The Perseverance of Adverse Experimental Effects After Post-Experimental Debriefings and Interventions

---

Doctoral Thesis

Submitted in Fulfillment of the Requirements for the Degree

Doctor of Natural Sciences

Faculty of Human and Business Sciences

Saarland University

---

by

Stefanie Manuela Miketta

Saarbrücken, Germany

January 2018

Dekan:

Univ.-Prof. Dr. Cornelius König

Berichterstatter:

Univ.-Prof. Dr. Malte Friese

Univ.-Prof. Dr. Cornelius König

Tag der Disputation: 22.06.2018

Summary

Psychologists often use highly impactful experimental manipulations that are designed to temporarily reduce participants' well-being, e.g., by threatening their self-worth. These manipulations can be crucial for experimentally answering certain research questions. At the same time, researchers have long been concerned about participants' well-being after those manipulations. Official ethical guidelines of several psychological societies even emphasize that investigators are *obliged* to protect participants from harm. To solve the dilemma of using harmful manipulations while not wanting to harm participants, researchers usually provide a debriefing at the end of each experiment to eliminate detrimental effects of the manipulation.

Little is known, however, about the effectiveness of post-experimental debriefings despite the fact that modern psychological research widely relies on it. Findings from a different line of research suggest that experimentally induced beliefs can persevere even after the underlying information is discredited. Given that perseverance of already discredited beliefs is possible, the question arises whether adverse effects caused by experimental manipulations can also persevere – even if post-experimental debriefings seek to undo them. This question seems particularly warranted since the majority of debriefings simply retract false information provided earlier. The discrepancy between the field's trust in the effectiveness of debriefings on the one hand and the scarcity of empirical evidence for this assumption on the other hand calls for thorough research addressing the question of potential perseverance of adverse experimental effects.

The present thesis investigates this question. Part I examines the effectiveness of post-experimental debriefings and interventions in eliminating the adverse effects of well-established ego threat manipulations on psychological well-being and hostile attributions. As expected, participants who received negative feedback compared to positive or no feedback reported lower well-being and showed more hostile attributions. Our results suggest that debriefing procedures failed to undo the harm caused by ego threat manipulations. Instead,

participants left the laboratory with unwanted aftereffects that persisted for one day and longer. An internal meta-analysis confirmed these findings.

Part II focuses on a different, equally well-established manipulation: experimentally induced ostracism, which is known for affecting various indicators of well-being in a negative way. After having been excluded as opposed to included by ostensible other participants in a virtual ball-tossing game, participants reported decreased positive and increased negative mood. A *Revised Outcome Debriefing*, the allegedly most effective kind of post-experimental debriefing, did not change this pattern. A follow-up assessment of undesirable aftereffects of the manipulation revealed perseverance of those effects for several hours after the end of the experimental session.

Part III employed the trauma film paradigm, a frequently used procedure in experimental trauma research. In this deception-free paradigm, participants are being shown highly aversive film clips in order to induce symptoms of a posttraumatic stress disorder. Participants' levels of both clinically relevant posttraumatic stress and subclinical well-being were examined over the course of three weeks after a trauma film manipulation and a subsequent debriefing procedure. While participants' level of posttraumatic stress strongly decreased over the course of three weeks, it was not fully eliminated 20 days after the manipulation. Additionally, participants still reported study-related aftereffects of the trauma film manipulation three weeks after having received a debriefing.

In sum, the present thesis challenges the generally assumed effectiveness of post-experimental debriefings and interventions in undoing the detrimental effects caused by experimental manipulations – and thereby in meeting psychological societies' guidelines to avoid any harm caused by participation in research studies. These results have implications for psychological researchers – given the field's ubiquitous reliance on the (alleged) effectiveness of debriefings – and raise general questions about research ethics.

Zusammenfassung

Psychologen verwenden häufig besonders wirkungsvolle experimentelle Manipulationen, welche entwickelt wurden, um vorrübergehend das Wohlbefinden der Versuchsteilnehmer zu senken, z. B. durch Bedrohen ihres Selbstwertgefühls (*ego threat*). Diese Manipulationen sind oft unverzichtbar zur experimentellen Beantwortung bestimmter Forschungsfragen. Gleichzeitig zeigen sich Forscher schon seit langem besorgt über das Wohlbefinden von Versuchspersonen nach der Teilnahme an Studien, die solche Manipulationen beinhalten. Die offiziellen ethischen Richtlinien verschiedener psychologischer Gesellschaften weisen Forscher sogar explizit an, Versuchsteilnehmer vor Schaden zu schützen. Um das Dilemma zwischen der Nutzung potenziell schädlicher Manipulationen einerseits und dem Bestreben, Versuchspersonen keinen Schaden zuzufügen, andererseits zu lösen, verwenden Forscher nach dem Ende eines Experiments üblicherweise ein so genanntes *Debriefing* um negative Effekte der Manipulation zu eliminieren.

Trotz der Tatsache, dass sich moderne psychologische Forschung stark auf die Effektivität solcher Debriefings verlässt, ist bis heute ungeklärt, ob Debriefings überhaupt tatsächlich effektiv wirken. Forschungsergebnisse aus anderen Gebieten legen nahe, dass experimentell induzierte Annahmen sogar dann noch bestehen bleiben, wenn die zugrundeliegende Information widerrufen wurde. Angesichts dieser möglichen Perseveranz widerrufener Informationen stellt sich die Frage, ob auch die durch eine experimentelle Manipulation ausgelösten negativen Effekte, welche durch das Debriefing aufgehoben werden sollen, weiter anhalten können. Da die Mehrzahl experimental-psychologischer Debriefings daraus besteht, früher gegebene Falschinformation zu widerrufen, scheint diese Frage besonders naheliegend. Die Diskrepanz zwischen dem Vertrauen der Forschungsgemeinde in die Effektivität von Debriefings zur Wiederherstellung des Wohlbefindens der Versuchsteilnehmer auf der einen und dem Mangel an empirischer Evidenz für diese

Annahme auf der anderen Seite wirft dringliche Fragen über die mögliche Perseveranz schädlicher Effekte experimenteller Manipulationen auf.

Die vorliegende Dissertation widmet sich diesen Fragen. Teil I untersucht, inwiefern post-experimentelle Debriefings und andere Interventionen die negativen Effekte bekannter und weitverbreiteter Ego-Threat-Manipulationen in Bezug auf Wohlbefinden und feindselige Attributionen (*hostile attributions*) der Versuchsteilnehmer eliminieren können. Erwartungsgemäß berichteten die Studienteilnehmer, die negatives Feedback erhalten hatten, von geringerem Wohlbefinden und zeigten mehr feindselige Attribution als jene, die positives Feedback bekommen hatten. Unsere Resultate legen nahe, dass die verwendeten Debriefings an der Aufhebung der schädlichen Effekte der Ego-Threat-Manipulationen scheiterten. Stattdessen verließen die Studienteilnehmer das Labor mit ungewollten Nachwirkungen, die noch mindestens einen Tag andauerten. Eine interne Meta-Analyse bestätigt diese Ergebnisse.

Teil II der Dissertation richtet den Schwerpunkt auf eine andere, ebenfalls weitverbreitete Manipulation: experimentell induzierte soziale Zurückweisung (*ostracism*), eine Manipulation welche bekannt dafür ist, verschiedene Indikatoren psychologischen Wohlbefindens negativ zu beeinflussen. Versuchspersonen, die von angeblichen anderen Teilnehmern bei einem virtuellen Ballspiel ausgeschlossen wurden, berichteten über verringerten positiven und erhöhten negativen Affekt im Vergleich zu Versuchspersonen, die im Spiel integriert wurden. Die Verwendung des angeblich wirksamsten Debriefings, das *Revised Outcome Debriefing*, konnte diesen Effekt nicht signifikant reduzieren. Eine spätere Befragung der Studienteilnehmer zeigt eine Perseveranz der negativen Nachwirkungen des Experiments auch mehrere Stunden nachdem die Teilnehmer das Labor verlassen hatten.

Teil III untersucht die Verwendung sogenannter Traumafilme, die häufig in der experimentellen Traumaforschung eingesetzt werden. In diesem Paradigma, welches im Gegensatz zu den beiden vorherigen keine Täuschung der Versuchspersonen benötigt, wurden den Versuchspersonen stark aversive Filmausschnitte gezeigt, mit dem Ziel, Symptome einer

posttraumatischen Belastungsstörung zu induzieren. Nach der Traumafilm-Manipulation und dem anschließenden Debriefing wurden im Laufe von drei Wochen sowohl klinisch relevante Symptome einer posttraumatischen Belastungsstörung als auch subklinisches Wohlbefinden der Versuchsteilnehmer untersucht. Obwohl die posttraumatischen Stresssymptome der Teilnehmer im Verlauf dieser drei Wochen stark abnahmen, waren sie auch zwanzig Tage nach dem Ende der Manipulation noch nicht vollständig eliminiert. Die Studienteilnehmer berichteten zusätzlich auch drei Wochen nach dem Debriefing noch von Nachwirkungen in Bezug auf die Traumafilm-Manipulation,

Zusammengenommen stellt die vorliegende Dissertation die Fähigkeit von Debriefings und anderen Interventionen, potenziell schädliche Effekte experimenteller Manipulationen aufzuheben und damit die ethischen Richtlinien psychologischer Gesellschaften zum Schutz von Versuchspersonen zu erfüllen, in Frage. In Anbetracht des allgegenwärtigen Vertrauens der psychologischen Forschungsgemeinde in die (vermeintliche) Effektivität von Debriefings haben diese Ergebnisse Implikationen für psychologische Forscher und werfen Fragen zur Forschungsethik in unserem Feld auf.

Acknowledgements

First, I would like to thank my supervisor Malte Friese for always taking the time to give me advice, for many fruitful discussions, and not least for giving me the freedom to pursue my research interests and to find my own answers. I am grateful to Cornelius König for agreeing to review this thesis. Furthermore, I want to thank Alexandra Gräbener and Tanja Michael for the friendly and constructive collaboration on our joint project.

I consider myself lucky to have met many great people at the university that added to a nice working atmosphere. In particular, I want to thank Janet Wessler, Lina Gieseler, David Loschelder, Hannah Schade, Marcel Weber, Dorota Reis, and Regine Bader for their advice on work and other things in life. Many thanks go to my research assistant of several years, Alexander Hart, whose reliable support was invaluable. Furthermore, I am thankful to my friend Peter for his language tips and to Nida Bajwa for his advice on official regulations. I sincerely appreciate the community of the *Exzellenzprogramm für Wissenschaftlerinnen* and the friendly network of support it provides.

Finally, I am very grateful to my close friends for their encouragement and support, especially to Fabienne for countless care-packages and to Michael for always being there.

