: The role of instructions in an audit litigation setting. *Auditing: A Journal of Practice & Theory*, *21*(2), 7–20. doi:10.2308/aud.2002.21.2.7

Cook, J. & Lewandowsky, S. (2011). *The Debunking Handbook*. Zugriff am 22. Januar 2019 unter https://skepticalscience.com/docs/Debunking_Handbook.pdf

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston.

Cronbach, L. J. & Furby, L. (1970). How we should measure "change"—Or should we? *Psychological Bulletin*, *74*, 68–80. doi:10.1037/h0029382

Davies, M. F. (1993). Field-dependence and hindsight bias: Output interference in the generation of reasons. *Journal of Research in Personality*, *27*, 222–237. doi:10.1006/jrpe.1993.1016

De keersmaecker, J., Roets, A., Pennycook, G. & Rand, D. G. (2018). *Is the illusory truth effect robust to individual differences in cognitive ability, need for cognitive closure, and cognitive style?* Manuskript in Vorbereitung.

Dechêne, A., Stahl, C., Hansen, J. & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*, 238–257. doi:10.1177/1088868309352251

Diedenhofen, B. & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4), e0121945. doi:10.1371/journal.pone.0121945

DiFonzo, N., Beckstead, J. W., Stupak, N. & Walders, K. (2016). Validity judgments of rumors heard multiple times: The shape of the truth effect. *Social Influence*, *11*, 22–39. doi:10.1080/15534510.2015.1137224

Ecker, U. K. H., Hogan, J. L. & Lewandowsky, S. (2017). Reminders and repetition of misinformation: Helping or hindering its retraction? *Journal of Applied Research in Memory and Cognition*, *6*, 185–192. doi:10.1016/j.jarmac.2017.01.014

Erdfelder, E., Cüpper, L. & Auer, T.-S. (2006). Multinomiale Verarbeitungsbaummodelle. In J. Funke & P. Frensch (Hrsg.), *Handbuch der Allgemeinen Psychologie – Kognition* (S. 760–768). Göttingen: Hogrefe.

Fazio, L. K., Brashier, N. M., Payne, B. K. & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002. doi:10.1037/xge0000098

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19*(4), 25–42. doi:10.1257/089533005775196732

Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology*, *97*, 185–195. doi:10.2307/1422594

Grenier, J. H., Peecher, M. E. & Piercey, M. D. (2007). *Judging auditor negligence: Debiasing interventions, outcome bias, and reverse outcome bias*. Zugriff am 12. Oktober 2018 unter https://ssrn.com/abstract=1015523

Harley, E. M., Carlsen, K. A. & Loftus, G. R. (2004). The "saw-it-all-along" effect: Demonstrations of visual hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 960–968. doi:10.1037/0278-7393.30.5.960

Hasher, L., Goldstein, D. & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112. doi:10.1016/S0022-5371(77)80012-1

Hawkins, S. A. & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, *19*, 212–225. doi:10.1086/209297

Hedge, C., Powell, G. & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. doi:10.3758/s13428-017-0935-1

Johar, G. V. & Roggeveen, A. L. (2007). Changing false beliefs from repeated advertising: The role of claim-refutation alignment. *Journal of Consumer Psychology*, *17*, 118–127. doi:10.1016/S1057-7408(07)70018-9

Kantner, J. & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*, 1163–1177. doi:10.3758/s13421-012-0226-0

Kim, C. (2001). *The role of individual differences in general skepticism in the illusory truth effect*. Unveröffentlichte Dissertation, University of Cincinnati.

Kirby, K. N. (2009). One-year temporal stability of delay-discount rates. *Psychonomic Bulletin & Review*, *16*, 457–462. doi:10.3758/PBR.16.3.457

Kruglanski, A. W. & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review*, *103*, 263–283. doi:10.1037//0033-295X.103.2.263

Ladowsky-Brooks, R. L. (2010). The truth effect in relation to neuropsychological functioning in traumatic brain injury. *Brain Injury*, *24*, 1343–1349. doi:10.3109/02699052.2010.506856

Law, S., Hawkins, S. A. & Craik, F. I. M. (1998). Repetition-induced belief in the elderly: Rehabilitating age-related memory deficits. *Journal of Consumer Research*, *25*, 91–107. doi:10.1086/209529

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science*, *359*, 1094–1096. doi:10.1126/science.aao2998

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*, 361–366. doi:10.1101/lm.94705

Loftus, E. F., Miller, D. G. & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19–31. doi:10.1037/0278-7393.4.1.19

Maio, G. R. & Esses, V. M. (2001). The need for affect: Individual differences in the motivation to approach or avoid emotions. *Journal of Personality*, *69*, 583–615. doi:10.1111/1467-6494.694156

McGlone, M. S. & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. *Psychological Science*, *11*, 424–428. doi:10.1111/1467-9280.00282

Meier, B. & Perrig, W. J. (2000). Low reliability of perceptual priming: Consequences for the interpretation of functional dissociations between explicit and implicit memory. *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *53*, 211–233. doi:10.1080/713755878

Michalkiewicz, M. & Erdfelder, E. (2016). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition*, *44*, 454–468. doi:10.3758/s13421-015-0567-6

Moritz, S., Köther, U., Woodward, T. S., Veckenstedt, R., Dechêne, A. & Stahl, C. (2012). Repetition is good? An internet trial on the illusory truth effect in schizophrenia and nonclinical participants. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*, 1058–1063. doi:10.1016/j.jbtep.2012.04.004

Mutter, S. A., Lindsey, S. E. & Pliske, R. M. (1995). Aging and credibility judgment. *Aging and Cognition*, *2*, 89–107. doi:10.1080/13825589508256590

Nadarevic, L. (2010). *Die Wahrheitsillusion*. Berlin: Köster.

Nadarevic, L. & Aßfalg, A. (2017). Unveiling the truth: Warnings reduce the repetition-based truth effect. *Psychological Research*, *81*, 814–826. doi:10.1007/s00426-016-0777-y

Nadarevic, L. & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, *23*, 74–84. doi:10.1016/j.concog.2013.12.002

Oeberst, A. & Blank, H. (2012). Undoing suggestive influence on memory: The reversibility of the eyewitness misinformation effect. *Cognition*, *125*, 141–159. doi:10.1016/j.cognition.2012.07.009

Pacini, R. & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*, 972–987. doi:10.1037//0022-3514.76.6.972

Pennycook, G., Cannon, T. D. & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*, 1865–1880. doi:10.1037/xge0000465

Pohl, R. F. (1999). *Hindsight bias: Robust, but not reliable*. Unveröffentlichtes Manuskript, Justus-Liebig-Universität, Gießen, Deutschland.

Pohl, R. F. & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes*, *67*, 49–58. doi:10.1006/obhd.1996.0064

Raghunathan, T. E., Rosenthal, R. & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*, 178–183. doi:10.1037/1082-989X.1.2.178

Reber, R. & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338–342. doi:10.1006/ccog.1999.0386

Roggeveen, A. L. & Johar, G. V. (2002). Perceived source variability versus familiarity: Testing competing explanations for the truth effect. *Journal of Consumer Psychology*, *12*, 81–91. doi:10.1207/S15327663JCP1202_02

Schlink, S. & Walther, E. (2007). Kurz und gut: Eine deutsche Kurzskala zur Erfassung des Bedürfnisses nach kognitiver Geschlossenheit. *Zeitschrift für Sozialpsychologie*, *38*, 153–161. doi:10.1024/0044-3514.38.3.153

Schwartz, M. (1982). Repetition and rated truth value of statements. *American Journal of Psychology*, *95*, 393–407. doi:10.2307/1422132

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*, 332–348. doi:10.1207/s15327663jcp1404_2

Sharpe, D. & Adair, J. G. (1993). Reversibility of the hindsight bias: Manipulation of experimental demands. *Organizational Behavior and Human Decision Processes*, *56*, 233–245. doi:10.1006/obhd.1993.1053

Silva, R. R., Garcia-Marques, T. & Mello, J. (2016). The differential effects of fluency due to repetition and fluency due to color contrast on judgments of truth. *Psychological Research*, *80*, 821–837. doi:10.1007/s00426-015-0692-7

Skurnik, I., Yoon, C., Park, D. C. & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*, 713–724. doi:10.1086/426605

Smith, J. B. & Batchelder, W. H. (2010). Beta-MPT: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. doi:10.1016/j.jmp.2009.06.007

Stanislaw, H. & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, & Computers*, *31*, 137–149. doi:10.3758/BF03207704

Sundar, A., Kardes, F. R. & Wright, S. A. (2015). The influence of repetitive health messages and sensitivity to fluency on the truth effect in advertising. *Journal of Advertising*, *44*, 375–387. doi:10.1080/00913367.2015.1045154

Swire, B., Ecker, U. K. H. & Lewandowsky, S. (2017). The role of familiarity in correcting inaccurate information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*, 1948–1961. doi:10.1037/xlm0000422

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 219–230. doi:10.1037/0278-7393.33.1.219

Unkelbach, C. & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22–38. doi:10.1016/j.concog.2008.09.006

Webster, D. M. & Kruglanski, A. W. (1994). Individual differences in need for cognitive closure. *Journal of Personality and Social Psychology*, *67*, 1049–1062. doi:10.1037/0022-3514.67.6.1049

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*, 345–422. doi:10.2307/1167368

Williams, R. H. & Zimmerman, D. W. (1983). The comparative reliability of simple and residualized difference scores. *Journal of Experimental Education*, *51*, 94–97. doi:10.1080/00220973.1982.11011846

Williams, R. H., Zimmerman, D. W. & Mazzagatti, R. D. (1987). Large sample estimates of the reliability of simple, residualized, and base-free gain scores. *Journal of Experimental Education*, *55*, 116–118. doi:10.1080/00220973.1987.10806443

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

Wilson, T. D., Centerbar, D. B. & Brekke, N. (2002). Mental contamination and the debiasing problem. In T. Gilovich, D. Griffin & D. Kahneman (Hrsg.), *Heuristics and biases: The psychology of intuitive judgment* (S. 185–200). New York, NY: Cambridge University Press.

Zimmerman, D. W. & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154. doi:10.1111/j.1745-3984.1982.tb00124.x

Zumbo, B. D. (1992). The comparative reliability of simple and residualized difference scores: A corrigendum. *Journal of Experimental Education*, *61*, 81–83. doi:10.1080/00220973.1992.9943852

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Hrsg.), *Advances in social science methodology* (Bd. 5, S. 269–304). Greenwich, CT: JAI Press.

Anhang: Einzelarbeiten

Manuskript 1:

Calio, F., Nadarevic, L., & Musch, J. (2019). Is the truth effect an individually stable phenomenon? An assessment of its test-retest stability. *Manuscript submitted for publication*.

Die im o.g. Manuskript beschriebenen Experimente wurden unter meiner Federführung gemeinsam geplant. Das in beiden Studien verwendete Aussagenmaterial wurde von mir erstellt und in einer Voruntersuchung normiert. Ich war für die Umsetzung des Web-Experimentes (Experiment 1) über das Befragungstool Unipark sowie für die Erstellung der in Experiment 2 eingesetzten Testhefte verantwortlich. Die Daten beider Experimente wurden ebenfalls von mir ausgewertet. Das Manuskript wurde von mir verfasst und wiederholt von Jochen Musch und mir überarbeitet; Anmerkungen von Lena Nadarevic flossen ebenfalls mit ein.

Manuskript 2:

Calio, F., Nadarevic, L., & Musch, J. (2019). Debiasing the truth effect: Exploring the effects of warnings on judgments of truth. *Manuscript submitted for publication*.

Die im o.g. Manuskript beschriebenen Experimente wurden unter meiner Federführung gemeinsam geplant. Lena Nadarevic stellte das im Experiment eingesetzte Aussagenmaterial zur Verfügung. Ich war für die Umsetzung des computergestützten Laborexperimentes über das Befragungstool Unipark sowie für die Datenerhebung und -auswertung verantwortlich. Das Manuskript wurde von mir verfasst und wiederholt von Jochen Musch und mir überarbeitet; Anmerkungen von Lena Nadarevic flossen ebenfalls mit ein.

Is the truth effect an individually stable phenomenon? An assessment of its test-retest stability

Frank Calio[1], Lena Nadarevic[2], & Jochen Musch[1]

[1] University of Duesseldorf

[2] University of Mannheim

Author note

Frank Calio and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf, Germany; Lena Nadarevic, Department of Psychology, School of Social Sciences, University of Mannheim, Germany.

Correspondence concerning this article should be addressed to Frank Calio, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany. E-mail: frank.calio@uni-duesseldorf.de

Abstract

The finding that repeating a statement typically increases its perceived validity is referred to as the *truth effect*. There is as yet little research on individual differences in the magnitude of the effect and its correlates, and it has yielded rather mixed results. However, any search for replicable correlations between the truth effect and other cognitive or personality variables is bound to fail if the truth effect is not a stable phenomenon and cannot be measured reliably on an individual level. We therefore conducted two experiments investigating the test-retest stability of the truth effect. To operationalize the magnitude of the effect, Experiment 1 employed the between-items criterion and Experiment 2 employed the within-items criterion (Dechêne, Stahl, Hansen, & Wänke, 2010). In both experiments, the truth effect's test-retest stability was very low. These findings are cause for concern regarding the usefulness of established indices of the truth effect for personality and individual difference research. The results are discussed in the context of related research on the reliability of measures used in other experimental paradigms.

Keywords: Truth effect, Illusory truth, Validity rating, Individual differences, Test-retest stability

Word count: 173

Is the truth effect an individually stable phenomenon? An assessment of its test-retest stability

A large body of research accumulated over the past 40 years has shown that repeating a statement typically increases its perceived validity. This phenomenon, which is now referred to as the *truth effect*, was first described by Hasher, Goldstein, and Toppino (1977) and has been observed for a wide variety of stimuli, including correct and false trivia statements, opinion statements, and product-related claims (for a review, see Dechêne, Stahl, Hansen, & Wänke, 2010). The most widely accepted explanation of the truth effect is based on *processing fluency*, defined as the "metacognitive experience of ease during information processing" (Dechêne et al., 2010, p. 240). According to this account, the truth effect occurs because repeated statements are processed more fluently than novel stimuli and feelings of processing ease are typically used as a cue for validity in judgments of truth (Unkelbach, 2007). Empirical support for the role of processing fluency in truth judgments was provided by Reber and Schwarz (1999), who manipulated processing fluency by varying the color contrast of statements that were presented to a group of participants. Statements which were easier to read and process due to better color contrast were more likely to be judged as true. Unkelbach (2007) argued that people rely on processing fluency when judging the truth of a statement because fluency is an ecologically valid cue for truth. Recently, Unkelbach and Rom (2017) reconsidered the role of processing fluency in the truth effect and proposed a referential theory of the repetition-induced truth effect. It assumes that people evaluate the truth of a statement based on the number and the coherence of corresponding references in memory. Specifically, the theory posits that presenting a statement like "Manfred von Richthofen, also known as the 'Red Baron', was a German fighter pilot during World War I" either activates corresponding references (e.g., "Red Baron", "World War I") and their links within a localized network, or prompts the construction of new references and links if the statement contains new information. A statement that has been presented before therefore has more corresponding and coherently linked references than a new statement. This results in more fluent processing of the statement, while also increasing the probability that it is judged as true.

Although the truth effect has been studied intensively and is considered to be a robust phenomenon occurring under a large variety of conditions (Dechêne et al., 2010), there are still important open questions. In particular, it has been noted that there is a dearth of research on individual differences in the magnitude of the truth effect and its correlates (Arkes, Boehm, & Xu, 1991; Dechêne et al., 2010). The rare exceptions include two studies by Arkes et al. (1991) and Boehm (1994), in which the magnitude of the truth effect was found not to be associated with participants' need for cognition, defined as "an individual's tendency to engage in and enjoy effortful cognitive endeavors" (Cacioppo, Petty, & Kao, 1984, p. 306). In an advertising context, Sundar, Kardes, and Wright (2015) observed a truth effect only for participants scoring high on the Need for Affect scale by Maio and Esses (2001), presumably because consumers who are sensitive to their feelings are more likely to pay attention to and rely on metacognitive cues such as processing ease when judging the truth of a product claim. In another study, DiFonzo, Beckstead, Stupak, and Walders (2016) reported a small effect of dispositional skepticism on the magnitude of the truth effect, indicating that participants high in skepticism were less influenced by repetition. De keersmaecker, Roets, Pennycook, and Rand (2018) investigated the relationship between the truth effect and several cognitive as well as personality variables. In this study, the truth effect was not moderated by participants' cognitive ability or need for cognitive closure, defined as the "desire for a firm answer to a question and an aversion toward ambiguity" (Kruglanski & Webster, 1996, p. 264). Moreover, the magnitude of the truth effect was unaffected by individual differences in the Cognitive Reflection Test, a test that assesses an individual's tendency to engage in "miserly information processing" (Toplak, West, & Stanovich, 2014, p. 147). Participants in the study by De keersmaecker et al. also completed the Rational-Experiential Inventory by Pacini and Epstein (1999).While the truth effect was found to be independent of participants' level of rational thinking, a moderating effect of experiential thinking was observed in one experiment, but did not replicate in a second study (De keersmaecker et al., 2018). Moreover, a meta-analysis by Dechêne et al. (2010) revealed that the truth effect is not moderated by participants' age.

Apart from these findings, some researchers have investigated the relationship between the magnitude of the truth effect and several neuropsychological and psychiatric variables. Moritz et al. (2012), for example, found that for emotional material, the truth effect was positively correlated with self-reports of positive symptoms typically present in people with schizophrenia. This correlation was observed for both healthy participants and participants with a probable diagnosis of schizophrenia. Ladowsky-Brooks (2010) examined the truth effect in a sample of individuals with traumatic brain injury. In this study, there were no consistent patterns of correlations between scores in several neuropsychological tests (e.g., memory tests, tests of executive functioning) and two different operationalizations of the truth effect's magnitude.

In summary, research on personality and cognitive correlates of the truth effect is still scarce. However, a successful search for cognitive or personality correlates of the truth effect requires one essential precondition to be met. Replicable relations with cognitive or personality traits can only be found for phenomena that are sufficiently reliable and stable on an individual level. Recently, Michalkiewicz and Erdfelder (2016) stressed the importance of this notion when investigating individual differences in the use of the recognition heuristic (Goldstein & Gigerenzer, 2002). They argued that before trying to explain variability in the use of the recognition heuristic in terms of personality variables or cognitive traits, it should be investigated whether use of the recognition heuristic is stable on an individual level in the first place, because no replicable relations with other personality variables or cognitive traits are to be expected if use of the recognition heuristic varies haphazardly within individuals across time and situations. In the same vein, no replicable correlation between the magnitude of the truth effect and any cognitive or personality trait is to be expected if the truth effect is not consistent over time.

A few studies have already investigated the stability of various constructs in the area of judgment and decision making. Kantner and Lindsay (2012, 2014), for example, found evidence for the temporal and cross-situational stability of response bias in recognition tasks. Michalkiewicz and Erdfelder (2016) demonstrated that the use of the recognition heuristic is stable across time and situations. Other studies (e.g., Beck & Triplett, 2009; Kirby, 2009) have reported evidence for the temporal stability of delay discounting, the effect that an outcome which is remote in time tends to have less value than a more immediate outcome. We sought to extend this research by investigating whether the truth effect is an individually stable phenomenon that correlates across two or more points in time. To this end, we assessed the test-retest stability of the truth effect in two experiments.

Experiment 1

The first study was conducted online and consisted of two sessions separated by approximately one week. Each session comprised three phases: In an initial exposure phase, participants were asked to assign a set of trivia statements to different categories of knowledge. Participants then completed a distractor task, and finally judged the validity of two sets of statements, one of which had already been presented during the initial category-sorting task.

Based on this experimental procedure, it was possible to assess the truth effect by comparing validity ratings between the set of repeated statements and the set of statements that participants had not encountered before. This comparison is known as the *between-items criterion* (Dechêne et al., 2010). Conducting two sessions allowed us to obtain *two* separate truth effect indices per participant. The test-retest stability of the truth effect was then computed as the correlation between these two indices.

Method

Participants. Participants were recruited by sending email invitations to members of a non-commercial online research panel. The panel consisted of volunteers who had previously indicated an interest in taking part in psychological research conducted by the University of Duesseldorf. Out of the 246 participants who started the study, 212 finished the first session. All participants who provided complete data in the first session were contacted again approximately one week after their participation and were invited to take part in the second session. We analyzed only the data of the 169 participants who also finished the second session. Two of these participants had to be excluded because they failed a seriousness check by indicating that they had not participated seriously in at least one of the two experimental sessions (Aust, Diedenhofen, Ullrich, & Musch, 2013). The data from one additional participant were discarded because he had completed the first session twice. Therefore, the final sample consisted of 166 participants, 90 of which were female. Participants' age ranged from 21 to 62 years (M = 40.43, SD = 12.45). All participants were native speakers of German.

Materials. To obtain normative data on the perceived validity of items in a pretest, we collected 404 trivia statements that could either be true or false and that covered different domains of knowledge, including geography, history, politics and sports. The validity of each statement was evaluated by at least 23 participants on a 6-point rating scale ranging from 1 (*definitely false*) to 6 (*definitely true*). Due to ceiling effects, a truth effect is not to be expected for statements that are obviously true from the outset. Therefore, following common practice in research on the truth effect, we selected 80 statements (40 true, 40 false) that were judged to be ambiguous with regard to their truth status for the main study (e.g., "The actor Keanu Reeves was born in Lebanon"). Validity ratings for these statements varied between M = 3.04 and 4.00, with standard deviations ranging from 0.84 to 1.90. We assigned these statements to four stimulus sets A, B, C, and D so that each set contained 10 false and 10 true statements. Special care was taken to match all item sets with regard to their mean perceived validity (M_A = 3.55, M_B = 3.55, M_C = 3.55, M_D = 3.57) and their mean standard deviations (SD_A = 1.36, SD_B = 1.37, SD_C = 1.36, SD_D = 1.36).

Procedure. Participants took part in two sessions of 20 minutes each that were separated by approximately one week. Both sessions consisted of an initial category-sorting task, a filler task and a validity-rating task. Within each session, half of the statements in the final validity-rating task had previously been presented in the category-sorting task, whereas the other half of the statements had not been shown before. The truth status of the statements was manipulated orthogonally to their repetition status; half of the statements were true, whereas the other half of the statements were false.

At the beginning of Session 1, participants answered some basic demographic questions. Afterwards, they were informed that they were going to see a collection of statements that could either be true or false, and that it would be their task to assign each statement to one of six different categories of knowledge (geography, flora & fauna, politics & history, science, entertainment, other). Participants were then presented with 32 statements. The first six and the last six statements served as buffer items; in between, 10 false and 10 true statements from one of the four stimulus sets were presented in random order. After participants had assigned each statement to one of the six categories of knowledge, a nonverbal filler task followed that lasted for a fixed interval of 10 minutes. Participants were then introduced to their final task: They were informed that they would again see true and false statements and that this time, all statements should be rated for validity on a 6-point scale ranging from 1 (*definitely false*) to 6 (*definitely true*). Furthermore, participants were told that the statements would be selected randomly from a large collection of statements, and that due to this random sampling they might encounter some statements they were already familiar with. Participants then rated the validity of statements from two items sets, one of which had already been presented during the initial category-sorting task. The 2 x 20 = 40 statements were presented in random order. Finally, participants were asked to indicate whether they had participated seriously (Aust et al., 2013).

Invitations to the second session were dispatched individually about one week after each participant had completed the first part of the study. The distribution of the response interval was skewed because not every participant responded to this invitation immediately; the median time interval between the two sessions was 7.27 days. Importantly, however, the pattern of the main results reported below did not change when, for example, only the 127 participants who responded after a maximum of 10 days were included in the analysis.

The basic structure of the second session was identical to that of the first session: Participants first performed the category-sorting task for 10 true and 10 false statements from a new stimulus set. Like in Session 1, the statements were presented in random order and were enclosed by six primacy and six recency buffer items. Participants then worked on a non-verbal filler task for 10 minutes. Finally, participants rated the validity of statements from two item sets, one of which had already been presented during the category-sorting task. The 2 x 20 = 40 statements were presented in random order. Importantly, for each participant, there was no overlap in stimulus sets between Sessions 1 and 2. At the end of the second session, participants again indicated whether they had participated seriously. Finally, they were thanked and debriefed.

Counterbalancing and computation of individual truth effect indices. We counterbalanced all stimulus sets across sessions and judgment tasks according to the plan shown in Table 1. Each participant was randomly assigned to one of the eight counterbalancing conditions. Participants in the first counterbalancing condition, for example, started Session 1 by performing the category-sorting task for the statements in Set A. Later, these participants rated the validity of statements in Sets A and B. Because the between-items criterion of the truth effect is usually calculated as the difference in validity ratings between a set of repeated statements and another set of new statements (Dechêne et al., 2010), individual truth effect indices were then determined as the difference in mean validity ratings for the repeated Set A and the non-repeated Set B. In Session 2, participants in the first counterbalancing condition completed the category-sorting task for statements in Set C, and later rated the validity of statements in Sets C and D. For these participants, individual truth effect indices in Session 2 were then determined as the difference in mean validity ratings for the repeated Set C and the non-repeated Set D.

Table 1

Counterbalancing of Stimulus Sets A, B, C, and D across sessions and judgment tasks in Experiment 1

		Session 1		Session 2	
Counterbalancing condition	*N*	category sorting	validity rating	category sorting	validity rating
1	22	A	AB	C	CD
2	21	B	BA	D	DC
3	19	A	AB	D	DC
4	18	B	BA	C	CD
5	25	C	CD	A	AB
6	21	D	DC	B	BA
7	25	D	DC	A	AB
8	15	C	CD	B	BA

Calculating raw differences between validity ratings is but one way of operationalizing the truth effect. To assess the generalizability of our findings, we also computed two additional indices of the truth effect’s magnitude. First, following Unkelbach (2007), we used a signal-detection theory (SDT) framework to compute bias parameters in order to measure participants’ tendency to more readily accept previously shown statements. Second, we adopted a linear regression approach to calculate residual scores as an indicator of whether participants' validity ratings for repeated statements were higher or lower than would have been expected on the basis of their validity ratings for non-repeated statements.

Results and Discussion

We used R (Version 3.5.1; R Core Team, 2018) for all analyses reported below. Prior to conducting the main analysis, we compared validity ratings across the eight counterbalancing conditions with univariate analyses of variance. Validity ratings did not differ between counterbalancing conditions in either the first or the second session (all *p*s > .05). We therefore pooled across all counterbalancing conditions in the following analyses.

To check if a truth effect was present in Session 1, we conducted a repeated measures ANOVA using statements' repetition status (repeated vs. new) and their truth status (true vs. false) as independent variables, and participants' mean validity ratings as the dependent variable. There was only a significant main effect of repetition status, $F(1, 165) = 31.71$, $p < .001$, $\eta_p^2 = .16$, indicating that in Session 1, participants indeed assigned higher validity ratings to repeated statements ($M = 3.83$, $SD = 0.49$) than to new statements ($M = 3.64$, $SD = 0.42$). To test for the presence of a truth effect in Session 2, we conducted an analogous 2 (repetition status: repeated vs. new) x 2 (truth status: true vs. false) repeated measures ANOVA. Again, there was only a significant main effect of repetition status, $F(1, 165) = 28.30$, $p < .001$, $\eta_p^2 = .15$. As in Session 1, participants assigned higher validity ratings to repeated statements ($M = 3.78$, $SD = 0.47$) than to new statements ($M = 3.60$, $SD = 0.43$).

As our primary concern was the test-retest stability of the truth effect, we calculated an individual truth effect index for each participant in each session by subtracting the mean validity rating for new statements from the mean validity rating for repeated statements. This computation of raw mean differences is the most straightforward way to calculate a truth effect index based on the between-items criterion (Dechêne et al., 2010). In the first session, 64% of all participants provided higher validity ratings for repeated statements compared to new statements, as indicated by a truth effect index that was numerically larger than zero. The same pattern was observed in the second session, where 63% of all participants provided higher validity ratings for repeated statements compared to new statements. However, the main question was whether participants were consistent in their tendency to exhibit a truth effect across both experimental sessions. As can be seen in Figure 1, the truth effect was not stable at all on an individual level and truth effect indices were virtually uncorrelated across sessions, $r = .04$, $p = .579$. Thus, the magnitude of a participant's truth effect in Session 1 did not predict the magnitude of the effect in Session 2.

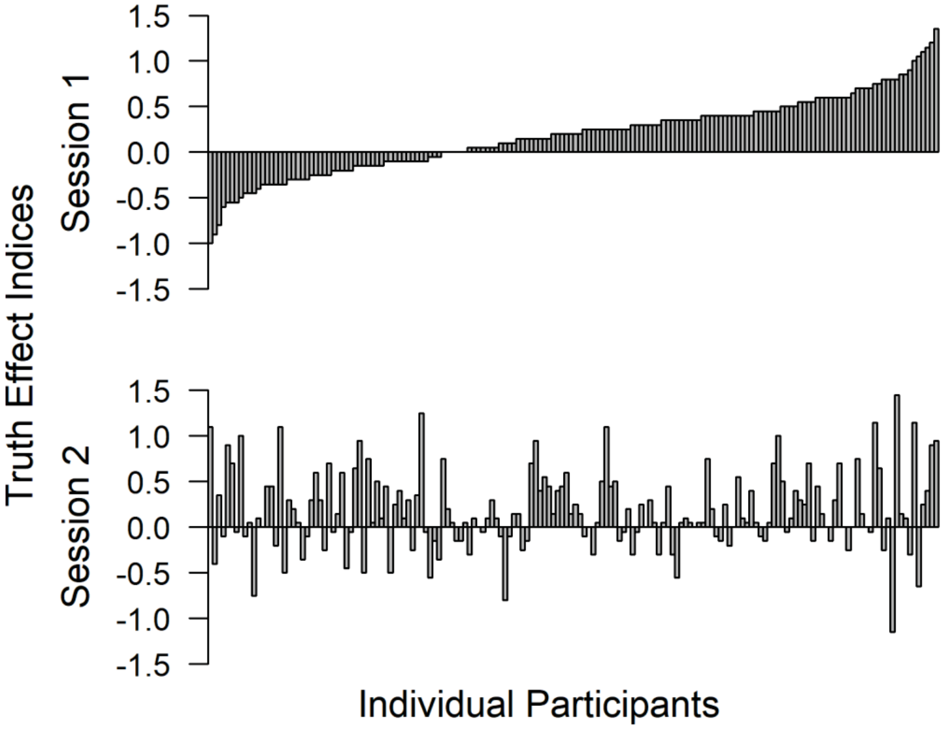

Figure 1. Magnitude of the truth effect per participant in the first (top) and in the second session (bottom) of Experiment 1. Each bar represents the difference between a participant's mean validity rating for repeated statements and his or her mean validity rating for new statements. Positive values indicate that participants assigned higher validity ratings to repeated statements. In both parts of the figure, participants are sorted according to their Session 1 truth effect indices (in ascending order).