Content

List of Figures and Tables

Introduction

Primum non nocere

— Latin aphorism of unknown origin[1]

The famous aphorism of unknown origin (C. M. Smith, 2005) can be translated to "first, do no harm" or even "above all, do no harm". It is often expressed as the central axiom of medical ethics, capturing the very essence of the Hippocratic Oath (Lilienfeld, 2007; C. M. Smith, 2005). While there is no widely used equivalent motto for psychological ethics, psychological associations subscribe to similar principles of refraining from causing harm. For instance, the *Ethical Principles of Psychologists and Code of Conduct* of the American Psychological Association (APA) state: "Psychologists take reasonable steps to avoid harming their … research participants" (2002, p. 1065). The British Psychological Society (BPS) declares in its *Code of Ethics and Conduct* "Psychologists value their responsibilities … including the avoidance of harm" (2009, p. 18). Similarly, the German Psychological Society (Deutsche Gesellschaft für Psychologie [DGPs]) published official professional guidelines for psychologists that mandate acting with the intention to protect people's well-being ("Psychologinnen und Psychologen handeln im Sinne des Wohls und Wohlbefindens der Menschen"; 2016, p. 5) and the avoidance of causing harm ("Sie vermeiden es, Schaden zuzufügen"; p. 7).

Despite the clear directives from several professional psychological societies to avoid causing harm, empirical psychological research often employs experimental manipulations

[1] The style and formatting of the present thesis is generally based on the guidelines specified in the 6th edition of the *Publication Manual of the American Psychological Association* (2010). Following the American Psychological Association's advice for dissertations, the style of the present thesis slightly differs from the guidelines for journal articles in order to allow for better readability and presentation.

that could be considered harmful for participants (Lilienfeld, 2007; Sharpe & Faye, 2009). For instance, participants might get intentionally ostracized, receive highly unfavorable feedback, get rejected by ostensible other participants, or are shown aversive and disturbing images. Such experimental manipulations often have rather worrisome effects on participants, e.g., they can lead to lowered self-esteem (e.g., Gonsalkorale & Williams, 2007; Heatherton & Vohs, 2000; Horton & Sedikides, 2009; Jamieson, Harkins, & Williams, 2010; Vohs & Heatherton, 2001, 2004; Williams, Cheung, & Choi, 2000), intrusive thoughts (e.g., Bourne, Mackay, & Holmes, 2013; Chou, La Marca, Steptoe, & Brewin, 2014; Clark et al., 2014; Graebener, Michael, Holz, & Lass-Hennemann, 2017; Nixon, Nehmy, & Seymour, 2007), and negative mood (e.g., Gonsalkorale & Williams, 2007; Heatherton & Vohs, 2000; Vohs & Heatherton, 2004; Williams et al., 2000).

The use of highly impactful and possibly harmful experimental manipulations is often justified in two different ways: (1) These manipulations are crucial for experimentally answering certain research questions that cannot be investigated otherwise, such as the questions how people seek to restore threatened self-worth or how they cope with traumata. (2) The employment of a post-experimental debriefing procedure is argued to be a remedy against potential adverse effects (Aronson & Carlsmith, 1968; Baumrind, 1985; Blanck, Bellack, Rosnow, Rotheram-Borus, & Schooler, 1992; D. S. Holmes, 1976a, 1976b; Nestor & Schutt, 2014; Sharpe & Faye, 2009; S. S. Smith & Richardson, 1983; Walster, Berscheid, Abrahams, & Aronson, 1967). Such post-experimental debriefings – the acts of discrediting any potentially employed deception and providing research participants with information about the study they just participated in – are the result of a long and turbulent historical development and have become the standard solution for restoring participants' well-being at the end of an experiment (Sharpe & Faye, 2009).

However, the fact that there was – and to this day still is – no comprehensive factual evidence for the effectiveness of debriefings in restoring participants' well-being was met

with great suspicion by several psychological scientists. While some generally questioned the field's blind trust in the effectiveness of debriefings in the absence of evidence to substantiate that trust (Sharpe & Faye, 2009; Tesch, 1977), others straightforwardly argued that debriefing procedures might fail at removing unwanted effects of experimental manipulations (Anderson, Lepper, & Ross, 1980; Baumrind, 1985; D. S. Holmes, 1976b; Ross, Lepper, & Hubbard, 1975; Tesch, 1977; Walster et al., 1967).

Moreover, an abundance of research illustrates a phenomenon called *belief perseverance*, people's general tendency to maintain a once formed belief or impression even if the underlying information and thereby the evidential basis of the belief or impression has been explicitly discredited (e.g., Anderson & Kellam, 1992; Anderson et al., 1980; Anderson, New, & Speer, 1985; Carroll, 1978; Choi & Nisbett, 2000; M. F. Davies, 1982, 1997; Edwards & Smith, 1996; Fleming & Arrowood, 1979; Greitemeyer, 2014; Guenther & Alicke, 2008; H. M. Johnson & Seifert, 1994; Lord, Lepper, & Preston, 1984; Lord, Ross, & Lepper, 1979; Munro & Ditto, 1997; Nestler, 2010; Nyhan & Reifler, 2010; Ross et al., 1975; Ross, Lepper, Strack, & Steinmetz, 1977; Sherman & Kim, 2002). The well-documented evidence for belief perseverance further challenges the idea that discrediting information that was given at one point during an experiment (e.g., false feedback) will prevent participants from continuing to believe in that information. If however, participants did continue to believe in information like false negative feedback, the potential adverse effects caused by that feedback would arguably persevere as well. Hence, research findings about belief perseverance cast reasonable doubts about the effectiveness of post-experimental debriefings in undoing potential harm caused by experimental manipulations.

Seeing that debriefings as a remedy against adverse experimental effects are often treated as psychological researchers' moral safety net in guaranteeing their participants' well-being, effectiveness of debriefings in undoing potential harm is of crucial importance. Yet, to this day, there is no conclusive answer to the question of debriefings' effectiveness. Do post-

experimental debriefings reliably remove adverse effects caused by experimental manipulations and thereby restore participants' well-being? The present thesis aims to answer this important question with respect to four different highly impactful experimental manipulations.

In the following, I will first give an overview of the historical development of post-experimental debriefings as a remedy against potential adverse experimental effects. That section of the thesis will illustrate the history of research ethics in the field of social psychology and document how debriefings as post-experimental interventions were the subject of criticism but eventually gained widespread acceptance. Second, I will provide an overview of potential mechanisms underlying the phenomenon of belief perseverance to illuminate the challenges that debriefing procedures have to face and to explain why unfounded trust in the effectiveness of post-experimental debriefings might be highly problematic. In three parts, I will then present eight studies (and one meta-analytical summary of five of those studies) that investigate the effectiveness of different post-experimental debriefings and interventions in removing experimental effects and restoring participants' well-being after four different experimental manipulations. The thesis closes with a discussion of the findings and their implications both with respect to the present research question and the perseverance of negative information and experiences in general.

The History of Debriefings and Research Ethics in Social Psychology

In psychological research, the term "debriefing" describes a procedure that usually takes place at the end of an experimental session and provides research participants with information about the experiment they just participated in (Howitt & Cramer, 2011). The debriefing procedure is intended to protect the well-being of participants and minimize or even undo the harm that the research procedures might have caused them (American Psychological Association, 2002; Aronson & Carlsmith, 1968; D. S. Holmes, 1976a, 1976b). While nowadays debriefings are a widely accepted procedure that most researchers deem important (Sharpe & Faye, 2009), this has not always been the case. Instead, today's acceptance of debriefings and concern about participants' well-being are the result of a complicated and turbulent development that lasted several decades and involved a heated debate amongst the scientific community.

Origins of the Term "Debriefing"

The term "debriefing" did not originate in the field of psychological research but was originally coined by the British military, initially describing the interrogation of Royal Air Force pilots after they had returned from missions in World War II (Rule, 1957). It was later adopted by the US military in the Korean war (Rule, 1957). In a meaning closer to the modern psychological use of the word, it was also first used by the military in the 1960s during the Vietnam war (Harris, 1988) where it was defined as "a factual review of events, and individual and unit reactions to those events" (Bartone & Adler, 1995, p. 2). The focus shifted from simply interrogating pilots and soldiers to providing them with an opportunity to reflect upon their personal experience (Tannenbaum & Cerasoli, 2013): "In the process of reviewing events, feelings may be expressed and problems may be defused. The main point of the Debriefing is to review the chronology of events, to give soldiers an opportunity to clear up any confusion, and to facilitate a healthy cognitive reframing and integration of their experiences" (Bartone & Adler, 1995, p. 2).

Early History – 1960s

In contrast to the military's approach, for a long time, psychological researchers did neither use the term "debriefing", nor the concept of an obligatory intervention that provides their participants with a review of the events that they had experienced (Harris, 1988). Until the late 1940s, there had been very little concern about the ethical treatment of research participants in general (American Psychological Association, 1952; Berg, 1954; Rich, 1952). In contrast, the APA had a committee on the ethical treatment of animals in research since 1925 (R. A. Smith & Davis, 2003), partially as a reaction to accusations of animal rights activists (Berg, 1954). The treatment of human participants in research studies, however, had only garnered small, local protests about specific studies including stress and frustration manipulations and was thus considered as unproblematic in comparison (Berg, 1954).

There was a wide-held trust among the scientific community that psychological researchers would follow an unwritten – but somehow implied – equivalent of the Hippocratic oath (Harris, 1988), rendering a more formal ethics code obsolete. In fact, this mindset can be seen in the early work of the Committee on Scientific and Professional Ethics, which was created by the APA in 1938 to decide whether or not the APA needed a formal ethics code for its members. Despite immediately receiving complaints about unethical behavior of psychologists on a regular basis (while not being formally authorized to handle complaints at all), in 1940, the committee reported its recommendation to not adopt an ethics code because "it did not feel that the time was ripe" (American Psychological Association, 1952, p. 426).

The end of World War II and the revelations about the Nazis' cruel treatment of human subjects for research purposes that were brought to light during the Nuremberg trials shattered psychologists' fundamental trust in the ethical treatment of research participants and raised general concerns about the use of human subjects in psychological research (Berg, 1954; Harris, 1988). In response to the growing worry about unethical behavior of researchers and the rapid growth of the field of psychological research, in 1947, the Committee on

Scientific and Professional Ethics revised their previous decision and strongly recommended the adoption of an ethics code (American Psychological Association, 1952). In 1949, the newly founded Committee on Ethical Standards for Psychology started working on developing an official ethics code (Sanford, 1952), resulting in the Ethical Standards of Psychologists (American Psychological Association) that was first published in 1953. The code focused on different areas of psychological work with scientific research being only one small part of it. Nonetheless, the code stated that "only when a problem is significant and can be investigated in no other way, is the psychologist justified in exposing human subjects to emotional stress or other possible harm" (American Psychological Association, 1953, p. 12).

Remarkably, both the development and the introduction of the APA's ethics code were met with tremendous resistance among the scientific community (Adair, Dushenko, & Lindsay, 1985; Adair, Lindsay, & Carlopio, 1983; Bobbitt, 1952; West & Gunn, 1978). While some psychological researchers demonstrated little consideration for participants' well-being, for example by arguing that an ethics code was useless for scientists since they do not deal with "clients" (e.g., Hughes, 1952, p.441), others even opposed the idea of teaching graduate students ethical standards because they considered it valueless and a waste of time for the students (Hackman, 1952). Another group of researchers worried that an ethics code would mean a cheap codified substitute for individual conscience (Sanford, 1952). For instance, Calvin Hall argued that "decent mature people do not need to be told how to conduct themselves" (1952, p. 430). In the end, the debate did not reach a conclusion and dwindled.

1960s – 1980s

A decade after the introduction of the APA's first ethics code, the discussion about the necessity of such a code specifically for psychological research regained traction after the publication of Stanley Milgram's (1963) study on obedience. The study stirred up a debate about the ethical treatment of research participants (Harris, 1988; Sharpe & Faye, 2009; S. S. Smith & Richardson, 1983; Ullman & Jackson, 1982) with Diana Baumrind (1964) leading

the way and accusing Milgram of most likely having seriously harmed his participants. Milgram (1964) defended his study and his use of deception against Diana Baumrind's accusations by emphasizing the use of a post-experimental intervention to restore participants' well-being: a debriefing. In fact, his usage of the now well-known and commonly used term "debriefing" is ironically the first published use of the word in the psychological community (Harris, 1988).

Overall, Milgram's study sparked an outcry about the ethical implications of psychological research and the use of deception on the one hand, while his use of a debriefing procedure on the other hand provided a possible solution to the dilemma of manipulating – and possibly deceiving – participants while not wanting to seriously harm them (Adair et al., 1985; Tesch, 1977). Both lines of discussion that followed the publication of Milgram's obedience study – the debate about ethical treatment of research participants and the debate about the use of debriefing procedures – led to a big controversy among the scientific community, marked by heated arguments from the beginning and pronounced resistance, especially from the field of social psychology (Adair et al., 1985).

Debate about research ethics. The debate about ethical treatment of research participants was fueled by the growth of experimental social psychology in the 1960s that called several topics of psychological research methods into question, with ethical issues and potential harm to participants being one of them (Adair et al., 1983; Elms, 1975; Tesch, 1977; Weber & Cook, 1972; West & Gunn, 1978). The growing use of deception in social psychological studies (for example, Seeman, 1969, found that in 1948, 23.8% of the studies published in the Journal of Personality involved deception, compared to 43.9% in 1963) added to that development. Lastly, the lack of a detailed formal ethics code specifically about psychological research, as well as the ongoing reported abuses in studies led to a revolution in the discipline in the late 1960s and early 1970s (Adair et al., 1983; Elms, 1975; Weber & Cook, 1972). This shift was marked by the APA's development of a new ethics code

specifically concerning the treatment of human participants in psychological research and by the US government demanding Institutional Review Boards (IRBs) for every study that received their funding (Adair et al., 1985, 1983; American Psychological Association, 1973).

In 1973, the APA published its Ethical Principles in the Conduct of Research with Human Participants, the first ethics code specifically addressing psychological research. In particular, the code stated that "the ethical investigator protects participants from physical and mental discomfort, harm, and danger." (American Psychological Association, p. 2). Furthermore, the code stipulated that any kind of deception occurring during the study must be addressed and any harm caused by the research procedure must be undone:

> After the data are collected, ethical practice requires the investigator to provide the participant with a full clarification of the nature of the study and to remove any misconceptions that may have arisen.… Where research procedures may result in undesirable consequences for the participant, the investigator has the responsibility to detect and remove or correct these consequences, including, where relevant, long-term aftereffects. (American Psychological Association, 1973, p. 2)

The controversy about ethical treatment of research participants, sparked by Milgram's obedience study (1963) and fueled by the announcement of the development of a new ethics code, did not cease after the code was published. Instead, both the development and the content of the new ethics code were subject to vivid discussions. One point of view was that the development of ethical guidelines was hasty as long as it had not been proven that manipulations were actually harmful for participants (e.g., Gergen, 1973), while other researchers considered this view as contradictory to basic human rights (e.g., Baumrind, 1971; A. F. Johnson, 1974). Another group acknowledged the possible negative effects to the well-being of participants but considered it a necessary evil (Aronson & Carlsmith, 1968; Baumrind, 1971). For instance, Kaufmann (1967) argued that if the insights of research are valuable enough, inflicting harm on participants is justified. Additionally, he found it invalid

and hypocritical that experimental psychologists criticized harmful manipulation while they "almost without exception, have at times gathered data involving pain, anxiety, or deprivation of their subjects" (p. 322). Some researchers even went as far as arguing that it might be beneficial for participants to realize their own flaws with the "help" of a harsh manipulation (e.g., Baron, 1981; Kaufmann, 1967). In contrast, other researchers not only applauded the general idea of a new ethics code but argued for an even stricter code (e.g., Baumrind, 1971, 1972) or pointed out that researchers' care for participants' well-being was part of the implicit contract between participants and experimenters (Epstein, Suedfeld, & Silverstein, 1973).

The growing interest in ethical aspects of psychological research also led to the first explicit mention of ethical treatment of research participants in a psychological textbook: in 1968, the *Handbook of Social Psychology* stated that researchers are ethically responsible for their participants' well-being (Aronson & Carlsmith, 1968). The authors recommended attempting to avoid the use of measures which cause discomfort and to debrief participants after experiments involving deception to reduce distress and restore their well-being.

Debate about post-experimental debriefings. Simultaneously to the debate about research ethics in general, the use of a post-experimental debriefing to counteract the possibly negative effects of research procedures that was advocated by Aronson & Carlsmith (1968) as well as by the APA's new ethics code (1973) caused another large controversy among the scientific community. A substantial number of researchers came out in favor of debriefing procedures, e.g., by demanding that all participants should be debriefed (e.g., Baumrind, 1985; D. S. Holmes, 1976a; Perry & Abramson, 1980; S. S. Smith & Richardson, 1983). Some even proposed that attending a debriefing session should come with an incentive for participants and that no manuscript should be accepted for publication if the authors did not include a clear statement that participants had been debriefed (e.g., Adair et al., 1985; McNamara & Woods, 1977; Perry & Abramson, 1980). A few researchers tried very hard to develop good debriefing procedures, e.g., Mills (1976) reported having developed her

debriefing procedure, which is longer than three pages of plain text, over the course of 20 years. Aronson & Carlsmith (1968) proposed that a debriefing should ideally last longer than the experimental session itself.

In contrast, numerous researchers questioned or even opposed the use of a post-experimental debriefing procedure for a variety of reasons. Some researchers rejected the idea of being limited in their scientific work by the obligation to employ additional procedures and felt that, if incompatible, successfully conducting their studies was more important than protecting participants' well-being (e.g., Baron, 1981; Menges, 1973). For instance, Resnick & Schwartz (1973) stated "we should be highly cautious in imposing additional methodological limits on our already overtaxed scientific methods" (p. 138). Similarly questioning methodological feasibility of debriefings, a substantial number of researchers expressed great concern that informing participants about the deception they were subjected to would make them more suspicious towards experimental procedures and thereby reduce the number of future naive subjects (e.g., Brock & Becker, 1966; Farrow, Lohss, Farrow, & Taub, 1975; Harris, 1988; Horka & Farrow, 1970; Lichtenstein, 1970; Sieber, 1983b).

Because of concerns regarding future participants' suspicion, some researchers actively decided against debriefing participants. For example, Zimbardo (1965) explicitly stated that he did not debrief participants after deception and manipulated feedback because his university did not have an honor code and he did not want to ruin naive participants for further studies. Similarly, Sieber (1983b) argued that participants should not be debriefed if they participated involuntarily, if the debriefing might reflect badly on the researcher, or if other naive participants are needed. Other researchers generally suspected that participants might actually be content with not learning the truth about the experiment they participated in (e.g., Menges, 1973) or even speculated that participants "enjoy an element of risk and nondisclosure" (Resnick & Schwartz, 1973, p. 137). Furthermore, many researchers were concerned that a debriefing might have negative as opposed to restorative effects on

participants and could additionally make them feel anger and dislike for experimenters (e.g., Adair et al., 1985; Aronson & Carlsmith, 1968; Brock & Becker, 1966; D. S. Holmes, 1976a; Menges, 1973; Mills, 1976; Seeman, 1969; Sieber, 1983a, 1983b). One group of researchers even suggested that a debriefing might trigger paranoid ideas in otherwise non-suspicious participants (Resnick & Schwartz, 1973).

Despite the disagreement on the topic in the scientific community, debriefings started to gain acceptance as both a terminology and practice during the years of the debate (Harris, 1988; S. S. Smith & Richardson, 1983). This development was associated with many researchers' unverified assumption that debriefings work, at least if done correctly (Baron, 1981; D. S. Holmes, 1976a; McFarland, Cheam, & Buehler, 2007; Mills, 1976; Ross et al., 1975; Sharpe & Faye, 2009).

Based on the assumption of debriefings' effectiveness, some psychological scientists used the employment of a post-experimental debriefing procedure as a carte blanche for conducting research that was generally considered to be harmful. For instance, one group of researchers induced partial deafness in participants through posthypnotic suggestion – not telling them why they were suddenly deaf – to find out whether they would develop symptoms of paranoia (Zimbardo, Andersen, & Kabat, 1981). Several other researchers still used Milgram's obedience paradigm (e.g., Kilham & Mann, 1974; Mantell, 1971; White, 1979). One research group even used the paradigm on children as young as six years old (Shanab & Yahya, 1977) and stated in their article that "no subject appeared to be affected adversely throughout the experiment" (p. 532), while at the same time describing participants' "intense tension" and "trembling" (p. 534).

The growing trust in debriefings' effectiveness, however, was not rooted in factual evidence (Tesch, 1977), a circumstance that was met with suspicion by some psychological scientists. Walster and colleagues (1967) found that "most experimenters seem to operate under the conviction that a thorough debriefing will accomplish the goal of returning the

subject to his preexperimental state" (Walster et al., 1967, p. 371). Similarly, Frederick Tesch stated that "it apparently became an article of professional faith that debriefing sufficiently minimized or eliminated any serious vilification of devious experimenters" (1977, p. 218). He further criticized the scientific community for what he deemed irrational behavior:

> On the one hand, we devise marvelous manipulations and hone them for maximum impact upon our participants. On the other hand, we apparently assume that the effects produced conveniently cease when the participants leave our experiments. Have we discovered the best of all possible worlds, in which events happen when we wish and do not when we turn away from them? It is perhaps this puzzling assumption that permits researchers to accept the 'eraser function' without second thoughts and without the database we require for most other statements about human behavior. (p. 219)

In addition to some researchers pointing out the field's blind trust in the effectiveness of debriefings, others straightforwardly suggested that debriefing procedures might fail at undoing the effects of an experimental manipulation (Anderson et al., 1980; Baumrind, 1985; D. S. Holmes, 1976a; Ross et al., 1975; Tesch, 1977; Walster et al., 1967).

Consequences of the debates. Fueled by the debate about post-experimental debriefings but in defiance of the general debate about research ethics, the 1970s and 1980s were marked by a reluctant but noticeable increase in the adoption of debriefing procedures but also in an increased use of deception, with no change in the types and intensity of deception (Adair et al., 1985). This development is documented in several survey studies.