To ensure that our results regarding the test-retest stability of the truth effect were not dependent on a particular operationalization of the effect's magnitude, we calculated two additional indices for susceptibility to the truth effect. First, we computed an index based on bias parameters from signal detection theory; second, we used linear regression to calculate residual scores.

Signal detection analysis has already proved to be a useful tool in previous studies investigating the truth effect (Unkelbach, 2007). It provides estimates for two theoretically independent parameters, discrimination ability (d') and response bias (c). In an SDT analysis of the truth effect, discrimination ability serves as an index of a participant's knowledge, that is, his or her ability to discriminate between true and false statements. The response bias parameter represents participants' tendency to be either conservative or liberal when judging the truth of a statement. In studies investigating the truth effect, SDT indices are computed separately for repeated and new statements. A truth effect is present if validity judgments are more liberal for repeated statements than for new statements. Knowledge regarding a statement's actual truth status should influence d', but not c. Assessing the test-retest stability of the truth effect on the basis of SDT bias parameters therefore ensures that estimates of the truth effect's stability are not contaminated by participants' knowledge. Following Paulhus, Harms, Bruce, and Lysy (2003), we calculated SDT parameters by using an iterative procedure to account for the fact that validity judgments in the present study were not binary in nature. To this end, we dichotomized participants' validity ratings at each of the five possible cut-off points on the 6-point response scale (i.e., 1, 2, 3, 4, and 5). Ratings above the respective cut-off point were considered to be "true" responses. For each participant, we then computed a hit rate and a false-alarm rate at each cut-off point. The hit rate was defined as the proportion of true items for which participants provided a "true" judgment, i.e., a rating above the respective cut-off point; the false-alarm rate was defined as the proportion of false items for which participants provided a "true" judgment. To prevent hit and false-alarm rates of zero or one, we followed the loglinear approach by Hautus (1995). In a next step, hit and false-alarm rates were used to compute an index of discrimination ability (d') and an index of bias (c) at each cut-off point; we then averaged these indices across cut-off points to obtain an overall index of discrimination ability and an overall index of bias for each participant. All computational steps were conducted separately for both sessions and for new and repeated statements. This resulted in four estimates of discrimination ability and four estimates of bias per participant.

In Session 1, discrimination ability was not significantly different from zero for both repeated (d'_{rep} = -0.01) and new statements (d'_{new} = 0.03), *p*s ≥ .358. This finding confirms that we successfully selected statements participants had no prior knowledge of. More importantly, however, validity judgments were found to be more liberal for repeated (c_{rep} = -0.19) than for new statements (c_{new} = -0.08), $t(165) = 5.46$, $p < .001$, indicating the occurrence of a truth effect. In Session 2, the same pattern of results emerged: Discrimination ability did not differ from zero for either repeated (d'_{rep} = 0.03) or new statements (d'_{new} < -0.01), *p*s ≥ .485. Again, however, truth judgments were more liberal for repeated (c_{rep} = -0.16) than for new statements (c_{new} = -0.06), $t(165) = 4.81$, $p < .001$. To determine the test-retest stability of the truth effect based on SDT bias parameters, we first subtracted participants' bias indices for new statements from their bias indices for repeated statements separately for each session. The correlation between the resulting two indices (Δc_{t1}, Δc_{t2}) provided an additional estimate of the truth effect's stability. Corroborating the previous analysis based on raw difference scores, this correlation was found to be low and insignificant, $r = .05$, $p = .547$.

To complement our analysis of the truth effect's test-retest stability based on raw difference scores and SDT bias parameters, we also assessed the truth effect's stability based on residual scores to address statistical concerns regarding the potential unreliability of raw difference scores in correlational research (for an overview, see Zumbo, 1999). Residual scores have been suggested as an alternative to raw difference scores and have sometimes been found to be slightly more reliable (e.g., Caruso, 2004; Williams, Zimmerman, & Mazzagatti, 1987). We argue that residual scores can be used as an indicator of individual differences in the truth effect because they allow for investigating whether validity ratings for repeated statements differ from what would be expected on the basis of validity ratings for non-repeated statements. To obtain residual scores, we predicted mean validity ratings for repeated statements from mean validity ratings for non-repeated statements in each of the two sessions by means of linear regression. We then interpreted the residuals as indices of the truth effect's magnitude. This operationalization of the magnitude of the truth effect differs conceptually from the computation of raw difference scores: A positive residual indicates that validity ratings for repeated statements are higher than would have been expected on the basis of validity ratings for non-repeated statements. Likewise, a negative residual indicates that validity ratings for repeated statements are lower than would have been expected on the basis of validity ratings for non-repeated statements. To reassess the stability of individual differences in the truth effect, we correlated residual scores across sessions and found retest stability to be low, albeit significantly higher than zero, $r = .16$, $p = .034$. Using the formulae provided by Raghunathan, Rosenthal, and Rubin (1996) as implemented in the R package cocor (Diedenhofen & Musch, 2015), this correlation was also found to be significantly larger than the corresponding correlation of $r = .04$ that was observed for raw difference scores, $z = -3.14$, $p = .002$.

Taken together, the results of Experiment 1 indicate a surprisingly low test-retest stability of the truth effect. This finding held true irrespective of whether susceptibility to the truth effect was operationalized as a raw difference score between mean validity ratings for repeated and non-repeated statements, or whether it was computed using bias parameters from signal detection theory. The test-retest stability of the truth effect was found to be only slightly higher when calculations were based on residual scores. However, even in this case the stability of the truth effect was well below test-retest stabilities reported for other phenomena in the area of judgment and decision making (e.g., Kantner & Lindsay, 2012; Kirby, 2009; Michalkiewicz & Erdfelder, 2016).

Experiment 2

In both sessions of Experiment 1, participants rated the validity of two different sets of statements, one of which had already been presented before. The presence of a truth effect was tested by comparing the validity ratings for the new set of statements and the repeatedly presented set of statements. This way of operationalizing the truth effect employs a *between-items criterion* (Dechêne et al., 2010). However, the truth effect can also be assessed by requiring participants to evaluate the truth of the same set of statements twice. This alternative operationalization of the truth effect employs a *within-items criterion* (Dechêne et al., 2010). In Experiment 2, we tested whether the low test-retest stability of the truth effect we observed in Experiment 1 would replicate when using the within-items criterion, which is arguably a purer measure of the truth effect because it does not confound the magnitude of the effect with potential a priori differences in the credibility of item sets.

For the truth effect to be assessed based on the within-items criterion, participants need to evaluate the validity of a given set of statements at two different points in time. This way of measuring the truth effect has been successfully used in a large number of previous studies (see Dechêne et al., 2010). However, Nadarevic and Erdfelder (2014) observed no truth effect when they asked participants to judge the validity of the same set of statements twice within a very short time interval of only 10 minutes. This was probably because participants were either still capable of reproducing their original judgments in the second validity-rating phase, or because performing the same validity-rating task twice in close succession induced skepticism and caused participants to discount processing fluency as a cue for validity. We therefore took care that there was a sufficiently large time interval of one week between the validity-rating tasks. Because determining the truth effect's test-retest stability requires not one, but two indices of the truth effect's magnitude, we asked participants to perform the standard validity-rating task on three occasions separated by one week each. One set of statements that was first presented in Session 1 was repeated in Session 2, and a different set of statements that was first presented in Session 2 was repeated in Session 3. Due to this experimental procedure, two individual truth effect indices could be calculated using the within-items criterion. The first index compared validity ratings collected for the set of statements that was first presented in Session 1 and recurred in Session 2; the second index compared validity ratings collected for a different set of statements that was first presented in Session 2 and then recurred in Session 3. As in Experiment 1, we computed the correlation of these two indices to determine the test-retest stability of the truth effect.

Method

Participants. Psychology students at the University of Duesseldorf participated in the experiment for course credit. On three successive weeks, test booklets in paper-and-pencil format were distributed at the beginning of a lecture on Differential Psychology. At the start of each session, participants were asked to tag their test booklet with a self-generated code that allowed us to preserve participants' anonymity when matching their booklets across sessions. The number of participants who completed the test booklets in Sessions 1, 2 and 3 was 154, 163, and 149, respectively. One hundred and twenty-one participants attended all three sessions. We discarded the data from three participants who indicated that they had misunderstood the instructions.[1] Data from another two participants also had to be discarded because they were erroneously given a wrong test booklet in one of the sessions. Our final sample therefore consisted of 116 participants. With respect to gender, 93 of these participants indicated being female, and 22 indicated being male. One participant did not answer the question. To protect participants' anonymity, age was assessed in age categories only. Six percent of the participants were below 18 years of age, 78% reported being between 18 and 24, and 12% reported being between 25 and 30 years of age. Only 3% of the participants indicated being older than 30 years of age.

Materials. We used the same four stimulus sets (A, B, C, D) as in Experiment 1. Test booklets in each of the three sessions contained two of these sets. In addition to the 2 x 20 = 40 critical statements, we also presented eight unique and easy-to-solve filler statements per session, four of which were true and four of which were false. By including easy filler statements, we tried to maintain participants' motivation by fostering their impression that their knowledge was helpful in judging the factual truth of the presented statements. Thus, the test booklets in each of the three sessions consisted of a total of 40 critical and 8 filler statements, all of which had to be rated for validity. However, validity ratings for the filler items were not included in any of the later analyses.

Procedure. The experiment was announced as a quiz being conducted to select items for a tricky knowledge test. At the beginning of Session 1, all participants were informed that they would have to rate the validity of true and false statements on three successive occasions. We

[1] However, not excluding these participants did not change any of the results reported below.

also briefed participants that they might be uncertain when evaluating the truth of some of the statements, but that they should nevertheless judge the validity of each statement to the best of their knowledge. At the start of Sessions 2 and 3, we additionally told participants that all statements were selected randomly from a large pool of statements, and that due to this random sampling they might encounter some statements they were already familiar with. In each session, participants then rated the validity of 40 critical and 8 filler statements on a 6-point scale ranging from 1 (*definitely false*) to 6 (*definitely true*). Half of the 40 critical statements presented in Session 1 were presented again in Session 2, and 20 new statements first presented in Session 2 were presented again in Session 3, along with 20 additional new statements. In Session 3, participants also provided some basic demographic data and indicated whether they had looked up the truth status of any of the statements they had been presented with in one of the three sessions. Finally, all participants were thanked and debriefed. Each of the three sessions lasted about 12 to 16 minutes; the time intervals between Sessions 1 and 2 and between Sessions 2 and 3 were 8 and 7 days, respectively.

Counterbalancing and computation of individual truth effect indices. To control for stimulus-specific effects, we randomly assigned participants to one of two counterbalancing conditions (see Table 2), using their month of birth as a randomization device. Because participants were assigned to one of two counterbalancing conditions and took part in three consecutive sessions, six different test booklets were needed. We prepared three versions of each test booklet to control for stimulus-specific order effects. For each version, the filler items were placed at the same fixed positions, whereas the critical statements were arranged in a different random order.

Table 2

Counterbalancing of Stimulus Sets A, B, C, and D across sessions in Experiment 2

Counterbalancing condition	*N*	Session 1	Session 2	Session 3
1	56	AC	CD	DB
2	60	AD	DC	CB

In the first counterbalancing condition, participants started the experiment in Session 1 by rating the validity of statements in Sets A and C. In Session 2, the same participants evaluated the truth of statements in Sets C and D. Finally, participants ended the experiment in Session 3 by rating the validity of statements in Sets D and B. Thus, a first truth effect index based on the within-items criterion could be computed using the validity ratings for statements in Set C, whereas a second index could be computed using the validity ratings for statements in Set D. In the second counterbalancing condition, the order of the assignment of Sets C and D was reversed (see Table 2). In this condition, a first truth effect index was computed using the validity ratings for statements in Set D, and a second index was computed using the validity ratings for statements in Set C. Validity ratings for statements in Sets A and B were not used to compute any within-items index as statements in these sets were presented only once throughout the experiment. As in Experiment 1, we calculated the truth effect's test-retest stability based on (a) raw difference scores, (b) SDT bias parameters and (c) residual scores.

Results and Discussion

Seven participants admitted that they had read up on the truth status of one or two of the critical statements in between sessions. We therefore discarded the corresponding validity judgments for these participants, resulting in a loss of 17 validity ratings. In a next step, we compared the validity ratings across the two counterbalancing conditions for each of the three sessions using *t*-tests for independent samples. Mean validity ratings did not differ across counterbalancing conditions in any session (all *p*s > .05). We therefore pooled across both counterbalancing conditions in all subsequent analyses.

The truth effect was assessed according to the within-items criterion, i.e., by comparing validity ratings for participants' first and second encounter with a given set of statements (Dechêne et al., 2010). Because statements that were presented only once throughout the experiment are not needed for assessing the within-items criterion of the truth effect, we did not include the corresponding validity ratings in the following analyses. Validity ratings were also excluded for repeated statements whenever a given participant failed to judge a particular statement twice, that is, whenever there were missing values on at least one occasion. Across participants, this applied to 44 out of the 4640 (0.95%) statements that had been shown twice throughout the experiment.

To test whether repetition led to an increase in the perceived validity of statements that had been shown in both Sessions 1 and 2, we conducted a repeated measures ANOVA using experimental session (Session 1 vs. Session 2) and statements' truth status (true vs. false) as independent variables, and participants' mean validity ratings as the dependent variable. As expected, there was a significant main effect of session, $F(1, 115) = 12.50$, $p < .001$, $\eta_p^2 = .10$, confirming that participants assigned higher validity ratings to statements during their second occurrence in Session 2 (*M* = 3.79, *SD* = 0.32) than during their first occurrence in Session 1 (*M* = 3.69, *SD* = 0.35). The results also showed that participants found it difficult to judge the assertions' truth status; they even assigned somewhat higher validity ratings to false (*M* = 3.79, *SD* = 0.35) than to true statements (*M* = 3.69, *SD* = 0.38), *F*(1, 115) = 5.55, *p* = .020, η_p^2 = .05. Importantly, however, the truth effect was not moderated by the factual correctness of statements since there was no interaction between session and truth status, *F* < 1.