An early study (Menges, 1973) analyzed nearly 1000 articles published in APA journals in 1971. Out of these almost 1000 articles, 17% reported the use of deception (40.3% of the ones published in the *Journal of Personality and Social Psychology* [JPSP]). Only 10% of the studies reported the use of a debriefing procedure. No article mentioned the employment of any other intervention or effort to restore participants' well-being. Menges

furthermore stated that he did not expect that "the use of nonvolunteer subjects or the use of deception will decline, even though they seem inherently incompatible with APA's call for 'openness and honesty'" (1973, p. 1034). A few years later, McNamara & Woods (1977) investigated the use of deception and debriefings in all articles published in JPSP from 1971 to 1974. They found that 57% of the studies reported deception and that 32% mentioned the use of a debriefing procedure, which depicts an increase in both compared to Menges' study.

The continuous use of both deception and debriefings in the late 1970s and early 1980s was documented in another study (Adair et al., 1985) that compared to what extent the use of deception and debriefings, respectively, were reported in several journals in both 1979 and 1983. In some journals, the percentage of studies that mentioned deception or debriefings, respectively, remained similar (e.g., in 1979, 57% of the studies published in the *Personality and Social Psychology Bulletin* [PSPB] reported deception and 29% of the studies referred to a debriefing compared to 54.4% and 24.6%, respectively, in 1983). Other journals showed a pronounced increase in the report of both deception and debriefing: In 1979, 59% of the studies published in the *Journal of Experimental Social Psychology* (JESP) reported the use of deception and 49% of the studies mentioned the use of a debriefing procedure. In 1983, however, as much as 81.2% of the studies published in JESP reported the use of deception and 63.8% mentioned the employment of a debriefing procedure.

The use of post-experimental debriefings was furthermore documented in a study by Ullman & Jackson (1982) that examined all articles that had been published in JPSP over the course of 14 years. They showed that in 1966, 19% of the articles had reported the use of a debriefing procedure compared to 47% of the articles that were published in 1980. The authors suggest that the increased use of debriefings was due to the controversy surrounding Milgram's obedience study, the introduction of the APA's ethics code, and the rise of institutional review panels (in fact, studies with college students as participants reported debriefing participants more often).

The ethical development in the scientific field also affected the content of textbooks: Adair and colleagues (1983) analyzed 46 experimental psychology textbooks that were published between 1970 and 1981 and contained sections on methods for conducting research with human participants. The authors found a slow growth in the mention of debriefing procedures: between 1970 and 1975, post-experimental debriefing procedures were mentioned in only 28% of the textbooks and their description was covered using on average as little as 1.5 lines of the entire textbook. In contrast, between 1976 and 1981, post-experimental debriefing procedures were mentioned in 32% of the books and their description took up on average 16.5 lines of the textbook.

In general, between the 1960s and the 1980s, both the use of deception and debriefings in psychological research increased. Nevertheless, the employment of post-experimental debriefing procedures did not become ubiquitous.

Post 1980s

After the heated debate during the 1960s and 1970s, the controversy about ethical guidelines, deception, and post-experimental debriefings quietened down over the next decades without the scientific community actually achieving agreement on those topics. Even though today, the APA and most local ethics committees require researchers to debrief participants, the use of a debriefing procedure is reported in only some scientific articles and the nature of the employed debriefing is rarely ever described (Adair et al., 1985; Sharpe & Faye, 2009). A study by Sharpe & Faye (2009) analyzing articles that were published in the years 2006 and 2007 showed that only 32% of the articles published in JPSP and only 6.7% of articles published in the *Journal of Traumatic Stress* (JTS) reported the use of a debriefing procedure. When asked explicitly in a subsequent survey, 46% of the authors who had published in JTS admitted not having debriefed their participants at all. In fact, 10.6% of the authors whose articles were published in JPSP and 16% of the authors whose articles were published in JTS found debriefings generally unnecessary.

Moreover, today there is still no conclusive evidence for the effectiveness of post-experimental debriefings in restoring participants' well-being: "At best, there is a small body of studies that suggests a more extensive debriefing may provide greater relief than a standard debriefing" (Sharpe & Faye, 2009, p. 441). This is due to the fact that in addition to the vast decrease in discussions about debriefing procedures in general since the 1970s, the specific question of the effectiveness of debriefings was and still appears to be of very little interest to the scientific community (Adair et al., 1983; Sharpe & Faye, 2009). Thus, even today, we only have a "lack of evidence for the effectiveness of debriefing coupled with an unfounded faith in the ability of debriefing to erase ethical problems" (Sharpe & Faye, 2009, p. 444).

The Mechanism of Perseverance of Discredited Information

> *The human understanding when it has once adopted an opinion (either as being the received opinion or as being agreeable to itself) draws all things else to support and agree with it. And though there be a greater number and weight of instances to be found on the other side, yet these it either neglects and despises, or else by some distinction sets aside and rejects.*
>
> — Francis Bacon, *Novum Organum Scientiarum*

In psychological research, a phenomenon similar to the one portrayed in the previous quote by British philosopher Francis Bacon is called *belief perseverance.* This term describes people's tendency to stick to a once formed belief even if the underlying information and thereby the evidential basis of the belief is explicitly discredited (Anderson et al., 1985; M. F. Davies, 1993; Guenther & Alicke, 2008; Lieberman & Arndt, 2000; Nestler, 2010).

Research on Perseverance and Persistence

Research on belief perseverance focuses on a plethora of areas, such as the persistence of lay theories in light of contradicting evidence (e.g., Anderson & Kellam, 1992; Anderson et al., 1980, 1985; Lord et al., 1979), the persisting belief in retracted scientific information (M. F. Davies, 1993, 1997; Greitemeyer, 2014; Munro & Ditto, 1997; Nestler, 2010), and the lasting effects of false feedback even after its revocation (e.g., M. F. Davies, 1982; Fleming & Arrowood, 1979; Guenther & Alicke, 2008; McFarland et al., 2007; Ross et al., 1975).

Additionally, there is a rather recent branch of research on the so-called "debunking" of "misinformation" and its tendency to persist (Ecker, Lewandowsky, Chang, & Pillai, 2014; Ecker, Lewandowsky, Fenton, & Martin, 2014; Ecker, Lewandowsky, Swire, & Chang, 2011; H. M. Johnson & Seifert, 1994; Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). In the misinformation paradigm, participants usually read a news article about a fabricated event,

which contains a piece of causal information that, at some point during the experiment, gets retracted for half of the participants. Afterwards, participants are asked to make inferences about the fabricated event with the number of references to the discredited piece of information being the crucial dependent variable. Overall, the misinformation paradigm is concerned with the revocation of only one single piece of causal information that is embedded in an article (Ecker, Lewandowsky, Fenton, et al., 2014).

In contrast, the belief perseverance paradigm is concerned with the retraction of the complete set of information, e.g., the whole article instead of just one small falsehood in it. A special – and classic – case of belief perseverance is investigated in the debriefing paradigm (Anderson et al., 1985; Birgegard & Sohlberg, 2008; McFarland et al., 2007): participants initially receive false feedback about their performance in a task. This feedback then gets fully retracted in a debriefing session. The critical dependent variable is usually participants' estimate of their own abilities in areas related to the task they completed (e.g., M. F. Davies, 1982; Fleming & Arrowood, 1979; Guenther & Alicke, 2008; McFarland et al., 2007; Ross et al., 1975).

Explaining Belief Perseverance

Belief perseverance has been investigated in numerous studies and seems to be a very robust phenomenon (Anderson & Kellam, 1992; Anderson et al., 1985; Cobb, Nyhan, & Reifler, 2013; M. F. Davies, 1997; Guenther & Alicke, 2008; Nestler, 2010; Wilson & Brekke, 1994) (Anderson & Kellam, 1992; Anderson et al., 1985; Cobb et al., 2013; M. F. Davies, 1997; Guenther & Alicke, 2008; Nestler, 2010; Wilson & Brekke, 1994). Yet, the process behind it appears complex, is still poorly understood (Chan, Jones, Hall Jamieson, & Albarracín, 2017; M. F. Davies, 1982; Lewandowsky et al., 2012; Sherman & Kim, 2002), and subject to controversy (Munro & Ditto, 1997). In general, the different approaches to explain the process behind belief perseverance can be roughly divided into three categories: a motivational approach, a cognitive approach, and an affective approach.

Proponents of the motivational approach argue that belief perseverance emerges because people are generally motivated to reject information that is inconsistent with their original beliefs. The cognitive approach explains perseverance with several cognitive biases and mechanisms such as primacy, the anchor effect, and the availability heuristic. Supporters of the affective approach view belief perseverance as a result of the experienced reality of people's feelings that cannot be discredited. Advocates of all three approaches seek to support their arguments by providing scientific evidence and often rigorously reject alternative explanations. However, it is noteworthy that even among proponents of the same approach opinions about the specifics of the perseverance process vastly differ.

Motivational explanations for belief perseverance. The motivational approach to explain belief perseverance is based on the idea that people filter information with respect to whether or not it matches their beliefs and desires: if new information is consistent with people's original beliefs, they trust, support, and assimilate it; if, in contrast, a new piece of information contradicts people's original beliefs, they question, discount, and reject it (Guenther & Alicke, 2008; Kruglanski & Freund, 1983; Lord, Lepper, & Preston, 1984; Lord et al., 1979). According to the motivational approach, all judgments about the validity, reliability and relevance of new information are heavily biased by people's beliefs and wishes (Lord et al., 1984, 1979). Therefore, when people are presented with a hypothesis that is incongruent with their wishes they become very sensitive to information that is disconfirmatory to this hypothesis. This heightened sensitivity leads to an assimilation of disconfirming evidence which then allows for a quick rejection of the undesired hypothesis (Kruglanski & Freund, 1983). By the same logic, when people are presented with a hypothesis that is congruent with their wishes they become very *in*sensitive to disconfirmatory information, which facilitates acceptance of the desirable hypothesis (Kruglanski & Freund, 1983).

One point of criticism regarding the motivational approach is concerned with the strength of people's pre-study beliefs (e.g., Ecker, Lewandowsky, Fenton, et al., 2014; Ross et al., 1975): research on belief perseverance usually only investigates the strength of people's beliefs *after* the retraction of the original piece of information. If, however, belief perseverance occurs because people are motivated to uphold their beliefs, one could argue that the strength of their motivation and thereby the strength of the perseverance effect should depend on the intensity of their *pre*-study beliefs. By contrast, the results of two studies suggest that participants' beliefs *before* receiving the original piece of information did not influence the effectiveness of its later retraction (Ecker, Lewandowsky, Fenton, et al., 2014). Additionally, in the debriefing paradigm, observers of a study including the revocation of prior given false feedback showed as much perseverance as the actual recipients of the feedback (Ross et al., 1975). Since observers and recipients of the feedback quite likely did not share the exact same pre-study beliefs and the same background knowledge, the authors concluded that background knowledge and pre-study beliefs do not play a critical role in the perseverance process.

A different argument against the validity of the motivational approach is not just made by opponents of the approach (M. F. Davies, 1993) but acknowledged by some of its proponents (e.g., Guenther & Alicke, 2008) as well: in light of people's general aim to maintain the most positive self-image possible it seems puzzling that not only positive but also negative feedback can persevere. If people's motivation to uphold their beliefs was the key aspect of perseverance, failure experiences should not persevere since people with a healthy self-esteem should clearly be motivated to maintain a positive self-image. The easiest way to maintain such a positive image in an experimental setting that includes false negative feedback would be to simply accept the feedback's retraction.