To test whether repetition led to an increase in the perceived validity of statements that had been shown in both Sessions 2 and 3, we conducted another repeated measures ANOVA using experimental session (Session 2 vs. Session 3) and statements' truth status (true vs. false) as independent variables, and participants' mean validity ratings as the dependent variable. Again, there was a significant main effect of session, *F*(1, 115) = 51.48, *p* < .001, η_p^2 = .31, confirming that participants assigned higher validity ratings to statements during their second occurrence in Session 3 (*M* = 3.79, *SD* = 0.32) than during their first occurrence in Session 2 (*M* = 3.60, *SD* = 0.28). False statements (*M* = 3.73, *SD* = 0.34) were again judged to be slightly more credible than true statements (*M* = 3.66, *SD* = 0.35). However, this difference failed to reach significance, *F*(1, 115) = 3.75, *p* = .055, η_p^2 = .03. More importantly, the truth effect was again not moderated by the factual correctness of statements, since there was no interaction between session and truth status, *F* < 1.[2]

[2] Apart from assessing the truth effect via the within-items criterion, the design of Experiment 2 also allowed to assess the effect via the between-items criterion. The results of additional analyses based on the between-items criterion confirmed the findings of our main analyses:

To assess the test-retest stability of the truth effect, we calculated two within-items indices of the truth effect for each participant based on raw difference scores. The first index represented the mean shift in validity ratings for a given set of statements from Session 1 to Session 2. This index was numerically larger than zero for 57% of participants, indicating that the majority of participants provided higher validity ratings to statements when they were shown for the second time. The second index represented the mean shift in validity ratings from Session 2 to Session 3 for a different set of statements. This index was numerically larger than zero for 69% of the participants. While these findings confirmed the occurrence of a truth effect, our primary concern was whether the effect was stable on an individual level. As can be seen in Figure 2, this was not the case. The magnitude of the truth effect participants exhibited for statements presented in Sessions 1 and 2 was not predictive of the magnitude of the truth effect they exhibited for statements presented in Sessions 2 and 3, *r* = .12, *p* = .191.

Repeated statements were judged to be significantly more credible than new statements both in Session 2 and in Session 3 (all *p*s < .001).

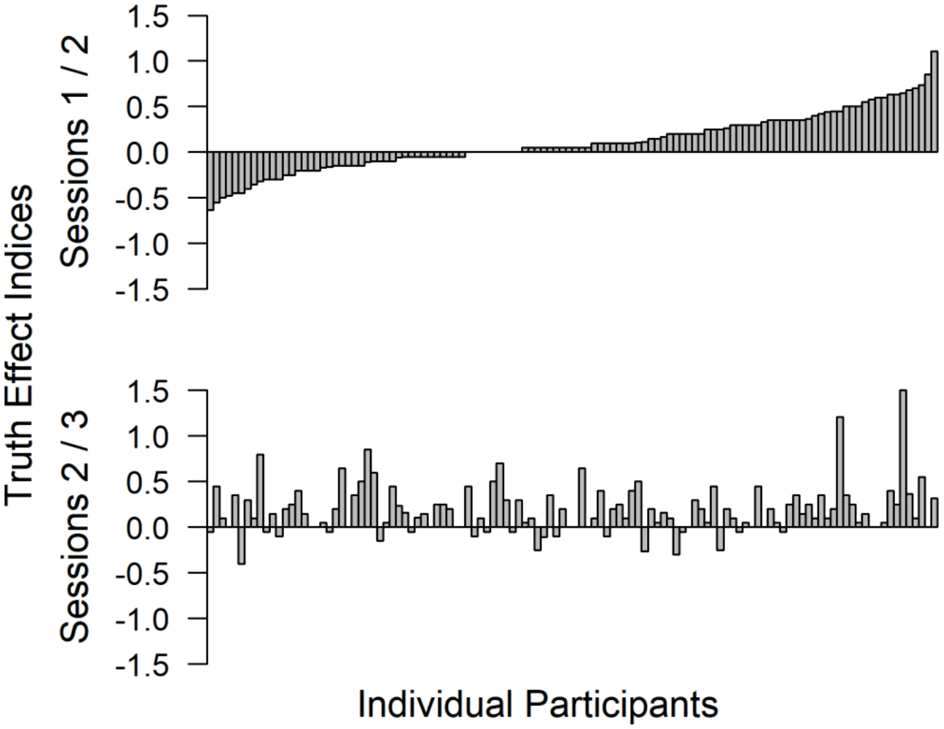

Figure 2. Participants' individual truth effect indices based on raw mean differences in Experiment 2. Bars either represent the shift in validity ratings for statements that were presented in the first and the second session (top) or the shift in validity ratings for a different set of statements that were presented in the second and the third session (bottom). Positive values indicate that participants assigned higher validity ratings to statements during their second presentation. In both parts of the figure, participants are sorted according to the size of their truth effect calculated on the basis of validity ratings from Sessions 1 and 2 (in ascending order).

As in Experiment 1, we supplemented our main analysis by calculating SDT parameters and residual scores as two alternative indices for susceptibility to the truth effect. Computation of SDT parameters followed the same basic procedure described in Experiment 1 and resulted in four estimates of discrimination ability and four estimates of bias per participant. Discrimination ability for statements that had been shown for the first time in Session 1 did not differ significantly from zero (d'_{new} = -0.03, p = .300). However, when calculations were based on the validity ratings collected during participants' second encounter with the same statements in Session 2, discrimination ability was marginally worse than zero (d'_{rep} = -0.05, p = .061). These results mirror the main effect of statement type that emerged in the ANOVA and confirm that participants tended to assign higher validity ratings to factually false statements, resulting in negative discrimination ability indices. More importantly, however, participants' response bias was more liberal for statements during their second occurrence in Session 2 (c_{rep} = -0.17) than during their first occurrence in Session 1 (c_{new} = -0.11), $t(115) = 3.61$, $p < .001$, indicating that a truth effect had occurred. A similar pattern of results emerged when discrimination ability and response bias indices were based on validity ratings for statements that were initially presented in Session 2 and then recurred in Session 3. Neither discrimination ability index differed significantly from zero (d'_{new} = -0.03, d'_{rep} = -0.05, ps ≥ .141), but a significant truth effect occurred; participants' response bias was more liberal for statements during their second presentation in Session 3 (c_{rep} = -0.17) than during their first presentation in Session 2 (c_{new} = -0.06), $t(115) = 7.33$, $p < .001$. To assess the test-retest stability of the truth effect within the SDT framework, we generated two truth effect indices by calculating the differences between each pair of response bias parameters that were computed based on validity ratings for the same set of statements ($\Delta c_{\text{t1\_t2}}$, $\Delta c_{\text{t2\_t3}}$). The correlation between these indices again indicated that the stability of the truth effect was very low, r = .06, p = .499.

Finally, we also assessed the truth effect's test-retest stability based on residual scores. In a first linear regression, we predicted mean validity ratings for repeated statements in Session 2 based on mean validity ratings for the same statements previously shown in Session 1. Likewise, in a second linear regression, we predicted mean validity ratings for repeated statements in

Session 3 based on validity ratings for the same statements previously shown in Session 2. We thus obtained two residuals per participant, which we then correlated to estimate the stability of individual differences in the magnitude of the truth effect. Although this correlation was also rather low, $r = .27$, $p = .003$, it was again found to be significantly larger than the corresponding correlation of $r = .12$ that was observed for raw difference scores, $z = -2.70$, $p = .007$. This result parallels the findings from Experiment 1.

The results of Experiment 2 reveal that although a truth effect was present at two different points in time on an aggregate level, the test-retest stability of the truth effect was again rather low. This was true irrespective of whether the truth effect was measured using raw difference scores or SDT bias parameters. Similar to Experiment 1, the truth effect's stability was found to be somewhat higher when calculations were based on residual scores; even in this case, however, the stability of the truth effect was almost negligible.

General Discussion

Several authors have noted a dearth of research on individual differences in the magnitude of the truth effect and its correlates (Arkes et al., 1991; Dechêne et al., 2010). In the present study, we therefore set out to test an essential precondition for detecting replicable associations between the magnitude of the truth effect and other cognitive or personality variables. Replicable relations with other variables can only be expected if the truth effect is temporally stable on an individual level (cf. Michalkiewicz & Erdfelder, 2016). If the truth effect is not stable, however, any associations with cognitive or personality variables that are occasionally observed are unlikely to replicate. Interestingly, this is what happened in a recent study by De keersmaecker et al. (2018); whereas individuals high in experiential thinking were found to exhibit a significantly larger truth effect in one experiment, this finding did not replicate in a second experiment.

By assessing the test-retest stability of the truth effect, we examined whether people consistently fall prey to the influence of repetition when asked to judge the truth of ambiguous statements. In both of our experiments, the truth effect's test-retest stability was found to be much lower than stability coefficients reported for other phenomena from the area of judgment and decision making (e.g., Kantner & Lindsay, 2012; Kirby, 2009; Michalkiewicz & Erdfelder, 2016). In particular, test-retest stability was found to be very low regardless of whether the truth effect was measured using the between-items criterion (Experiment 1) or the within-items criterion (Experiment 2). Moreover, although stability was shown to be slightly higher when based on residual scores, the finding of low test-retest stability was obtained for all three operationalizations of the truth effect's magnitude that were used in the present study. Finally, the finding of low test-retest stability was not dependent on whether the truth effect was assessed in an online experiment or a classroom setting, lending further support to the generalizability of our results.

The low test-retest stability of the truth effect suggests that the effect is not an individually stable phenomenon. However, low stability coefficients might also result from the use of unreliable measures (Crocker & Algina, 1986). The question of how the truth effect can be measured reliably on an individual level is of major importance when investigating the effect from an individual difference perspective. The present study contributed to answering this question by scrutinizing several different operationalizations of the truth effect. The most intuitive way of measuring the effect was based on the calculation of simple difference scores. In the past, such scores have been criticized for being inherently unreliable (for a review, see

Zumbo, 1999), even though this criticism may only apply under specific conditions (Rogosa & Willett, 1983; Zimmerman & Williams, 1982). In particular, the reliability of the difference between two measures X and Y typically decreases with an increasing correlation between both measures and with decreasing differences in their variances (Williams & Zimmerman, 1996). In our experiments, we calculated differences between validity ratings for repeatedly presented statements and new statements (Experiment 1) or between validity ratings for statements during their first and their second occurrence (Experiment 2). Validity ratings for new and repeated statements had similar variances and were highly correlated (Experiment 1: $r_{\text{t1}} = .59$, $r_{\text{t2}} = .55$; Experiment 2: $r_{\text{t1/t2}} = .56$, $r_{\text{t2/t3}} = .59$), which is probably not an uncommon finding in experimental studies on the truth effect. Although such circumstances do not preclude an acceptable reliability of simple difference scores, they do make it more difficult to achieve satisfactory reliability (cf. Williams & Zimmerman, 1996). Residual scores can be more reliable than raw difference scores under specific conditions (Williams & Zimmerman, 1983; Zumbo, 1992). Empirically, however, the reliability of residual scores has been found to only slightly exceed the reliability of raw differences (e.g., Caruso, 2004; Williams et al., 1987), and it is worth noting that the truth effect's test-retest stability turned out to also be low for residual scores in the present study.

Taken together, these findings and considerations suggest that current operationalizations of the truth effect have to be reconsidered if the effect is to be measured reliably on an individual level. Studies that aim to investigate the truth effect from an individual difference perspective will have to address these reliability concerns. It might be possible—albeit costly—to improve the test-retest stability of the truth effect by using a much larger number of items since—all other things being equal—using a larger number of items increases reliability (e.g., Buchner &

Wippich, 2000; see also Nunnally & Bernstein, 1994). Future research should also examine the test-retest stability of more advanced operationalizations of the truth effect based on multinomial processing tree (MPT) modeling (Fazio, Brashier, Payne, & Marsh, 2015; Unkelbach & Stahl, 2009). MPT models allow to estimate the probability with which specific cognitive processes occur and contribute to observed behavior. MPT models of the truth effect, for example, can be used to estimate the probability with which participants rely on processing fluency when asked to judge the truth of a (repeated) statement. Whereas traditional MPT models provide probability estimates for the occurrence of cognitive processes on a group level, recently developed hierarchical models explicitly account for participant heterogeneity by providing parameter estimates for individual participants, making these models particularly well-suited for research on individual differences (Heck, Arnold, & Arnold, 2018; Smith & Batchelder, 2010). It might therefore be worthwhile to use hierarchical MPT models if they can help to more reliably assess the truth effect's magnitude on an individual level. However, these models require categorical data instead of the rating-scale data collected in the present experiments.

When investigating an effect from an individual difference perspective, it should be kept in mind that an effect being robust does not imply that it is also reliable. In a recent paper, Hedge, Powell, and Sumner (2018) have dealt with this conundrum, which also applies to other areas of research in cognitive psychology (see, e.g., Pohl, 1999). They noted that researchers in experimental psychology often implicitly assume that cognitive tasks that have turned out to be "'reliable' workhorses" (p. 1167) will also serve well as objective measures of individual variation. To explain why this is not necessarily the case, Hedge et al. pointed out that the term "reliability" is used differently in experimental and correlational research. According to Hedge et al., an effect is considered to be "reliable" in experimental research if it is exhibited by the majority of participants, replicates well and is of consistent size across studies. Arguably, this kind of reliability might better be referred to as "robustness", because in correlational research the term reliability is typically used in a different way. Here, the term "reliability" refers to the degree to which a measure consistently ranks individuals. Hedge et al. further noted that cognitive paradigms in experimental research are usually developed and selected for providing robust effects, which is why they are typically accompanied by low between-subjects variability. A low between-subjects variability, however, necessarily reduces reliability—in the sense in which this term is used in correlational research—and therefore poses a major impediment to finding replicable relations between the magnitude of an effect and potentially related cognitive or personality traits (Hedge et al., 2018). Following this reasoning, it is entirely possible than an effect can be robust and unreliable at the same time.

To summarize, the present study provides evidence for a surprisingly low test-retest stability of the truth effect. This finding should be a cause of concern for studies that aim to investigate the effect from an individual difference perspective. Any search for replicable relations between the truth effect and other cognitive or personality variables is bound to fail if the truth effect is not an individually stable phenomenon and cannot be measured reliably on an individual level.

References

Arkes, H. R., Boehm, L. E., & Xu, G. (1991). Determinants of judged validity. *Journal of Experimental Social Psychology*, *27*, 576–605. doi:10.1016/0022-1031(91)90026-3

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*, 527–535. doi:10.3758/s13428-012-0265-2

Beck, R. C., & Triplett, M. F. (2009). Test-retest reliability of a group-administered paper-pencil measure of delay discounting. *Experimental and Clinical Psychopharmacology*, *17*, 345–355. doi:10.1037/a0017078

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin*, *20*, 285–293. doi:10.1177/0146167294203006

Buchner, A., & Wippich, W. (2000). On the reliability of implicit and explicit memory measures. *Cognitive Psychology*, *40*, 227–259. doi:10.1006/cogp.1999.0731

Cacioppo, J. T., Petty, R. E., & Kao, C. F. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment*, *48*, 306–307. doi:10.1207/s15327752jpa4803_13

Caruso, J. C. (2004). A comparison of the reliabilities of four types of difference scores for five cognitive assessment batteries. *European Journal of Psychological Assessment*, *20*, 166–171. doi:10.1027/1015-5759.20.3.166

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory*. New York, NY: Holt, Rinehart and Winston.

De keersmaecker, J., Roets, A., Pennycook, G., & Rand, D. G. (2018). *Is the illusory truth effect robust to individual differences in cognitive ability, need for cognitive closure, and cognitive style?* Manuscript in preparation.