Opposing this argument, one group of researchers claims that people's self-esteem might simply depend on being right: for instance, once participants have received feedback

they will discard every piece of subsequent contradicting information because acknowledging the falsehood of the initial feedback would mean acknowledging having been fooled (Kruglanski & Freund, 1983). Another group of researchers explains the perseverance of negative feedback by emphasizing the strength of the perseverance effect in general (Guenther & Alicke, 2008). According to them, the perseverance effect might simply be capable of overpowering people's attempts to restore a positive self-image. Therefore, they view people's desire to dismiss negative information in order to protect a favorable self-image as a moderator of the perseverance effect. Indeed, they find that after positive feedback and its subsequent retraction, both the recipients of the feedback and observers of the scene showed the same degree of perseverance. By contrast, after retracted negative feedback, observers displayed a higher degree of perseverance than the recipients of the negative feedback themselves. The authors concluded from their finding that people show indeed less perseverance of a piece of information if it is detrimental to their self-image.

Cognitive explanations for belief perseverance. The cognitive approach of explaining belief perseverance is both the most prevalent one and the most diverse: while its rather large group of proponents agrees that the process behind belief perseverance is a cognitive one, they differ in their opinion on its exact nature. The various parties do, however, share the conviction that cognitive processes create new information, which becomes independent of the original piece of evidence and, therefore, perseveres even after the original information has been retracted.

For instance, some researchers believe that the original false piece of information is used to make subsequent inferences and to create associations that can lead to rumination about the original event (M. F. Davies, 1993; McFarland et al., 2007; Ross et al., 1975; Topolinski & Strack, 2010). Those associations stay fully intact after the original information has been discredited and can thus interfere with people's attempts to mentally correct this information. Within the debriefing paradigm, some researchers suggest that after the

revocation of prior given false feedback, participants continue to ruminate and involuntarily use the false feedback as an anchor when wondering about their actual score (e.g., McFarland et al., 2007).

Others reject the anchor effect as an explanation for perseverance of false feedback based on their findings about the relationship between perseverance and cognitive resources (e.g., Fleming & Arrowood, 1979): since people with low cognitive resources and little information especially rely on anchors (Tversky & Kahneman, 1974), mental distraction should facilitate the use of anchors. Thus, if the perseverance effect could indeed be explained by people's use of anchors, people should experience *more* perseverance of information under low cognitive resources. However, when participants had been mentally distracted (by counting backwards from 200 by 3) after receiving a piece of information and before its retraction, they subsequently showed *less* perseverance of the initial information compared to participants who had not been distracted (Fleming & Arrowood, 1979).

Some researchers even fully reject the idea that rumination plays a significant role in the cognitive process underlying belief perseverance: When a piece of information is discredited immediately after its initial presentation, participants have no time to ruminate before the retraction. Nonetheless, participants in such an experimental condition showed as much perseverance as participants in a condition that included a five minute delay between initial presentation and discrediting of the piece of information (H. M. Johnson & Seifert, 1994). Hence, the opportunity to ruminate did not affect belief perseverance.

Another, much more widespread approach to account for belief perseverance is loosely related to the aforementioned propositions based on people's inferences and associations: the idea that the generation of causal explanations for an event leads to perseverance (Anderson & Kellam, 1992; Anderson et al., 1980, 1985; Chan et al., 2017; M. F. Davies, 1993, 1997; Ecker, Lewandowsky, Fenton, et al., 2014; Ecker et al., 2011; Fleming & Arrowood, 1979; H. M. Johnson & Seifert, 1994; Lewandowsky et al., 2012; Nestler, 2010;

Richter, Schroeder, & Wöhrmann, 2009; Ross et al., 1975). According to this idea, whenever people experience events – especially if these events are surprising, extreme, or unusual – they search for a causal explanation including antecedent elements that would likely lead to the experienced event. Those resulting antecedents become fully independent of the original event. Therefore, the newly created antecedent-consequence links necessarily lead to perseverance: if the consequence is discredited, the antecedents remain unchanged, are still salient and thus highly available, and will imply consequences similar to the one that has been discredited. Hence, the freshly generated causal network results in perseverance.

Several empirical findings support the idea that generating causal explanations leads to perseverance and give further insights into various underlying aspects of the process: For instance, participants who generated explanations for an outcome showed more perseverance after the outcome was discredited than participants who merely read provided explanations (M. F. Davies, 1997). The same pattern of results emerged when the provided explanations had been rated beforehand as being especially convincing and even when only hypothesized outcomes instead of actually experienced ones had been considered.

Similarly, in the debriefing paradigm, generating explanations for one's own performance has been shown to produce more perseverance than just being given time to think about the performance without clear instructions to engage in causal reasoning (Fleming & Arrowood, 1979). Interestingly, the results showed a direct correlation between the strength of the perseverance and the number of causal reasons participants produced for the initial outcome. In fact, some researchers argue that more reasoning – and thereby usually generating more explanations for an event – makes the created antecedents more salient and thereby leads to regarding the original event as a highly likely potential outcome, even when it has been discredited (e.g., Anderson & Kellam, 1992; Ross et al., 1975). By a similar logic, some researchers advocate the generation of explanations for a hypothetical opposite outcome

as a countermeasure for perseverance (e.g., Anderson & Kellam, 1992; H. M. Johnson & Seifert, 1994; Lord et al., 1984; cf. Nestler, 2010).

Other researchers reject the claim that more generated explanations lead to more perseverance and vice versa and point out the influence of the cognitive accessibility and availability of potential explanations (e.g., Carroll, 1978; Nestler, 2010). Empirical findings show that while being instructed to create a *few* explanations for an outcome leads to perseverance after its revocation, being instructed to create *many* explanations for an outcome can even fully inhibit perseverance (Nestler, 2010). In a similar vein, instructions to create many explanations for hypothetical opposite outcomes lead to perseverance of the actually observed outcome while instructions to create only a few explanations for the opposite can weaken the perseverance effect (Nestler, 2010). Therefore, the recommendation to generate counter-explanations as a remedy for perseverance is seen as highly problematic by some researchers because the effectiveness of this intervention vastly depends on the accessibility of those counter-explanations (e.g., Chan et al., 2017; Nestler, 2010).

Further evidence opposing the claim that generating causal explanations leads to perseverance shows that participants who were asked to only *imagine* an outcome experienced as much perseverance as participants who were asked to actively generate explanations for the outcome (Carroll, 1978). Hence, it was not the generation of explanations but the mere cognitive availability of the outcome as a concept that facilitated belief perseverance. Following this logic, the most prominent example of the debriefing paradigm – the finding that performance feedback can persevere after a debriefing and subsequently influence participants' perceptions of themselves (Ross et al., 1975) – can be interpreted in a different way: it is argued that recent events, such as the false feedback in the debriefing paradigm, are simply more salient and therefore more available and easier to recall, which causes biased judgement and thereby perseverance (Carroll, 1978). The retraction of initial information is thus, by definition, bound to fail: retraction necessarily involves repetition of

the initial information, which makes it more salient, more available and thereby strengthens its perseverance (Ecker et al., 2011; Lewandowsky et al., 2012; Lieberman & Arndt, 2000; Nestler, 2010). In the literature, this phenomenon is often called *backfire effect.*

One group of researchers, however, firmly opposes the view that the mere availability of initially given information causes its perseverance: in an experiment, H. M. Johnson & Seifert (1994) found that simply presented information did not persevere after its retraction whereas information that included causal explanations did persevere. Other researchers simply incorporate the concept of cognitive availability in their claim that the generation of causal explanations leads to perseverance. They argue that cognitive availability mediates perseverance because the most recently created or considered causal explanation is the most available one and thereby has the greatest influence on subsequent perseverance (Anderson et al., 1985).

A different, much smaller controversy about explaining the perseverance effect inverts the aforementioned argument that the most recently created information is the most important one in creating perseverance: it focuses on the primacy effect, the phenomenon that *earlier* encoded information can be remembered more easily than later encoded information. Proponents of explaining perseverance with the primacy effect claim that within the debriefing paradigm, participants naturally encode the false feedback first and subsequently disregard every piece of information that contradicts this feedback – including the debriefing (e.g., Kelley & Michela, 1980). Opponents of this view point out that the nature of perseverance is very different from the nature of the primacy effect: in the case of the primacy, new information is *added* to initially given information with the intention to *augment* it; in the case of belief perseverance, initial information is meant to get *subtracted*, which describes a *challenge* to the authenticity of the original piece of information (M. F. Davies, 1997; Ross et al., 1975). Additionally, primacy and perseverance differ in how rational their effects appear to be: while it is difficult to objectively define to what degree old

information should ideally change after the addition of new one, it seems clearly illogical to believe in information that has already been explicitly discredited (Anderson et al., 1980, 1985; M. F. Davies, 1997).

Affective explanations for belief perseverance. Proponents of the affective approach of explaining the perseverance effect claim that perseverance in general is an affective phenomenon. They build their arguments on the reasoning that affect is autonomous from facts and cognitions and therefore remains unaffected if the underlying factual basis is discredited (e.g., Sherman & Kim, 2002; Zajonc, 1980). According to Sherman and Kim (2002), cognitions can easily be changed by the invalidation of old and the provision of new information. A crucial component of that process is comparing the cognition in question with the – possibly contradicting – external reality of events. In contrast, given that affect has actually been experienced it constitutes an internal reality and is therefore true by definition. This leads to affective perseverance even after successful cognitive invalidation because affective invalidation is hardly possible.

As opposed to Sherman and Kim (2002), Ross, Lepper, and Hubbard (1975) rejected affective explanations for their findings: not only recipients of false performance feedback showed perseverance after the feedback had been discredited but also people who had merely observed the scene. Based on this pattern of results, the researchers concluded that the perseverance phenomenon is not based on affect since observers should not be emotionally affected by false feedback given to others. This view is shared by a number of proponents of cognitive explanations for belief perseverance (e.g., Lord et al., 1984; McFarland et al., 2007). Sherman and Kim (2002), on the other hand, argue that it is actually quite possible that observers were emotionally affected, though likely to a lesser extent than the recipients of the feedback. The authors support their argument by a set of studies demonstrating that participants even show affective perseverance with respect to content that is neither emotionally involving nor self-relevant: Participants were presented with Chinese symbols

and their respective alleged meaning that had clear positive or negative valence. Unsurprisingly, participants displayed a preference for symbols that were assigned positive meaning. When the alleged meaning was retracted and participant were informed about the true, neutral meaning of the symbols, participants were able to remember the new meaning but still showed a clear preference for symbols that were originally coupled with positive compared to negative meaning. Sherman and Kim (2002) reason that since the meaning of Chinese symbols holds no personal relevance for the participants, their role in the study was similar to the observers' role in the study by Ross et al. (1975). Thus, an affective process could be a valid explanation for the perseverance phenomenon.

A different view on the affective approach connects emotions with memory (Edwards & Smith, 1996): Affective reactions to a new piece of information might trigger a more extensive memory search for arguments for or against the validity of the information. Additionally, affect can lead to physiological arousal that might distract people and bias their judgement in a way that facilitates the generation of invalid (counter-)arguments.

An interesting finding that bolsters the potential role of affective processes in the perseverance phenomenon (cf. Sherman & Kim, 2002) comes from research on patients with Korsakoff's syndrome, a form of anterograde amnesia caused by chronic alcohol abuse. For instance, in one study, Korsakoff patients were provided with facial stimuli that were paired with fictional biographical information, which contained descriptions of either exclusively positive or exclusively negative attributes (M. K. Johnson, Kim, & Risse, 1985). Twenty days after the initial presentation, the Korsakoff patients were unable to recall any of the biographical information but still showed a clear preference for the facial stimulus that had initially been paired with the positive biography. Hence, the Korsakoff patients seemed to have experienced affective perseverance even in the absence of cognitive recollection.

The Value of Belief Perseverance

As suggested by the Francis Bacon quote at the beginning of this chapter, once formed

beliefs can be tremendously difficult to erase (Anderson et al., 1980; M. F. Davies, 1982; Ecker et al., 2011; Edwards & Smith, 1996; Guenther & Alicke, 2008; Lord et al., 1979; Markus & Wurf, 1987; Nestler, 2010; Ross et al., 1975; Sherman & Kim, 2002; Wilson & Brekke, 1994; Zajonc, 1980). People's adherence to already discredited information is often considered fundamentally illogical (Anderson et al., 1980, 1985; M. F. Davies, 1997; Nestler, 2010) but, as some researchers point out, it might be a reasonable reaction considering social rules and implications of conversational logic: People generally operate under the assumption that others tell the truth or that they at least believe in what they tell someone. In the case of discrediting prior given information, the other person openly admits to having told a lie and subsequently presents a new and different alleged truth. Since both the initial information and the subsequent correction are – from a point of conversational logic – almost equally likely to be wrong, it seems in fact very logical that people find it hard to make sense of such a situation and might eventually choose a way of reinterpretation that can lead to perseverance (H. M. Johnson & Seifert, 1994; Lewandowsky et al., 2012; Richter et al., 2009).

Some researchers even suggest that belief perseverance can be seen as a quite useful mechanism. For instance, Anderson and colleagues (1980) argue that the world would be a very chaotic one if people would challenge all of their beliefs whenever they receive new information. Choi and Nisbett (2000) even assert that giving up beliefs too easily and unconditionally accepting new information as true would lead to less epistemic curiosity and less engagement in information-seeking behavior.

The arguments for the potential usefulness of adhering to one's beliefs notwithstanding, perseverance has been shown to develop even when the original piece of evidence has been presented only very quickly for a single time under cognitive load, when it is minimal, weak, heavily logically flawed, or inconclusive (Anderson et al., 1980; Ecker et al., 2011). More so, attempts to retract information can lead to a backfire effect that ultimately strengthens the perseverance of that piece of information (Ecker et al., 2011; Lewandowsky et

al., 2012; Lieberman & Arndt, 2000; Nestler, 2010). Concluding a meta-analysis on the debunking of misinformation, Chan and colleagues (2017) even give researchers who attempt to erase prior misinformation the advice to "keep expectations low" (p. 1544). In the case of the debriefing paradigm with participants' well-being after potentially harmful experimental manipulations being at stake, this piece of advice seems, of course, almost impossible to accept.

The Present Research

The previous two chapters illustrated how difficult and controversial the development of debriefings as mandatory interventions against harmful manipulations had been and how powerful and steadfast the perseverance effect is. The fact that the heated debate about the employment of debriefing procedures quietened down without the scientific community actually achieving agreement on those topics seems almost like a calm *after* the storm. This relative calm after the previous tumultuous developments might explain why explicitly voiced concerns about the effectiveness of post-experimental debriefings had "little discernable impact on the research literature" (Sharpe & Faye, 2009, p. 432).

Regardless of what the reason for researchers' "comparatively little attention" (Sharpe & Faye, 2009, p. 433) to debriefings' effectiveness is – Tesch (1977) suggested that it could be convenience and fear of uncomfortable insights – attempting to find an answer to the question of debriefings' effectiveness is of great importance and should not be postponed. This is especially true since the use of possibly harmful experimental manipulations is often justified with the alleged effectiveness of post-experimental debriefing procedures (Aronson & Carlsmith, 1968; Baumrind, 1985; Blanck et al., 1992; D. S. Holmes, 1976a, 1976b; Nestor & Schutt, 2014; Sharpe & Faye, 2009; S. S. Smith & Richardson, 1983; Walster et al., 1967). Unfortunately, research on the perseverance effect and theories about the affective, cognitive, and motivational processes behind it raise doubts about the mere possibility of removing effects of something that has been said or has happened. Taken together, the field's great reliance on debriefings' effectiveness in undoing the potential harm caused by experimental manipulations on the one hand and the sheer strength of perseverance of discredited information on the other hand call for a thorough examination of debriefings' effectiveness in removing adverse experimental effects.

In three parts[2] which encompass eight empirical studies and one internal meta-analysis, the present research investigates the effectiveness of different debriefing procedures and an additional intervention in removing the adverse effects caused by four different experimental manipulations from the fields of social psychology and trauma research. Further, the present research investigates whether the experimental manipulations would cause any discernable effects that would last until after participants had left the laboratory.

Part I: Debriefed but Still Troubled? About the (In)Effectiveness of Post-Experimental Debriefings After Ego Threat

Ego threat, defined as a threat to people's self-esteem or self-image, is a construct that is frequently used in experimental manipulations in social psychological research. Apart from lowering self-esteem, ego threat is also known to reliably lead to negative emotions (Leary, Terry, Batts Allen, & Tate, 2009). How can such adverse effects of impactful ego threat manipulations be removed?

The so called *Revised Outcome Debriefing* is supposedly the most effective type of debriefing to eliminate the effects of deceptive information given to research participants in a study (McFarland et al., 2007). However, it has never been put to the test against potentially harmful manipulations like ego threat. Part I aims to close this gap in research and investigates the effectiveness of a Revised Outcome Debriefing in removing the negative effects caused by ego-threatening manipulations.

Additionally, the effectiveness of a different procedure in restoring participants' well-being is examined: a self-affirmation intervention. Self-affirmations are intended to restore people's positive self-image, especially when it has been threatened (Cohen & Sherman, 2014). Specifically, a self-affirmation can reduce rumination and increase various aspects of

[2] The order of the three parts in the present thesis was chosen with respect to theoretical considerations and does not correspond to the chronological order in which the research had been conducted.

well-being including mood and self-esteem (Cohen & Sherman, 2014; Klein, Blier, & Janze, 2001; Koole, Smeets, van Knippenberg, & Dijksterhuis, 1999). Self-affirmation interventions are usually not used at the end of an experiment to restore participants' well-being in lieu of or in addition to a post-experimental debriefing. However, considering its mitigating effects after threat, a self-affirmation's effectiveness as a post-experimental intervention is certainly worth investigating.

Part I addresses the immediate effects of adverse experimental ego threat manipulations and examines the effectiveness of a Revised Outcome Debriefing and a self-affirmation procedure, respectively, in removing those effects. Furthermore, it investigates the persistence of the effects by measuring participants' well-being up to two days after they had left the lab.

Part II: When Loneliness Sticks With You: Can a Debriefing Undo the Effects of Ostracism?

Part II focusses on a different experimentally employed concept, which is known as *ostracism*, a term that describes being ignored and excluded by others. In social psychological research, ostracism manipulations have been used in over 200 studies (Hartgerink, van Beest, Wicherts, & Williams, 2015) and have been shown to result in a large number of negative effects, such as negative mood (Gonsalkorale & Williams, 2007; Williams et al., 2000) and lowered self-esteem (Gonsalkorale & Williams, 2007; Jamieson et al., 2010; Williams et al., 2000).

Ostracism is usually manipulated by letting participants play a ball tossing game in which they are systematically excluded by other ostensible participants. In 2000, a virtual analogue, called *Cyberball* was invented with the intention to create an ostracism manipulation that would have *less* severe effects and be less traumatic for participants (Williams, 2007; Williams et al., 2000). Are the effects of such an ameliorated manipulation easy to overcome with the employment of a post-experimental debriefing? Part II addresses

this question and investigates the effectiveness of a Revised Outcome Debriefing in removing both the immediate and the potentially longer-lasting (several hours after the lab session) effects caused by the Cyberball manipulation.

Part III: Will a Debriefing Eliminate the Distress Caused by a Trauma Film Paradigm?

In contrast to Part I and Part II, Part III does not focus on an experimental manipulation that includes deception. Instead, it is centered on the trauma film paradigm, a procedure that is frequently employed in experimental trauma research. In the trauma film paradigm, participants watch short film clips with highly aversive content like accidents or extreme violence. This procedure is known to have a very high negative impact on participants and to reliably cause subclinical symptoms of a posttraumatic stress disorder (Bourne et al., 2013; Chou et al., 2014; Clark et al., 2014; Graebener, Michael, et al., 2017; Michael, Holz, & Lass-Hennemann, 2016; Nixon et al., 2007). Can such strong detrimental effects be eliminated by a debriefing procedure? Since it does not include any deception, trauma film research can obviously not employ post-experimental debriefings that discredit prior given false information like the Revised Outcome Debriefing used in Part I and Part II. Instead, the debriefing in Part III consisted of a session that offered participants the opportunity to discuss their experiences with the experimenter. The effectiveness of this debriefing procedure in eliminating clinically relevant posttraumatic stress as well as subclinical well-being was investigated at three, respectively two, different time points over the course of three weeks after the trauma film manipulation.

PART I

Debriefed but Still Troubled? About the (In)Effectiveness of Post-Experimental Debriefings After Ego Threat

Stefanie Miketta & Malte Friese

Saarland University

Manuscript submitted for publication to the *Journal of Personality and Social Psychology*.
(Currently under revision)

Abstract

Psychological researchers often use powerful experimental manipulations to temporarily reduce participants' well-being. In contrast, the APA mandates that research participants must be protected from any psychological or physical harm. Post-experimental debriefings are intended to eliminate potentially detrimental effects of experimental manipulations. However, research suggests that beliefs elicited in an experiment can persevere even when the underlying information is later explicitly declared as false. The present research investigates whether debriefings rectifying all falsehoods given to participants in the context of ego-threatening manipulations would re-establish participants' pre-study conditions. In five studies, participants received false feedback about their intelligence (Studies 1 & 5) or their attractiveness and likeability (Studies 2–4), completed dependent variables indicative of well-being (Studies 1, 2, 4, 5), or aggressive behavior and hostile attributions (Study 3), and were thoroughly debriefed in writing or in person (Studies 4 & 5). In the critical conditions, participants were debriefed *before* completing the dependent variables. Participants reported lower well-being (Studies 1, 2, 4, 5) and exhibited more hostile attributions (Study 3) after receiving negative compared to neutral or positive feedback. This pattern was largely unchanged when participants had already been debriefed. In Studies 4 and 5, similar results emerged even when participants were debriefed in person and when participants wrote a self-affirming essay before they completed the dependent variables. Follow-up assessments in Studies 2–5 revealed adverse effects of the manipulations that pertained at least one day after the end of the experimental session. An internal meta-analysis confirmed our findings. These results challenge the effectiveness of debriefings and raise ethical questions about current research practices.