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*, 238–257. doi:10.1177/1088868309352251

Diedenhofen, B., & Musch, J. (2015). Cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE*, *10*(4), e0121945. doi:10.1371/journal.pone.0121945

DiFonzo, N., Beckstead, J. W., Stupak, N., & Walders, K. (2016). Validity judgments of rumors heard multiple times: The shape of the truth effect. *Social Influence*, *11*, 22–39. doi:10.1080/15534510.2015.1137224

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002. doi:10.1037/xge0000098

Goldstein, D. G., & Gigerenzer, G. (2002). Models of ecological rationality: The recognition heuristic. *Psychological Review*, *109*, 75–90. doi:10.1037/0033-295X.109.1.75

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112. doi:10.1016/S0022-5371(77)80012-1

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d'. *Behavior Research Methods, Instruments, & Computers*, *27*, 46–51. doi:10.3758/BF03203619

Heck, D. W., Arnold, N. R., & Arnold, D. (2018). TreeBUGS: An r package for hierarchical multinomial-processing-tree modeling. *Behavior Research Methods*, *50*, 264–284. doi:10.3758/s13428-017-0869-7

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*, 1166–1186. doi:10.3758/s13428-017-0935-1

Kantner, J., & Lindsay, D. S. (2012). Response bias in recognition memory as a cognitive trait. *Memory & Cognition*, *40*, 1163–1177. doi:10.3758/s13421-012-0226-0

Kantner, J., & Lindsay, D. S. (2014). Cross-situational consistency in recognition memory response bias. *Psychonomic Bulletin & Review*, *21*, 1272–1280. doi:10.3758/s13423-014-0608-3

Kirby, K. N. (2009). One-year temporal stability of delay-discount rates. *Psychonomic Bulletin & Review*, *16*, 457–462. doi:10.3758/PBR.16.3.457

Kruglanski, A. W., & Webster, D. M. (1996). Motivated closing of the mind: "Seizing" and "freezing". *Psychological Review*, *103*, 263–283. doi:10.1037//0033-295X.103.2.263

Ladowsky-Brooks, R. L. (2010). The truth effect in relation to neuropsychological functioning in traumatic brain injury. *Brain Injury*, *24*, 1343–1349. doi:10.3109/02699052.2010.506856

Maio, G. R., & Esses, V. M. (2001). The need for affect: Individual differences in the motivation to approach or avoid emotions. *Journal of Personality*, *69*, 583–615. doi:10.1111/1467-6494.694156

Michalkiewicz, M., & Erdfelder, E. (2016). Individual differences in use of the recognition heuristic are stable across time, choice objects, domains, and presentation formats. *Memory & Cognition*, *44*, 454–468. doi:10.3758/s13421-015-0567-6

Moritz, S., Köther, U., Woodward, T. S., Veckenstedt, R., Dechêne, A., & Stahl, C. (2012). Repetition is good? An internet trial on the illusory truth effect in schizophrenia and nonclinical participants. *Journal of Behavior Therapy and Experimental Psychiatry*, *43*, 1058–1063. doi:10.1016/j.jbtep.2012.04.004

Nadarevic, L., & Erdfelder, E. (2014). Initial judgment task and delay of the final validity-rating task moderate the truth effect. *Consciousness and Cognition*, *23*, 74–84. doi:10.1016/j.concog.2013.12.002

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory*. New York, NY: McGraw-Hill.

Pacini, R., & Epstein, S. (1999). The relation of rational and experiential information processing styles to personality, basic beliefs, and the ratio-bias phenomenon. *Journal of Personality and Social Psychology*, *76*, 972–987. doi:10.1037//0022-3514.76.6.972

Paulhus, D. L., Harms, P. D., Bruce, M. N., & Lysy, D. C. (2003). The over-claiming technique: Measuring self-enhancement independent of ability. *Journal of Personality and Social Psychology*, *84*, 890–804. doi:10.1037/0022-3514.84.4.890

Pohl, R. F. (1999). *Hindsight bias: Robust, but not reliable*. Unpublished manuscript, Justus-Liebig-University, Giessen, Germany.

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Raghunathan, T. E., Rosenthal, R., & Rubin, D. B. (1996). Comparing correlated but nonoverlapping correlations. *Psychological Methods*, *1*, 178–183. doi:10.1037/1082-989X.1.2.178

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338–342. doi:10.1006/ccog.1999.0386

Rogosa, D. R., & Willett, J. B. (1983). Demonstrating the reliability of the difference score in the measurement of change. *Journal of Educational Measurement*, *20*, 335–343. doi:10.1111/j.1745-3984.1983.tb00211.x

Smith, J. B., & Batchelder, W. H. (2010). Beta-mpt: Multinomial processing tree models for addressing individual differences. *Journal of Mathematical Psychology*, *54*, 167–183. doi:10.1016/j.jmp.2009.06.007

Sundar, A., Kardes, F. R., & Wright, S. A. (2015). The influence of repetitive health messages and sensitivity to fluency on the truth effect in advertising. *Journal of Advertising*, *44*, 375–387. doi:10.1080/00913367.2015.1045154

Toplak, M. E., West, R. F., & Stanovich, K. E. (2014). Assessing miserly information processing: An expansion of the cognitive reflection test. *Thinking & Reasoning*, *20*, 147–168. doi:10.1080/13546783.2013.844729

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 219–230. doi:10.1037/0278-7393.33.1.219

Unkelbach, C., & Rom, S. C. (2017). A referential theory of the repetition-induced truth effect. *Cognition*, *160*, 110–126. doi:10.1016/j.cognition.2016.12.016

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22–38. doi:10.1016/j.concog.2008.09.006

Williams, R. H., & Zimmerman, D. W. (1983). The comparative reliability of simple and residualized difference scores. *Journal of Experimental Education*, *51*, 94–97. doi:10.1080/00220973.1982.11011846

Williams, R. H., & Zimmerman, D. W. (1996). Are simple gain scores obsolete? *Applied Psychological Measurement*, *20*, 59–69. doi:10.1177/014662169602000106

Williams, R. H., Zimmerman, D. W., & Mazzagatti, R. D. (1987). Large sample estimates of the reliability of simple, residualized, and base-free gain scores. *Journal of Experimental Education*, *55*, 116–118. doi:10.1080/00220973.1987.10806443

Zimmerman, D. W., & Williams, R. H. (1982). Gain scores in research can be highly reliable. *Journal of Educational Measurement*, *19*, 149–154. doi:10.1111/j.1745-3984.1982.tb00124.x

Zumbo, B. D. (1992). The comparative reliability of simple and residualized difference scores: A corrigendum. *Journal of Experimental Education*, *61*, 81–83. doi:10.1080/00220973.1992.9943852

Zumbo, B. D. (1999). The simple difference score as an inherently poor measure of change: Some reality, much mythology. In B. Thompson (Ed.), *Advances in social science methodology* (pp. 269–304). Greenwich, CT: JAI Press.

Debiasing the truth effect: Exploring the effects of warnings on judgments of truth

Frank Calio[1], Lena Nadarevic[2], & Jochen Musch[1]

[1] University of Duesseldorf

[2] University of Mannheim

Author note

Frank Calio and Jochen Musch, Department of Experimental Psychology, University of Duesseldorf, Germany; Lena Nadarevic, Department of Psychology, School of Social Sciences, University of Mannheim, Germany.

Correspondence concerning this article should be addressed to Frank Calio, Department of Experimental Psychology, University of Duesseldorf, Universitaetsstr. 1, 40225 Duesseldorf, Germany. E-mail: frank.calio@uni-duesseldorf.de

Abstract

The finding that presenting statements repeatedly typically increases their perceived truth has been referred to as the *truth effect*. The truth effect may play a role in the creation of false knowledge because repetition enhances the credibility of both true and false statements. Previous research has found that warning participants about the truth effect can successfully reduce, but not eliminate the effect. Extending these findings, we used a multinomial modeling approach to more closely investigate how warnings affect the cognitive processes that are assumed to underlie judgments of truth. In a laboratory experiment ($N = 167$), half of the participants were warned about the truth effect before judging the truth of repeated and new statements. Multinomial modeling analyses revealed that warning instructions did not affect the retrieval of knowledge or participants' guessing behavior relative to a control condition. However, warned participants exhibited a significantly reduced tendency to rely on processing fluency when judging a repeated statement's truth. These results are consistent with the notion that people tend to discount metacognitive experiences of processing ease when their informational value is questioned. Even though participants were provided with the prospect of a substantial financial reward in return for unbiased and accurate judgments of truth, warnings were only moderately successful in debiasing the truth effect, as the effect was mitigated but could not be eliminated.

Keywords: Truth effect, Truth judgments, Warning, Processing fluency, Multinomial processing trees

Word count: 222

Debiasing the truth effect: Exploring the effects of warnings on judgments of truth

In 1977, Hasher, Goldstein and Toppino investigated the effects of repetition on the perceived truth of trivia statements. They found that truth judgments for repeated statements increased over three successive sessions conducted at 2-week intervals, whereas truth judgments for non-repeated statements did not change. This credibility-enhancing effect of repetition has been replicated numerous times and is now referred to as the *truth effect* (for a review, see Dechêne, Stahl, Hansen, & Wänke, 2010). Given that the perceived truth of a statement can be increased by means as simple as repeated exposure, repetition can arguably be a powerful persuasive weapon. It has frequently been employed in politics, as for example by Cato the Elder ("Ceterum censeo Carthaginem esse delendam"; Hertwig, Gigerenzer, & Hoffrage, 1997), and in the field of advertising ("Carlsberg. Probably the best beer in the world"; Egan, 2007). Previous research has empirically confirmed that repetition increases belief in advertising claims (e.g., Hawkins & Hoch, 1992; Johar & Roggeveen, 2007; Law, Hawkins, & Craik, 1998; Roggeveen & Johar, 2002). Repetition has also been argued to play a crucial role in the propagation of and belief in myths and other widespread but untrue beliefs (Lilienfeld, Lynn, Ruscio, & Beyerstein, 2010). With the advent of social media platforms, where shared information spreads quickly, the link between repetition and credibility has become even more relevant. Recently, Pennycook, Cannon, and Rand (2018) investigated the effect of repetition on the perceived accuracy of *fake news*, which they defined as "entirely fabricated and often partisan content that is presented as factual" (p. 1865). Consistent with their hypothesis, they found that previously presented fake news headlines were rated as more accurate than novel headlines. Findings like these highlight the societal relevance of the link between repetition and perceived truth.

The most widely accepted explanation of the truth effect is based on *processing fluency*, defined as the "metacognitive experience of ease during information processing" (Dechêne et al., 2010, p. 240). According to this approach, the truth effect occurs because repetition increases the ease with which statements are processed and because people tend to use the metacognitive experience of ease as a cue for truth (Unkelbach, 2007). Empirical evidence for a direct influence of ease of processing on judgments of truth was provided by Reber and Schwarz (1999), who manipulated processing fluency by varying the color contrast of statements. Statements that were easy to process due to their high color contrast were more likely to be judged as true compared to statements that were more difficult to process due to their low color contrast. Importantly, fluent statements typically receive higher truth ratings than non-fluent statements (Alter & Oppenheimer, 2009), irrespective of whether processing fluency is induced by visual manipulations (e.g., Reber & Schwarz, 1999), linguistic manipulations (e.g., McGlone & Tofighbakhsh, 2000) or by repetition (e.g., Dechêne, Stahl, Hansen, & Wänke, 2009). In this paper, we will exclusively focus on the repetition-based truth effect.

From a societal point of view, it is a cause for concern that repetition not only increases the credibility of true information, but also strengthens the belief in factually false trivia statements (e.g., Gigerenzer, 1984), myths (Doland, 1999) and fake news (Pennycook et al., 2018). But how can people be protected from possible detrimental consequences of the truth effect? To date, only a few studies have explicitly tried to prevent or "debias" the truth effect. In some of them, participants were provided with feedback on the actual truth status of subsequently repeated statements (e.g., Brown & Nix, 1996; Mutter, Lindsey, & Pliske, 1995; Skurnik, Yoon, Park, & Schwarz, 2005). These studies demonstrated that feedback can have a marked impact on participants' truth judgments, although these effects seem to be relatively short-lived. Brown and

Nix (1996, Exp. 1), for example, found that repeated false statements were rated as significantly less true than new false statements when participants had received feedback about the repeated statements' actual truth status a week before; when the interval was one month, however, a standard truth effect was observed. These findings suggest that the credibility-enhancing effect of repetition may outlast the memory of feedback regarding a statement's actual truth status. Therefore, providing feedback may not be the most practical way to counter the truth effect. In a related study, Pennycook et al. (2018) explicitly labeled fake-news headlines as "disputed by 3rd party fact-checkers" during their first presentation. This did not abolish the truth effect that occurred for these headlines in two experiments, although the labeling led to a slight, non-significant reduction of the effect. In another study, Boehm (1994, Exp. 3) tried to debias the truth effect using an accountability manipulation. The experiment consisted of two sessions separated by a time interval of one week. Immediately before participants rated the truth of new and repeated statements in the second session, Boehm told half of the participants that they would have to justify their answers to a group of peers. Boehm expected that this accountability manipulation would lead to more elaborate processing, which might reduce or even eliminate the truth effect. However, this manipulation was not successful. Unfortunately, potential reasons for this outcome are difficult to identify, because Boehm provided very little information on the exact nature of his accountability manipulation.

The present study investigates whether explicit warnings about the repetition-based truth effect can reduce or even eliminate the effect. This question is strongly related to previous research on the efficacy of warnings in debiasing other memory or judgmental biases. A complete review of this literature is beyond the scope of this paper, but warnings have been used with varying degrees of success in previous debiasing studies. This may be because the efficacy of warnings in preventing unwanted influences on judgments critically hinges on whether several preconditions are met. For example, people must not only be aware of a biasing influence, but they must also be motivated and capable of exerting control over their judgments (Wilson & Brekke, 1994; see also Wilson, Centerbar, & Brekke, 2002).

Several studies have investigated whether warnings can prevent *hindsight bias*, a phenomenon that can be defined as “an overestimation of foresight knowledge following the receipt of outcome knowledge” (Harley, Carlsen, & Loftus, 2004, p. 962). Fischhoff (1977), for example, informed some of his participants about hindsight bias, provided them with strategic advice on how to reduce the biasing influence of outcome knowledge, and asked them to try to avoid the bias. However, this warning manipulation failed and left hindsight bias unaffected. Similar attempts to educate participants about the biasing nature of outcome knowledge did not successfully reduce hindsight bias either (e.g., Bernstein, Wilson, Pernat, & Meilleur, 2012; Davies, 1993; Harley et al., 2004; Pohl & Hell, 1996; Sharpe & Adair, 1993). Explicit warnings have also been used to try to reduce *outcome bias*, a phenomenon that is closely related to hindsight bias. When people are asked to evaluate the quality of a decision, outcome bias describes their tendency to “take outcomes into account in a way that is irrelevant to the true quality of the decision” (Baron & Hershey, 1988, p. 570). Simple warnings that inform participants about outcome bias and ask them to beware of its influence have not been successful in reducing this bias (e.g., Clarkson, Emby, & Watt, 2002; Grenier, Peecher, & Piercey, 2007).

The efficacy of warnings has also been explored with regard to the *misinformation effect*. This term denotes the finding that people’s recall of an event usually becomes less accurate when inconsistent information is presented after the event (e.g., Loftus, Miller, & Burns, 1978). In a recent meta-analysis, Blank and Launay (2014) reviewed 25 studies of the misinformation effect in which participants were exposed to misleading details after an event, but later also received a warning about the presence of misinformation. On average, post-warnings reduced the misinformation effect to less than half its size. Results were even more compelling for a subset of warning studies that implemented a so-called "enlightenment procedure" that not only informed participants about the presence of misinformation, but also explained the "scientific motivation and logic of the misinformation manipulation" (Blank & Launay, 2014, p. 79). Post-warning studies that made use of an enlightenment procedure, such as the study conducted by Oeberst and Blank (2012), completely eliminated the misinformation effect (Blank & Launay, 2014).