Keywords: Debriefing, Ego Threat, Perseverance

Introduction

They may forget what you said, but they will never forget how you made them feel.

— Carl W. Buehner (1898-1974)

Let us hope Carl Buehner was wrong. In various lines of psychological research participants are purposefully brought into situations that will make them *feel* bad. They relive traumatic events, are confronted with disgust- or anxiety-eliciting stimuli, or receive ego-threatening feedback about their intellectual abilities or physical appearance, to name just a few. What researchers *say* is the debriefing at the end of the study: it provides background information and is intended to undo all negative feelings potentially elicited during the study. However, do debriefings reliably accomplish that goal? What do we actually know about how participants feel when they leave the laboratory? With only a few exceptions, research has never systematically examined whether and when debriefings are effective. The present research seeks to close this gap and investigate the effectiveness of post-experimental debriefings in restoring participants' well-being after their ego has been threatened.

Several fields of psychological research have a long history of powerful experimental manipulations intended to purposefully, albeit temporarily, affect participants' well-being. For example, in the analogue trauma paradigm, participants watch film scenes of highly aversive content like the depiction of accidents or extreme (sexual) violence (E. A. Holmes, Brewin, & Hennessy, 2004; Nixon, Cain, Nehmy, & Seymour, 2009). Other manipulations target domains that are highly personally relevant to the participants. For example, in studies investigating the consequences of social rejection, participants are being blatantly ignored, explicitly excluded by others, or led to believe that in the future they will live a life alone without close social bonds (for meta-analyses, see Blackhart, Nelson, Knowles, & Baumeister, 2009; Gerber & Wheeler, 2009). In research investigating the consequences of

ego threat, participants receive negative feedback about their performance, intellectual abilities, or physical appearance (for an overview see Leary et al., 2009). Both social rejection and ego threat manipulations impose a temporary threat to participants' well-being and diminish their self-esteem.

Research Ethics

Psychological researchers have been concerned about the impact of experimental manipulations such as those just outlined above on participants' well-being for a long time. In 1953, the American Psychological Association (APA) published its first version of an ethics code: the *Ethical Standards of Psychologists.* It stated "only when a problem is significant and can be investigated in no other way, is the psychologist justified in exposing human subjects to emotional stress or other possible harm" (p. 12). The following decade was marked by a crisis in psychological research concerning several topics of research methods, with ethical issues being one of them (Elms, 1975; Weber & Cook, 1972). Some have argued that researchers should avoid any experimental procedures that may "involve loss of dignity, self-esteem, and trust in rational authority" (Baumrind, 1964). Others were more lenient, but demanded that potentially harmful manipulations should only be employed under precautions such as recruiting only psychologically stable participants (Kelman, 1967). The consensus of many researchers was that participants' well-being should not be harmed as a consequence of participating in psychological studies (D. S. Holmes, 1976b; Tesch, 1977; Walster et al., 1967). In line with this consensus, the tone of the *Ethical Standards of Psychologists* became stricter: "The investigator protects the participant from physical and mental discomfort, harm, and danger that may arise from research procedures" (American Psychological Association, 1981, p. 638). Remaining faithful to this standard, the current *Ethical Principles of Psychologists and Code of Conduct* states: "Psychologists take reasonable steps to avoid harming their … research participants" (American Psychological Association, 2002, p. 1065).

Arguably conflicting with this claim, some lines of research entail experimental

manipulations that are specifically designed to cause temporary harm (e.g., when participants are intentionally being socially excluded, rejected by others, insulted, subjected to physical pain, threatened in their ego and self-worth or shown disturbing images). Although researchers' proximate goal in these studies is to inflict harm on participants, it is of course not their ultimate goal. It is a means to an end: in many situations causing harm may be a crucial prerequisite for answering certain research questions such as how people respond to social exclusion, how they seek to restore a threatened self-worth, or how they cope with traumata. Not being willing to inflict this harm on participants would often times mean to not be able to investigate these and other research questions in the first place.

Debriefings

Throughout several decades, many researchers advocated the use of post-experimental debriefings to solve the dilemma of using (temporarily) harmful manipulations despite not wanting to cause participants any harm (Baumrind, 1985; Blanck et al., 1992; D. S. Holmes, 1976a, 1976b; Nestor & Schutt, 2014; Sharpe & Faye, 2009; S. S. Smith & Richardson, 1983). These debriefings are usually intended to (a) provide background information which could not be given prior to the study in the informed consent form without jeopardizing the study's aim, (b) clarify the use of deception (if any) and provide accurate information instead, and (c) generally undo any harm that may or may not have been induced.

Despite the general routine to offer a debriefing to research participants, there is no broadly accepted standard debriefing procedure that researchers could adapt to their needs. The vast majority of scientific publications do not provide information as to what exactly the debriefing that was used entailed. The APA (2002; see also Sharpe & Faye, 2009; S. S. Smith & Richardson, 1983; Sommers & Miller, 2013) points out the critical aspect of informing participants about the (potentially false) nature of the manipulation as the main component of debriefings. The underlying rationale of this advice is that informing participants about any false information that was provided during the study will undo the potentially harming

consequences this false information might have caused.

However, whether correcting false information that was given earlier in the study is sufficient to restore participants' well-being is unclear: Research suggests that even information that is explicitly declared as false can still persevere (e.g., Anderson et al., 1980; Greitemeyer, 2014; Guenther & Alicke, 2008; Nyhan & Reifler, 2010; Wyer, 2010). For example participants continue to believe in already revoked scientific findings (Greitemeyer, 2014) and in corrected political misconceptions (Nyhan & Reifler, 2010), and they hold on to impressions of persons even when the information causing these impressions has been discredited (Wyer, 2010).

The implication of these studies for the present research is clear: If originally given information can persevere after it has been corrected and explicitly declared as false, then this may also be true for information provided in the context of post-experimental debriefings that seek to countervail any incomplete or false information provided earlier.

Effectiveness of Debriefings

Up to this point we have established that (a) abundant psychological research seeks to temporarily impair participants' well-being, (b) post-experimental debriefings are the method of choice to undo detrimental effects of these manipulations before participants leave the laboratory, and (c) information that is explicitly revoked and declared as false can nevertheless persevere. This begs the question what we actually know about the effectiveness of debriefings in fulfilling their aim to restore participants' psychological state to pre-study conditions.

Considering the pivotal role of post-experimental debriefings in meeting the APA's requirements to minimize any harm caused by research procedures, it is disconcerting that to this date, it remains unknown whether debriefings actually do that. In general, research on the effectiveness of debriefings is surprisingly scarce (cf. McFarland et al., 2007; Ross et al., 1975; Silverman, Shulman, & Wiesenthal, 1970; Walster et al., 1967).

In a classic study, Ross and colleagues (1975) provided participants with false feedback about their social perceptiveness after they tried to distinguish fake suicide notes from real ones. Afterwards, participants were fully informed about the false feedback. The authors applied two types of debriefings: the *(Standard) Outcome Debriefing* explaining that participants had received *false* feedback about their social perceptiveness and the *Process Debriefing* explaining furthermore the so-called "perseverance effect" – that beliefs caused by false feedback can persevere and lead to incorrect self-perceptions. After receiving the debriefing, participants estimated their actual social perceptiveness. The (Standard) Outcome Debriefing was ineffective: participants who had received a poor social perceptiveness feedback estimated their social perceptiveness to be lower than participants who had received an excellent social perceptiveness feedback. The Process Debriefing was effective in eliminating this difference in self-perceptions that had been caused by the manipulation.

In a replication of this study, McFarland and colleagues (2007) found a similar pattern of results – an ineffective Standard Outcome Debriefing and a more effective Process Debriefing. The authors additionally developed a third type of debriefing procedure, the Revised Outcome Debriefing: participants not only learn that the feedback that they had received was false, but also that the test to measure their social perceptiveness was invalid and did not measure any underlying abilities in the first place. According to the authors this latter information could be crucial for eliminating the perseverance effect: McFarland and colleagues (2007) suspected that the Standard Outcome Debriefing did not work because participants might have been "wondering what their real score on the test is, and might even use the fake score as an anchor with which to estimate their real score" (p. 234). As the Revised Outcome Debriefing provides participants with the information that the test was invalid in the first place, they should not ruminate about their real test score. In fact, the Revised Outcome Debriefing not only proved to be effective in eliminating the effects of the manipulation but was even more effective than the Process Debriefing. Moreover, it was the

only kind of debriefing that successfully prevented post-study rumination in participants. Additionally, McFarland and colleagues (2007) speculated that contrary to the Process Debriefing, the Revised Outcome Debriefing might not require participants' engagement in controlled corrective processes and therefore be "more effective … among participants who are less motivated or able to engage in an effortful correction process" (p. 239).

The work by McFarland and colleagues (2007) suggests that the Revised Outcome Debriefing is the most effective type of debriefing to eliminate the effects of deceptive information given to research participants in a study. However, the effectiveness of the Revised Outcome Debriefing was hitherto only examined in the context of false feedback about the ability to distinguish fake from real suicide notes, an ability that is arguably not relevant for people's everyday lives. In contrast, some of the prominent psychological manipulations discussed earlier (e.g., analogue trauma, ego threat, social exclusion) are highly personally relevant and threatening. This personal relevance may also impact on the effectiveness of the debriefing procedure.

Taken together, the literature leaves researchers with conflicting hypotheses: On the one hand, the APA mandates the use of a debriefing to eliminate any harm that may have been caused (implying debriefings' effectiveness). In addition, many researchers advocate the use of a debriefing to minimize harm (Blanck et al., 1992; D. S. Holmes, 1976b, 1976a; Nestor & Schutt, 2014; Sharpe & Faye, 2009; S. S. Smith & Richardson, 1983). On the other hand, evidence for the effectiveness of debriefings is scarce, mixed and comes from studies in which no actual harm was inflicted on participants' well-being. Additionally, some studies failed to find any appreciable effects of debriefings (Silverman et al., 1970; Walster et al., 1967). To date, it is unknown which of these hypotheses holds true. Are debriefings capable of eliminating the detrimental effects of ego threat manipulations on participants' well-being?

The Present Research

The question whether debriefings are effective in eliminating adverse effects caused

by ego threat manipulations might be unanswered, but it is of great importance considering that ego-threatening manipulations are widely used and known to impair participants' well-being. This paper thus seeks to examine the effectiveness of different debriefing procedures in eliminating the harm caused by ego threat manipulations. In five studies, we sought to conceptually replicate the negative effects of two well-established ego-threatening manipulations (indirect failure feedback and direct negative social feedback) on indicators of subjective well-being such as mood and self-esteem (Studies 1, 2, 4, 5) or aggressive behavior and hostile attributions (Study 3). The main aim was to investigate whether a thorough written Revised Outcome Debriefing would eliminate those effects as intended. In an effort to examine the effects of different post-experimental interventions, we additionally investigated the effectiveness of both a verbal Revised Outcome Debriefing and a self-affirmation task in Studies 4 and 5. Furthermore, Studies 2–5 included a follow-up measurement investigating whether the ego threat manipulations would cause any discernable effects at least one day after participants had left the laboratory. Lastly, we conducted an internal meta-analysis of all studies. Internal meta-analyses are recommended because they allow for increased statistical power, more reliable effect size estimates and therefore more robust and cogent conclusions than individual studies (Braver, Thoemmes, & Rosenthal, 2014; Goh, Hall, & Rosenthal, 2016; Maner, 2014). There are no unreported studies in our file drawer (see Footnote 4). All studies were approved by the local ethics committee.