Recently, the effects of warnings have also been explored with regard to the truth effect. Nadarevic and Aßfalg (2017) used warnings that arguably fall into the category of enlightenment procedures (cf. Blank & Launay, 2014), as participants not only received information about the biasing nature of repetition on judged truth, but were also provided with detailed information about the scientific motivation and procedure of the study. Moreover, participants were explicitly prompted to prevent the truth effect. In the first experiment by Nadarevic and Aßfalg, participants assigned trivia statements to different categories of knowledge in an initial exposure phase and then rated the truth of new and repeated statements in a testing phase that took place one week later. Immediately before the testing phase, half of the participants received a warning about the truth effect and were asked to avoid it. Results showed that although the truth effect was reduced to less than half its size in the warning condition compared to the control condition, this difference failed to reach significance, possibly due to low statistical power (Nadarevic & Aßfalg, 2017). In a second experiment, Nadarevic and Aßfalg therefore recruited a much larger sample of participants. Moreover, they eliminated the retention interval between sessions and ensured that participants had understood the warning instructions properly by adding several control questions. Under these conditions, warning instructions significantly reduced the truth effect, but did not eliminate it. Nadarevic and Aßfalg interpreted these results as evidence for the notion that people have at least some control over their fluency-truth attributions.

In the present study, we wanted to investigate more closely how warnings affect the cognitive processes that presumably underlie judgments of truth. We therefore set up an experiment in which half of the participants received a warning about the truth effect immediately before judging the truth of repeated and new trivia statements. Similar to the warning manipulation used by Nadarevic and Aßfalg (2017), our warning explicitly briefed participants about the purpose of the study, informed them about the truth effect and asked them to try to resist the influence of repetition on judged truth. Going beyond the analyses reported by Nadarevic and Aßfalg, however, we analyzed the data using multinomial processing trees (MPT), which are a class of stochastic models that allow to disentangle and measure the cognitive processes underlying human behavior (Erdfelder et al., 2009). MPT models have already been applied in previous studies of the truth effect (e.g., Fazio, Brashier, Payne, & Marsh, 2015; Unkelbach & Stahl, 2009). Building on this work, the model we used in the present study allowed us to explore whether explicit warnings about the truth effect would affect (1) participants' reliance on processing fluency, (2) their retrieval of relevant knowledge, or (3) their guessing behavior when asked to judge the validity of a statement.

In addition to conducting a model-based analysis of the effects of warnings on the truth effect, we also wanted to explore why warnings have previously only reduced the effect but not eliminated it. As discussed by Nadarevic and Aßfalg (2017), it could be argued that participants may simply not have been motivated enough to sufficiently control their judgments of truth (cf. Wilson & Brekke, 1994). We tried to address this concern by using a financial incentive to increase participants' motivation to provide accurate judgments of truth and resist the credibility-enhancing effect of repetition. Moreover, we adapted the experimental procedure used by Nadarevic and Aßfalg by presenting participants not only with statements of relatively unknown validity but also with statements for which participants could likely identify the correct truth status considerably better than chance. This modification increased the ecological validity of the experimental setup, as it is highly unlikely that people exclusively encounter maximally ambiguous claims in their natural environment. Arguably, participants may also be able to counter the truth effect more successfully if they can draw on their stored knowledge instead of relying on heuristic cues such as processing fluency when evaluating a claim's validity (cf. Nadarevic & Aßfalg, 2017).

Method

Materials

To collect data for a model-based investigation of the effects of warnings on the truth effect, we set up a standard truth effect experiment in which 40 trivia statements served as stimuli. These statements were taken from a larger collection of statements that had been judged for truth by a student sample as part of an unpublished and unrelated study conducted by the second author at the University of Mannheim. We selected ten false statements with a proportion of "true" judgments (PTJ) between .09 and .26 and ten true statements with a PTJ between .76 and .93. These twenty statements were used as "easy" items in the present experiment because the majority of pretest participants had identified their truth status correctly. In addition, we selected ten true and ten false statements with a PTJ between .33 and .70. These statements served as “difficult” items in the present experiment because the truth status of these statements was considerably more difficult to determine.

We assigned the selected statements to two stimulus sets A and B, so that each set contained 5 easy false statements, 5 easy true statements, 5 difficult false statements, and 5 difficult true statements. Sets A and B were matched with regard to their mean PTJs (M_{A} = .49, M_{B} = .50) and with regard to the standard deviation of the PTJs (SD_{A} = .26, SD_{B} = .25). Moreover, special care was taken to match the two item sets with regard to mean PTJ within item categories (easy and true: M_{A} = .82, M_{B} = .84; easy and false: M_{A} = .16, M_{B} = .18; difficult and true: M_{A} = .54, M_{B} = .53; difficult and false: M_{A} = .42, M_{B} = .44).

Procedure & Design

The experiment was conducted in the laboratory and comprised an exposure phase, a testing phase and an additional knowledge check phase. Participants completed all three phases within a single session. After participants entered the laboratory, they were seated in a cubicle and started the computer-based experiment. Participants first provided demographic data and were randomly assigned to the warning or the control condition. In the subsequent exposure phase, participants in both conditions were informed that they were going to see a collection of true and false statements and that it would be their task to assign each statement to one of six different categories of knowledge (geography, flora & fauna, politics & history, science, sports & entertainment, other). Before they were allowed to proceed, all participants had to correctly answer a multiple-choice question testing their understanding of the task to make sure that they had read the instructions carefully. If they failed to provide the correct answer, participants were allowed to reread the instructions and were then asked to answer the question again. Next, the 20 statements from one of the two stimulus sets (i.e., A or B) were presented sequentially and in random order, and had to be assigned to the six categories of knowledge. Participants were thus familiarized with one of the two sets of trivia statements during the initial exposure phase.

Immediately after participants had categorized the 20 statements, the testing phase started. Participants in both conditions were informed that they would again be presented with true and with false statements and that this time, these statements should be judged as either *true* or *false*. Furthermore, participants were told that the statements were sampled from a large collection of statements, and that they might therefore encounter some statements they were already familiar with. Importantly, only the instructions in the warning condition contained an additional paragraph that informed participants about the truth effect and asked them to prevent it. This text, which was adapted from Nadarevic and Aßfalg (2017), read as follows (translated from German):

> Attention: Some of the statements that will be presented on the upcoming pages have already been shown in the first phase of the study. Other statements are new, which means that you have not encountered them before. Some statements appear repeatedly in this study because we are interested in the "illusion of truth". The illusion of truth denotes people's tendency to judge repeated statements as "true" more often than new statements, irrespective of the statements' actual truth status.
>
> In the upcoming task, half of the repeated statements as well as half of the new statements are true. Therefore, a statement's truth status is independent of whether the statement has been presented before or not. Thus, the illusion of truth would lead to errors in the upcoming task. You should therefore try to judge the truth of all statements as accurately as possible. Prevent the illusion of truth and try not to be influenced by whether or not you have encountered a statement before.

To make sure that participants in the warning condition were sufficiently motivated to prevent the truth effect, we provided them with the prospect of a substantial financial reward in return for accurate and unbiased judgments of truth:

> Please work carefully; your effort will be rewarded. For each true and for each false statement that you correctly classify as either true or false, you will earn one point. Upon completion of this study, we will award 30 euros to each of the 10 participants scoring the most points.

The two paragraphs presented above were surrounded by a red frame to graphically highlight their importance. After being introduced to their task, participants in the warning condition had to correctly answer two multiple-choice questions to make sure that they had understood the instructions. Specifically, the following questions were presented right underneath the warning text: "*What is the illusion of truth?*" and "*What is your task in the second part of the study?*". If participants failed to provide the correct answer for at least one of the two questions, the following error message was displayed:

> At least one of your answers was incorrect. Please remember that the illusion of truth denotes people's tendency to judge repeated statements as true more often than new statements, irrespective of the statements' actual truth status. Because the illusion of truth will lead to errors in the upcoming task, you should judge the truth of all presented statements as accurately as possible while at the same time not being influenced by whether or not you have encountered a statement before.

Participants in the warning condition were only allowed to proceed after they had provided the correct answers to both questions. After the instructions, warned and unwarned participants were presented with the trivia statements from both stimulus sets (i.e., Sets A and B), half of which had already been shown in the initial exposure phase. All 40 statements were presented sequentially and in random order. Participants had to classify each statement as either true or false.

The testing phase was followed by a knowledge check phase. This phase served as a manipulation check to make sure that participants were indeed more knowledgeable about the preselected “easy” statements than about the preselected “difficult” statements. To this end, participants answered 40 multiple-choice questions, each referring to one of the statements that had been shown during the study. Each question was presented along with three response options: the correct answer, a distractor and a “don’t know” answer option. If, for example, participants had encountered the true statement “Bratislava is the capital of Slovakia” in the testing phase, the question “Of which country is Bratislava the capital?” was presented along with the three response options “Slovakia”, “Slovenia” and “don’t know” in the knowledge check phase. If a false statement like “Stonehenge is located in the county of Berkshire” had been presented before, the question “What county is Stonehenge located in?” was presented along with the response options “Wiltshire”, “Berkshire” and “don’t know”.

In a post-experimental questionnaire, participants in both conditions were asked to indicate whether they had participated seriously (Aust, Diedenhofen, Ullrich, & Musch, 2013). Warned participants were additionally provided with several questions referring to the truth effect. Most importantly, participants had to indicate whether they had tried to prevent the effect. Participants were also asked to specify whether they thought that they had succeeded in doing so on a rating scale ranging from 1 (*not at all successful*) to 4 (*very successful*). Moreover, warned participants were prompted to describe their strategy for preventing the truth effect in written form, using their keyboard as input device. After completing the post-experimental questionnaire, all participants were thanked and debriefed. The median study duration was 16.9 minutes, with an interquartile range of 14.5 to 20.1 minutes.

In the testing phase, warned and unwarned participants judged the truth of both repeated and new statements. Each statement was either easy or difficult and either true or false. Thus, our experimental design comprised the between-subjects factor instruction (warning vs. control) as well as the three within-subjects factors repetition status (repeated vs. new), truth status (true vs. false) and difficulty (easy vs. difficult). To control for stimulus-specific effects, each participant was randomly assigned to one of two counterbalancing conditions. During the exposure phase, participants in the first counterbalancing condition assigned statements in Set A to different categories of knowledge, whereas participants in the second counterbalancing condition completed the same task for statements in Set B. In the testing phase, participants in both counterbalancing conditions then judged the validity of statements in both the repeated Set A (B) and the unrepeated Set B (A). The final knowledge check phase was identical for both counterbalancing groups in that all participants answered questions referring to statements in both sets.

Participants

Participants were recruited on the campus of the University of Duesseldorf. One hundred and seventy-one participants completed the study. We discarded the data from four participants in the warning condition because they indicated in the post-experimental questionnaire that they had not tried to prevent the truth effect.[1] The final sample therefore consisted of 167 participants (n_{warning} = 82; n_{control} = 85), all of whom indicated having taken part seriously. One hundred and twenty-nine participants were female (77%). Participants' ages ranged from 18 to 40 years (*M* = 22.63, *SD* = 4.49) and the majority of participants were native speakers of German (86%). One hundred and twenty-three participants were students of psychology (74%), 43 participants studied another subject and one person indicated not being a student. All psychology students received course credit for their participation.

Our primary interest was in the interaction of the between-subjects factor instruction (warning vs. control) and the within-subjects factor repetition (repeated vs. new). We therefore conducted a sensitivity analysis to determine the size of this interaction effect, which our study was able to detect with a power of $1 - \beta = .80$, given the sample size of $N = 167$, $\alpha = .05$ and a correlation between repeated measurements of .15.[2] This analysis, conducted using G*Power 3.1 (Faul, Erdfelder, Lang, & Buchner, 2007), indicated that we were able to detect an effect equal to or larger than $f = 0.14$, which corresponds to a fairly small effect according to the classification by Cohen (1988). Thus, our study was adequately powered to detect effects even smaller than those found in previous experiments investigating the effects of warnings on the truth effect ($f \geq 0.20$, cf. Nadarevic & Aßfalg, 2017).

[1] However, not excluding these participants did not change any of the results reported below.

[2] This was the size of the correlation between the proportion of "true" judgments (PTJ) for repeated statements and the PTJ for new statements that we observed in the present experiment.

Results

We used R (Version 3.5.0; R Core Team, 2018) and the R-packages *afex* (Version 0.22.1; Singmann, Bolker, Westfall, & Aust, 2018), *ggplot2* (Version 3.1.0; Wickham, 2016), *papaja* (Version 0.1.0.9709; Aust & Barth, 2018), *plyr* (Version 1.8.4; Wickham, 2011), and *reshape2* (Version 1.4.3; Wickham, 2007) for the majority of our analyses. Model-based analyses based on multinomial processing trees were conducted using the program multiTree (Moshagen, 2010).

Knowledge check

To make sure that we had selected an adequate sample of easy and difficult statements, we first analyzed the data from the knowledge check phase. For each participant, we computed the percentage of questions that had been answered correctly. Because we were interested in comparing participants' performance for questions referring to presumably easy and presumably difficult statements, percentages were calculated separately for both item types. On average, participants correctly answered a higher percentage of multiple-choice questions referring to presumably easy statements (69%) than referring to presumably difficult statements (22%), $t(166) = 34.81$, $p < .001$, $d_z = 2.69$. We also determined the percentage of questions for which participants selected the "don't know" response option. On average, this option was chosen more frequently for questions referring to presumably difficult statements (61%) than for questions referring to presumably easy statements (22%), $t(166) = -27.26$, $p < .001$, $d_z = 2.11$. Taken together, the data from the knowledge check phase confirmed that participants had no profound knowledge of the presumably difficult statements, but considerable knowledge regarding the presumably easy statements used in the present study.

Judgments of truth: ANOVA-based approach

To analyze participants' binary truth judgments from the testing phase using an ANOVA-based approach, we calculated the proportion of "true" judgments (PTJ) as the dependent variable. Comparing the mean PTJs between the two counterbalancing conditions using a t-test for independent samples showed no significant difference, $t(165) = 0.32$, $p = .746$, $d = 0.05$. We therefore pooled across both counterbalancing conditions in the following analyses.

To analyze the effects of warnings on the truth effect, we conducted a 2 (instructions: warning vs. control) x 2 (repetition status: repeated vs. new) x 2 (difficulty: easy vs. difficult) x 2 (truth status: true vs. false) mixed factorial ANOVA using the PTJ as the dependent variable. Not surprisingly, factually true statements were judged to be true significantly more often ($M = .70$, $SD = .11$) than false statements ($M = .44$, $SD = .14$), $F(1, 165) = 398.04$, $p < .001$, $\eta_p^2 = .71$. As expected, this main effect of truth status was qualified by an interaction of truth status and difficulty: For difficult statements, participants were unable to discriminate between true and false statements, as they judged an approximately equal number of true ($M = .56$, $SD = .18$) and false statements ($M = .59$, $SD = .18$) to be true. For easy statements, however, PTJs were substantially higher for true ($M = .84$, $SD = .13$) than for false statements ($M = .30$, $SD = .18$). The interaction of truth status and difficulty was significant, $F(1, 165) = 588.19$, $p < .001$, $\eta_p^2 = .78$, indicating that participants were able to retrieve relevant knowledge for easy but not for difficult statements. This finding corroborates the results obtained for the knowledge check data. More importantly, however, there also was a significant main effect of repetition status, $F(1, 165) = 29.81$, $p < .001$, $\eta_p^2 = .15$, indicating a truth effect. That is, participants judged repeated statements to be true more often ($M = .61$, $SD = .14$) than new statements ($M = .54$, $SD = .12$). This main effect was qualified by a significant interaction of instruction and repetition status, *F*(1, 165) = 4.90, *p* = .028, η_p^2 = .03. As illustrated in Figure 1, the truth effect in the warning condition was substantially smaller than the truth effect in the control condition (d_z = 0.27 vs. d_z = 0.57). However, warnings did not completely eliminate the truth effect, as warned participants still accepted significantly more repeated statements (*M* = .58, *SD* = .12) than new statements (*M* = .54, *SD* = .11), *t*(81) = 2.41, *p* = .018, d_z = 0.27. There also was a significant four-way interaction, *F*(1, 165) = 4.29, *p* = .040, η_p^2 = .03. To explore the pattern of this interaction, we calculated the size of the truth effect for each type of item (easy and false, easy and true, difficult and false, difficult and true) as the difference in PTJs between repeated and new statements. As can be seen from the results displayed in Figure 2, the effect of the warning manipulation was particularly strong for easy statements that were false.

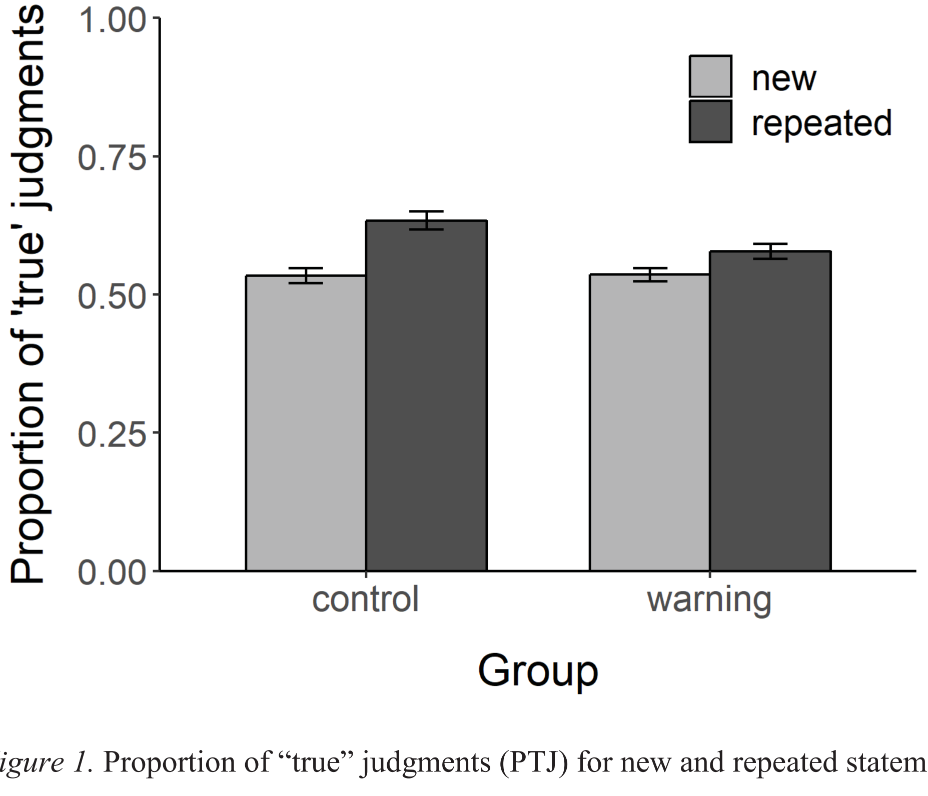

Figure 1. Proportion of "true" judgments (PTJ) for new and repeated statements in the control condition and in the warning condition. Error bars represent ± 1 standard error.

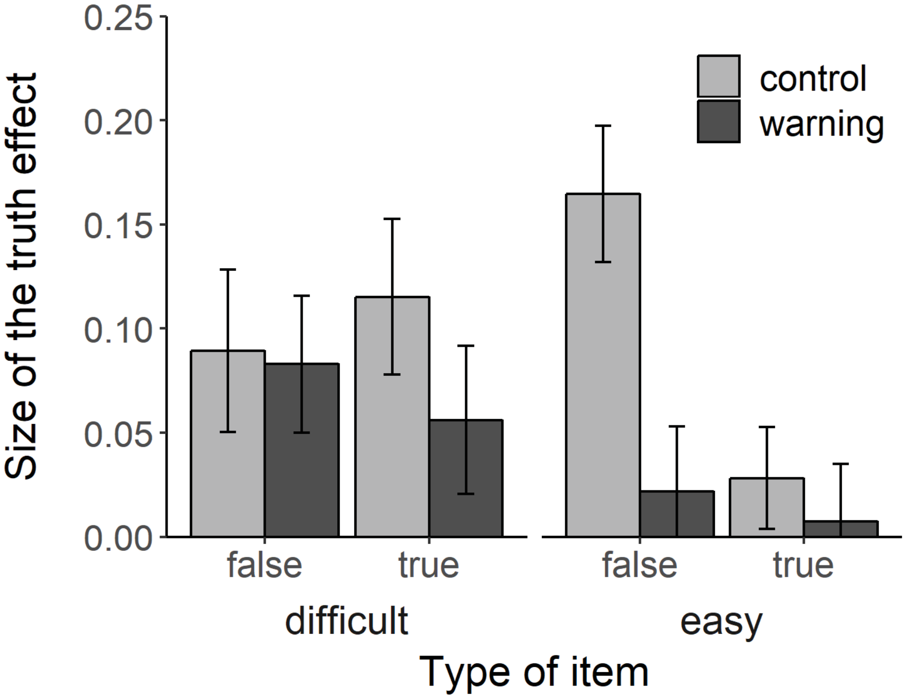

Figure 2. Size of the truth effect by item type in the control and in the warning condition. The size of the truth effect is operationalized as the difference between the proportion of “true” judgments (PTJ) for repeated and new statements. Error bars represent ± 1 standard error.

Judgments of truth: Multinomial modeling approach

Because the main goal of our study was to investigate how warnings influence the cognitive processes that are assumed to underlie judgments of truth, we also subjected our data to a model-based analysis. Recently, Fazio et al. (2015) introduced the *fluency-conditional model*, a multinomial model that allows to disentangle three different cognitive processes that presumably underlie truth judgments. Specifically, the model estimates the probabilities that a person relies on processing fluency (*F*-parameter), retrieves the relevant knowledge (*K*-parameter) or simply guesses “true” or “false” (*G*-parameter) when asked to judge a statement’s truth. For the following analyses, we used a parsimonious variant of the fluency-conditional model that is illustrated in Figure 3. This variant incorporates a fluency parameter for repeated statements only because processing fluency should be present for repeated statements, but not for new statements. Following Fazio and colleagues, the knowledge parameter was estimated separately for easy and difficult statements. Moreover, we estimated all parameters separately for the warning and the control condition. The parsimonious model fit the data well, $G^2(8)$ = 10.78, p = .215. Parameter estimates are displayed in Table 1.

As expected, the probability of knowledge retrieval was substantially higher for easy than for difficult statements; constraining the knowledge parameters to be equal for easy and difficult statements therefore resulted in a significant deterioration of model fit, $\Delta G^2(2)$ = 597.87, p < .001. This result provides evidence for the validity of the knowledge parameters. In a next step, we compared parameters between the warning condition and the control condition. The probability of knowledge retrieval for difficult statements did not differ between the two conditions as both parameters were estimated to be zero. Knowledge retrieval for easy statements also did not differ between the warning condition and the control condition, $\Delta G^2(1)$ = 0.21, p = .645. Moreover, the warning manipulation did not affect participants' guessing behavior, $\Delta G^2(1)$ < 0.01, p = .955. In the control condition as well as in the warning condition, guessing parameters were close to .50, thus mirroring the actual base rate of true statements. Most importantly, the warning instruction was found to affect participants' tendency to rely on processing fluency. Model fit worsened significantly when constraining the processing fluency parameter to be equal in the warning condition and in the control condition, $\Delta G^2(1)$ = 10.15, p = .001, supporting the notion that warned participants less frequently used fluency as a cue for truth than participants in the control condition. Taken together, the model-based analysis showed that warnings significantly decreased reliance on processing fluency when participants were asked to evaluate the truth of repeated statements. However, warning instructions did not affect participants' retrieval of knowledge or guessing behavior relative to the control condition.

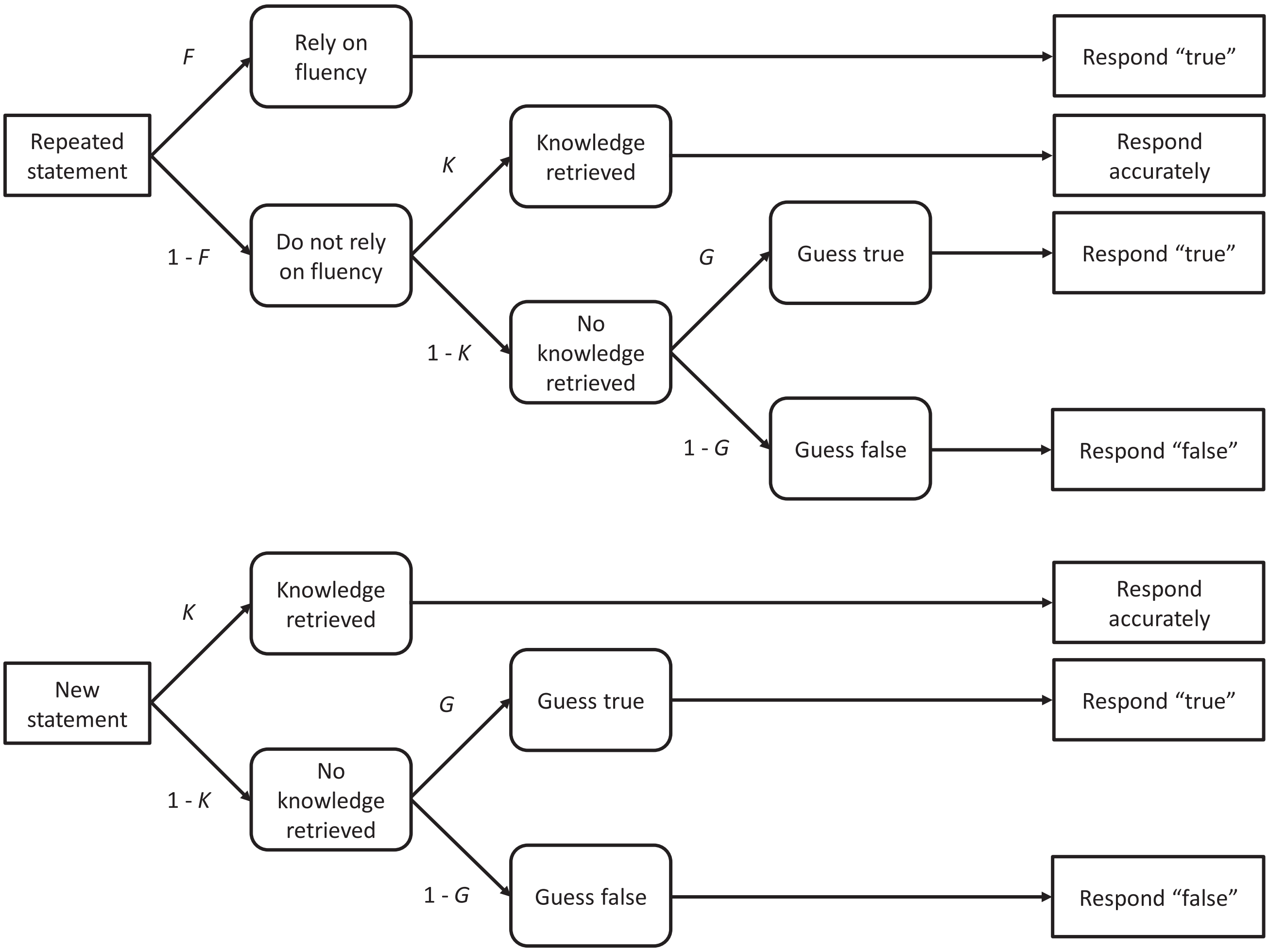

Figure 3. Multinomial model of the truth effect. Judgments of truth are expected to be determined by reliance on processing fluency (*F*), knowledge retrieval (*K*), and guessing (*G*).

Table 1

Multinomial model parameter estimates (with standard errors in parentheses) for the control condition and the warning condition

Condition	F	K_{easy}	$K_{\text{difficult}}$	G
Control	.22 (.03)	.58 (.02)	.00 (.03)	.55 (.01)
Warning	.09 (.03)	.59 (.02)	.00 (.03)	.55 (.01)

Note. F = probability of reliance on processing fluency, K_{easy} = probability of knowledge retrieval for easy statements, $K_{\text{difficult}}$ = probability of knowledge retrieval for difficult statements, G = probability of guessing “true”.

Post-experimental questionnaire

In the post-experimental questionnaire, participants in the warning condition were asked to describe the strategy they had used to prevent the truth effect during the testing phase. A large number of participants (38%) reported that they had tried to prevent the truth effect by relying on previous knowledge, logic or the plausibility of statements. Another group of participants (23%) indicated that they had explicitly considered the repetition status of statements when evaluating their truth status. These participants reported that they had treated repeated statements in a special way, for example by shifting their response criterion when judging the statements’ truth. Several other participants (20%) indicated that they had tried to ignore the repetition status of statements altogether. A fourth group of participants (10%) reported other strategies, such as reading every statement twice or taking plenty of time to judge the perceived truth of each statement. The remaining participants (10%) either indicated that they had not used any particular strategy to prevent the truth effect or provided no meaningful answer to the question.

Participants in the warning condition were also asked to indicate whether they thought that they had been successful in preventing the truth effect. Most participants (63%) believed they had been “rather successful” in countering the credibility-enhancing effect of repetition; two participants (2%) even believed that they had been “very successful” in doing so. However, almost a third of the warned participants (29%) were convinced that they had been rather unsuccessful in preventing the truth effect; four participants (5%) indicated not having succeeded at all. Interestingly, participants’ estimated success in preventing the truth effect was not associated with the size of the truth effect on an individual level, $r = -.14$, $p = .220$.

Discussion

Although the truth effect has been studied for more than four decades, astonishingly little is known about potential means to avoid the influence of repetition on perceived truth. In a recent study, Nadarevic and Aßfalg (2017) observed that explicit warnings significantly reduced the truth effect but did not eliminate it. Based on the assumption that the truth effect occurs due to people’s tendency to use processing fluency as a cue for truth (Unkelbach, 2007), Nadarevic and Aßfalg interpreted their results as evidence “that fluency-truth attributions can be controlled to some degree” (p. 823). However, model-based analyses are needed to investigate how exactly a warning affects the processes that presumably underlie judgments of truth. Extending previous studies, we therefore used a multinomial modeling approach that allowed us to estimate the probabilities with which warned and unwarned participants relied on (a) processing fluency, (b) previous knowledge and (c) guessing when evaluating the truth of statements. Comparing these parameters between the warning condition and the control condition revealed that warnings significantly reduced participants' probability of relying on processing fluency when evaluating the truth of repeated statements. Warnings, however, did not affect participants' guessing behavior or the probability with which they retrieved stored knowledge. This pattern of results is consistent with the notion that people are capable of discounting feelings of processing ease when their informational value is questioned (Alter & Oppenheimer, 2009; Schwarz, 2004). The results of the present study also replicate Nadarevic and Aßfalg's core findings: Warning people about the truth effect is sufficient to significantly reduce, but not completely eliminate the effect. Our results show that this even holds true when participants are given the prospect of a substantial financial incentive in return for accurate and unbiased judgments of truth.

Wilson and Brekke (1994) provided potential reasons for why the mere presence of a warning might not suffice to eliminate mental biases. They argued that for successful control of unwanted and biasing influences on judgment, several conditions have to be met. For example, people have to be aware of a potential bias, they have to know about its size and direction, and they have to be motivated to correct for its influence. Moreover, people need to have sufficient control over the relevant mental processes to successfully counter a bias. At best, they should be provided with a clear-cut strategy to help them counter unwanted influences on their judgment. In our experiment, we used warning instructions to make participants aware of the origin and nature of the truth effect. Moreover, we tried to ensure that participants were highly motivated to counter the effect by providing them with the prospect of a substantial financial reward in return for unbiased and accurate judgments of truth. However, even though participants were made aware of the truth effect and were motivated to combat its biasing influence, they might not have had the necessary means to fully counter the effect of repetition on judged truth. In our post-experimental questionnaire, we asked participants in the warning condition to describe the strategy they had used to prevent the truth effect. The variety of strategies reported suggests that there is no clear-cut, obvious approach to counter the effect of repetition on judged truth. This might have been one of the reasons why warnings did not successfully eliminate the truth effect in the present study. Moreover, the use of processing fluency as a source of information has been argued to occur in a largely automatic manner (Schwarz, 2004). Assuming that warnings operate through a deliberate discounting process, such a process would have to be employed reliably and consistently to fully overcome any automatic influence of processing fluency on judged truth.

Future studies should address some of the limitations of the present experiment. For example, in our study the testing phase immediately followed the exposure phase, meaning that participants likely remembered which statements had been presented in the exposure phase. Future studies should therefore investigate whether the efficacy of warnings for reducing the truth effect is moderated by the length of the retention interval. Currently, only the study by Nadarevic and Aßfalg (2017) addresses this issue. A comparison of the effect sizes in their two experiments suggests that warnings are equally effective in reducing the truth effect regardless of whether the warning instructions and the truth judgment phase immediately follow the exposure phase (Exp. 2) or take place one week after the exposure phase (Exp. 1). However, a more thorough investigation of the role of the retention interval seems warranted for two reasons: First, due to a lack of power, the effect of the warning instructions on the truth effect was not significant in the first experiment by Nadarevic and Aßfalg; second, an across-experiment comparison cannot replace a direct experimental manipulation of the length of the retention interval.

False information and “fake news” are quite common in everyday life and can be encountered in many places, including social media platforms on the internet (Allcott &

Gentzkow, 2017; Lazer et al., 2018). It is therefore of considerable political and societal importance to investigate how detrimental effects of the repeated exposure to misinformation can be avoided. Research on the truth effect has primarily explored two different approaches to trying to counter the influence of repetition on the perceived truth of (false) claims. The first approach relies on providing feedback on the actual truth status of statements. However, explicitly labeling statements as false during their first presentation does not seem to be enough to outweigh the effects of repetition, at least not in the long run (e.g., Brown & Nix, 1996; Skurnik et al., 2005). Moreover, a recent study by Pennycook et al. (2018) has shown that simply questioning the credibility of fake-news headlines by flagging them as “disputed by 3rd party fact-checkers” may not be an effective way of combating the truth effect either. In contrast to this line of research, the present study chose not to question the credibility of statements per se, but instead explicitly warned participants that repetition leads to an increase in the perceived validity of statements. Replicating Nadarevic and Aßfalg (2017), this approach substantially reduced the truth effect, but did not eliminate it completely. Therefore, combatting the truth effect in real-world settings, where “information rarely comes with warning labels” (Wilson & Brekke, 1994, p. 135), might be harder than expected. Given that the truth effect seems to be resilient to debiasing approaches to a considerable degree, it might be more fruitful to harness the credibility-enhancing effect of repetition instead of fighting it. People trying to combat fake news might best be advised to counter misinformation by repeatedly promoting true information that should best be presented in a way that makes it easy to process and understand (Cook & Lewandowsky, 2011). By pursuing this cumbersome approach, truth might prevail in the end—if it is repeated often enough.

References

Allcott, H., & Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, *31*(2), 211–236. doi:10.1257/jep.31.2.211

Alter, A. L., & Oppenheimer, D. M. (2009). Uniting the tribes of fluency to form a metacognitive nation. *Personality and Social Psychology Review*, *13*, 219–235. doi:10.1177/1088868309341564

Aust, F., & Barth, M. (2018). *papaja: Create APA manuscripts with R Markdown*. Retrieved from https://github.com/crsh/papaja

Aust, F., Diedenhofen, B., Ullrich, S., & Musch, J. (2013). Seriousness checks are useful to improve data validity in online research. *Behavior Research Methods*, *45*, 527–535. doi:10.3758/s13428-012-0265-2

Baron, J., & Hershey, J. C. (1988). Outcome bias in decision evaluation. *Journal of Personality and Social Psychology*, *54*, 569–579. doi:10.1037/0022-3514.54.4.569

Bernstein, D. M., Wilson, A. M., Pernat, N. L. M., & Meilleur, L. R. (2012). Auditory hindsight bias. *Psychonomic Bulletin & Review*, *19*, 588–593. doi:10.3758/s13423-012-0268-0

Blank, H., & Launay, C. (2014). How to protect eyewitness memory against the misinformation effect: A meta-analysis of post-warning studies. *Journal of Applied Research in Memory and Cognition*, *3*, 77–88. doi:10.1016/j.jarmac.2014.03.005

Boehm, L. E. (1994). The validity effect: A search for mediating variables. *Personality and Social Psychology Bulletin*, *20*, 285–293. doi:10.1177/0146167294203006

Brown, A. S., & Nix, L. A. (1996). Turning lies into truths: Referential validation of falsehoods. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 1088–1100. doi:10.1037/0278-7393.22.5.1088

Clarkson, P. M., Emby, C., & Watt, V. W.-S. (2002). Debiasing the outcome effect: The role of instructions in an audit litigation setting. *Auditing: A Journal of Practice & Theory*, *21*(2), 7–20. doi:10.2308/aud.2002.21.2.7

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Cook, J., & Lewandowsky, S. (2011). *The Debunking Handbook*. Retrieved from https://skepticalscience.com/docs/Debunking_Handbook.pdf

Davies, M. F. (1993). Field-dependence and hindsight bias: Output interference in the generation of reasons. *Journal of Research in Personality*, *27*, 222–237. doi:10.1006/jrpe.1993.1016

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2009). Mix me a list: Context moderates the truth effect and the mere-exposure effect. *Journal of Experimental Social Psychology*, *45*, 1117–1122. doi:10.1016/j.jesp.2009.06.019

Dechêne, A., Stahl, C., Hansen, J., & Wänke, M. (2010). The truth about the truth: A meta-analytic review of the truth effect. *Personality and Social Psychology Review*, *14*, 238–257. doi:10.1177/1088868309352251

Doland, C. A. (1999). *Repeating is believing: An investigation of the illusory truth effect* (Unpublished doctoral dissertation). State University of New York at Albany, New York.

Egan, J. (2007). *Marketing communications*. Andover, England: Cengage Learning.

Erdfelder, E., Auer, T.-S., Hilbig, B. E., Aßfalg, A., Moshagen, M., & Nadarevic, L. (2009). Multinomial processing tree models: A review of the literature. *Zeitschrift für Psychologie / Journal of Psychology*, *217*, 108–124. doi:10.1027/0044-3409.217.3.108

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*, 175–191. doi:10.3758/BF03193146

Fazio, L. K., Brashier, N. M., Payne, B. K., & Marsh, E. J. (2015). Knowledge does not protect against illusory truth. *Journal of Experimental Psychology: General*, *144*, 993–1002. doi:10.1037/xge0000098

Fischhoff, B. (1977). Perceived informativeness of facts. *Journal of Experimental Psychology: Human Perception and Performance*, *3*, 349–358. doi:10.1037/0096-1523.3.2.349

Gigerenzer, G. (1984). External validity of laboratory experiments: The frequency-validity relationship. *American Journal of Psychology*, *97*, 185–195. doi:10.2307/1422594

Grenier, J. H., Peecher, M. E., & Piercey, M. D. (2007). *Judging auditor negligence: De-biasing interventions, outcome bias, and reverse outcome bias*. Retrieved from https://ssrn.com/abstract=1015523

Harley, E. M., Carlsen, K. A., & Loftus, G. R. (2004). The “saw-it-all-along” effect: Demonstrations of visual hindsight bias. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 960–968. doi:10.1037/0278-7393.30.5.960

Hasher, L., Goldstein, D., & Toppino, T. (1977). Frequency and the conference of referential validity. *Journal of Verbal Learning and Verbal Behavior*, *16*, 107–112. doi:10.1016/S0022-5371(77)80012-1

Hawkins, S. A., & Hoch, S. J. (1992). Low-involvement learning: Memory without evaluation. *Journal of Consumer Research*, *19*, 212–225. doi:10.1086/209297

Hertwig, R., Gigerenzer, G., & Hoffrage, U. (1997). The reiteration effect in hindsight bias. *Psychological Review*, *104*, 194–202. doi:10.1037/0033-295X.104.1.194

Johar, G. V., & Roggeveen, A. L. (2007). Changing false beliefs from repeated advertising: The role of claim-refutation alignment. *Journal of Consumer Psychology*, *17*, 118–127. doi:10.1016/S1057-7408(07)70018-9

Law, S., Hawkins, S. A., & Craik, F. I. M. (1998). Repetition-induced belief in the elderly: Rehabilitating age-related memory deficits. *Journal of Consumer Research*, *25*, 91–107. doi:10.1086/209529

Lazer, D. M. J., Baum, M. A., Benkler, Y., Berinsky, A. J., Greenhill, K. M., Menczer, F., … Zittrain, J. L. (2018). The science of fake news. *Science*, *359*, 1094–1096. doi:10.1126/science.aao2998

Lilienfeld, S. O., Lynn, S. J., Ruscio, J., & Beyerstein, B. L. (2010). *50 great myths of popular psychology: Shattering widespread misconceptions about human behavior*. Chichester, England: Wiley-Blackwell.

Loftus, E. F., Miller, D. G., & Burns, H. J. (1978). Semantic integration of verbal information into a visual memory. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 19–31. doi:10.1037/0278-7393.4.1.19

McGlone, M. S., & Tofighbakhsh, J. (2000). Birds of a feather flock conjointly (?): Rhyme as reason in aphorisms. *Psychological Science*, *11*, 424–428. doi:10.1111/1467-9280.00282

Moshagen, M. (2010). MultiTree: A computer program for the analysis of multinomial processing tree models. *Behavior Research Methods*, *42*, 42–54. doi:10.3758/BRM.42.1.42

Mutter, S. A., Lindsey, S. E., & Pliske, R. M. (1995). Aging and credibility judgment. *Aging and Cognition*, *2*, 89–107. doi:10.1080/13825589508256590

Nadarevic, L., & Aßfalg, A. (2017). Unveiling the truth: Warnings reduce the repetition-based truth effect. *Psychological Research*, *81*, 814–826. doi:10.1007/s00426-016-0777-y

Oeberst, A., & Blank, H. (2012). Undoing suggestive influence on memory: The reversibility of the eyewitness misinformation effect. *Cognition*, *125*, 141–159. doi:10.1016/j.cognition.2012.07.009

Pennycook, G., Cannon, T. D., & Rand, D. G. (2018). Prior exposure increases perceived accuracy of fake news. *Journal of Experimental Psychology: General*, *147*, 1865–1880. doi:10.1037/xge0000465

Pohl, R. F., & Hell, W. (1996). No reduction in hindsight bias after complete information and repeated testing. *Organizational Behavior and Human Decision Processes*, *67*, 49–58. doi:10.1006/obhd.1996.0064

R Core Team. (2018). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Reber, R., & Schwarz, N. (1999). Effects of perceptual fluency on judgments of truth. *Consciousness and Cognition*, *8*, 338–342. doi:10.1006/ccog.1999.0386

Roggeveen, A. L., & Johar, G. V. (2002). Perceived source variability versus familiarity: Testing competing explanations for the truth effect. *Journal of Consumer Psychology*, *12*, 81–91. doi:10.1207/S15327663JCP1202_02

Schwarz, N. (2004). Metacognitive experiences in consumer judgment and decision making. *Journal of Consumer Psychology*, *14*, 332–348. doi:10.1207/s15327663jcp1404_2

Sharpe, D., & Adair, J. G. (1993). Reversibility of the hindsight bias: Manipulation of experimental demands. *Organizational Behavior and Human Decision Processes*, *56*, 233–245. doi:10.1006/obhd.1993.1053

Singmann, H., Bolker, B., Westfall, J., & Aust, F. (2018). *afex: Analysis of factorial experiments*. Retrieved from https://CRAN.R-project.org/package=afex

Skurnik, I., Yoon, C., Park, D. C., & Schwarz, N. (2005). How warnings about false claims become recommendations. *Journal of Consumer Research*, *31*, 713–724. doi:10.1086/426605

Unkelbach, C. (2007). Reversing the truth effect: Learning the interpretation of processing fluency in judgments of truth. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 219–230. doi:10.1037/0278-7393.33.1.219

Unkelbach, C., & Stahl, C. (2009). A multinomial modeling approach to dissociate different components of the truth effect. *Consciousness and Cognition*, *18*, 22–38. doi:10.1016/j.concog.2008.09.006

Wickham, H. (2007). Reshaping data with the reshape package. *Journal of Statistical Software*, *21*(12), 1–20. doi:10.18637/jss.v021.i12

Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29. doi:10.18637/jss.v040.i01

Wickham, H. (2016). *Ggplot2: Elegant graphics for data analysis*. New York, NY: Springer.

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

Wilson, T. D., Centerbar, D. B., & Brekke, N. (2002). Mental contamination and the debiasing problem. In T. Gilovich, D. Griffin, & D. Kahneman (Eds.), *Heuristics and biases: The psychology of intuitive judgment* (pp. 185–200). New York, NY: Cambridge University Press.

Eidesstattliche Versicherung

Ich versichere an Eides statt, dass die Dissertation von mir selbständig und ohne unzulässige fremde Hilfe unter Beachtung der „Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der Heinrich-Heine-Universität Düsseldorf“ erstellt worden ist. Ferner versichere ich, dass die Arbeit in der vorgelegten oder in ähnlicher Form bisher bei keiner anderen Fakultät als Dissertation eingereicht wurde. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, den 05.02.2019

Frank Calio