Study 1

In Study 1, we investigated whether the most effective kind of debriefing, the Revised Outcome Debriefing, would be able to remedy the effects of an ego threat manipulation on subjective well-being. Participants received *indirect failure feedback*: after completing an ostensible IQ test allegedly predictive of future academic success they were led to believe that they had performed poorly (*threat* conditions; e.g., Heatherton & Vohs, 2000; Vohs & Heatherton, 2001, 2004). In the neutral control conditions participants were not told that the

test was indicative of future academic success and did not receive performance feedback. We expected participants in the ego threat conditions to show signs of impaired subjective well-being as indicated by, among others, more negative mood and lowered state self-esteem. To investigate whether the Revised Outcome Debriefing would be capable of neutralizing these postulated effects of the ego threat manipulation, we debriefed half of the participants *before* the dependent variables were collected. This allowed for a comparison between participants who had already been debriefed and participants who had not been debriefed yet.

Method

Participants and design. We planned our sample size based on past research on the same ego threat manipulation and a dependent variable that was used in both previous research and the present study (state self-esteem; Heatherton & Vohs, 2000, Studies 1 and 2; Vohs & Heatherton, 2004, Study 1). We used the mean of the effect sizes of these three studies, weighted by sample size as the effect size estimate for the present study (f = .33). A power analysis based on the present 2 × 2 design using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) and aiming for at minimal power of 1-β = .80 revealed a minimal sample size of N = 75 participants (for α = .05). Therefore, we aimed to recruit at least 80 participants (20 per condition; see Simmons, Nelson, & Simonsohn, 2011), but as many participants as possible until the end of the semester.

One hundred students (53% females; mean age M = 20.91, SD = 2.29) of various disciplines from a German university (homogenous sample with respect to ethnicity) participated in exchange for two vouchers for a hot beverage at a café on campus and a piece of freshly baked cake. They were randomly assigned to one of four conditions of a 2 (ego threat condition: threat vs. no threat) × 2 (debriefing status (at time of DV measurement): debriefed vs. not debriefed yet) design and run in individual sessions that lasted about 20 minutes. Dependent variables were various measures indicative of subjective well-being. In past research, the ego threat manipulation used in this study reliably affected self-esteem and

various indicators of affective experience (Heatherton & Vohs, 2000; Vohs & Heatherton, 2001; 2004). Therefore, an explicit and an implicit self-esteem measure, as well as one measure of mood served as dependent variables. If participants would take the indirect failure feedback seriously, it should not only affect their mood and self-esteem but also their subjective comparison with others on the dimensions affected by the ego threat manipulation. We therefore additionally included a measure of self-perception as a dependent variable. Eleven participants were excluded because they indicated that they did not look at the false statistics provided to them as a critical part of the ego threat manipulation (see below). The manipulation could therefore not work as intended for these participants. The final sample consisted of 89 participants (54% females; mean age $M = 21.02$, $SD = 2.36$). To prevent the inclusion of participants who could become suspicious of the hypothesis, no psychology students were tested.

Procedure. After giving informed consent, participants in the *threat* conditions completed an ostensible IQ test. The experimenter left the room while the participant worked on the test. After four minutes, the experimenter returned and scored the test. Participants then received an ostensible scientific publication about the test that they had just completed.

Participants in the neutral control conditions were led to believe that they participated in a pilot study and were asked to test several items' difficulties. The experimenter returned after four minutes and simply asked participants whether they had liked the test.

Afterwards, participants in the *not debriefed yet* conditions completed the dependent variables (state self-esteem, mood, self-perception, implicit self-esteem) as well as demographics (age, gender, major, and whether they had ever participated in an IQ test). They then received the debriefing. Participants in the *debriefed* conditions received the debriefing before completing the dependent measures. In several pilot runs of the study, we measured the time that participants spent reading the debriefing information. In the *not debriefed yet* conditions, we then included a loading screen that lasted for the average of these

measurements (50 seconds). This ensured that for all participants the same amount of time passed between the manipulation and the completion of the dependent variables. After the debriefing and completion of the dependent measures, participants were thanked and dismissed.

Experimental manipulations.

Ego threat. To manipulate ego threat, we used an indirect failure task used by Vohs and Heatherton (2001, 2004; Heatherton & Vohs, 2000). Participants completed a version of the Remote Associates Test (RAT; Mednick, 1968). Specifically, they were asked to complete 12 German word puzzles. To solve these puzzles, participants had to find one word that connects three seemingly unrelated words (see Appendix for the list of items). To ensure the efficacy of the manipulation (which increases statistical power), threat was manipulated in three ways: First, participants in the *threat* conditions were told that performance on the test was a strong indicator of future academic achievement, and that therefore they should try to solve as many items as possible in the available time frame of four minutes. Second, items in the *threat* conditions were very difficult. In a pre-test based on 23 participants from the same population the average success rate was only 6.0% (*SD* = 8.7%). Third, after four minutes, the experimenter returned with a red pen and scored the test while making supposedly surprised remarks about the fact that the participant had solved so few items (Heatherton & Vohs, 2000; Vohs & Heatherton, 2001, 2004). The experimenter then left the room after providing the participant with an ostensible scientific journal article about the test so that they could "take a look at the correct answers". The article looked like an actual journal article and contained the correct answers to all questions plus the alleged standard average values other students achieve in this test (7.2 out of 12). These statistics were in fact incorrect and biased to suggest to participants that they had performed poorly in the test. The actual success rate of our participants in the *threat* conditions was 1.25 out of 12 (10.42%, *SD* = 1.39), thus considerably lower than the false standard score.

Participants in the *no threat* conditions were asked to test the difficulty of several word puzzles in a pilot study and to "just give it a try" for a couple of minutes. The items they received had been tested beforehand as being moderately easy (average success rate of 69.7%, *SD* = 17.5%) by the same 23 participants who pre-tested the difficult items. When the experimenter returned after four minutes, he/she asked participants whether they had liked the test. Because of the implied irrelevance of the test and the rather high, but not perfect success rate of 9.14 out of 12 (76.17%; *SD* = 3.11), participants in the *no threat* control conditions should have experienced neither an ego threat nor an ego boost.

Debriefing status at the time of the DV measurement. In the *not debriefed yet* conditions, participants completed the dependent variables *before* they received the debriefing, following previous research (Heatherton & Vohs, 2000; Vohs & Heatherton, 2001, 2004). In the *debriefed* conditions, participants were first debriefed and then completed the dependent variables.

Debriefing. Participants received a Revised Outcome Debriefing that has been shown to be the most effective type of debriefing (McFarland at al., 2007). The debriefing was presented on the computer screen. They read that contrary to what they had learned earlier (1) the test was not a real IQ test (*threat* conditions) or a pilot study (*no threat* conditions), respectively, (2) the test was not an indicator of academic achievement (*threat* conditions only) (3) the average score indicated in the ostensible scientific journal article was false (*threat* conditions only), and (4) the participant had been randomly assigned to receive very hard (*threat* conditions) or easy (*no threat* conditions) items, respectively. In order to continue with the study, participants had to confirm the following sentence by marking a checkbox: "I fully understand that the test was not a real IQ test (*no threat* conditions: pilot study) and that I have received very difficult (*no threat* conditions: easy) items". All participants confirmed this.

Dependent variables.

State self-esteem. To measure state self-esteem, we used the German version (Rudolph, Schröder-Abé, & Schütz, 2009) of the State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991). The SSES consists of 3 subscales (performance, appearance and social self-esteem, five items each). We reasoned that believing that one performed poorly in an IQ test should not affect state self-esteem relating to body image and other people's opinions, but performance self-esteem only. Therefore, we focused on the performance self-esteem subscale. Internal consistency for this subscale was $\alpha = .64$.

Mood. Participants indicated their current mood on 12 items taken from the Positive And Negative Mood Scale - Expanded Form (PANAS-X; Watson & Clark, 1999). Items were: ashamed, relaxed, at ease, proud, joyful, irritable, happy, cheerful, upset, downhearted, angry, sad (5-point rating scales, 1 = *very slightly or not at all*, 5 = *extremely*). After reverse-coding items indicating negative mood, internal consistency was good ($\alpha = .88$).

Self-perception. We used a modified version of the How I See Myself questionnaire (HSM; Taylor, Lerner, Sherman, Sage, & McDowell, 2003) to measure self-perception. Participants rated themselves on different qualities in comparison to the average student of their age and gender. Eight items addressed the qualities threatened by the false feedback (intelligent, capable, successful, incompetent, stupid, clever, talented, smart). They were embedded in eight distractor items addressing qualities irrelevant to the false feedback (impatient, unkempt, religious, athletic, fashionable, social, caring, loyal; 7-point scales, 1 = *A lot less than the average student*, 7 = *A lot more than the average student*). Internal consistency for the focal items was high ($\alpha = .90$).

Implicit self-esteem. To measure implicit self-esteem we used a self-esteem-IAT (Greenwald & Farnham, 2000). Participants sorted stimuli that were presented on the computer screen into four categories by pressing one of two response keys. Response latencies were collected for each response. Participants were instructed to react as fast as

possible while committing as few errors as possible. Category labels were “positive”, “negative”, “self”, and “other”. For all categories we used the items used by Dislich et al. (2012): Evaluative stimuli were five positive and five negative words addressing qualities threatened by the false feedback (positive: intelligent, smart, bright, capable, sharp; negative: stupid, slow-witted, untalented, dense, foolish). The “self” and “other” categories consisted of five generic stimuli each (self: me, self, own, mine, my; other: not-me, you, theirs, they, yours). The IAT consisted of 7 blocks. Blocks 1, 2, and 5 were practice blocks. Blocks 3, 4, 6, and 7 were critical blocks. In blocks 1 and 5, participants sorted stimuli into the “self” and “other” categories. In block 2, participants sorted evaluative stimuli into their respective categories. In blocks 3 and 4, the “self”/”other” categories were each combined with an evaluative category and shared a response key with it. In blocks 6 and 7 the “self”/”other” categories were each combined with the evaluative category that they were not combined with in blocks 3 and 4. It was counterbalanced whether the combination self+positive or the combination self+negative was encountered first. Blocks 1, 2, 3, and 6 contained 20 trials each, blocks 4, 5 and 7 contained 40 trails each. The inter-trial interval was 250 ms. The mean error rate was 7.77%. The D_1 score (Greenwald, Nosek, & Banaji, 2003) was used to determine the IAT effect for implicit self-esteem. A positive score reflects a stronger association between oneself and positive relative to negative concepts, thus indicating a higher implicit self-esteem.

Results

Analytic strategy. For each dependent variable, we ran a 2 (ego threat condition: threat vs. no threat) × 2 (debriefing status: debriefed vs. not debriefed yet) ANOVA. In each case, we first examined the predicted simple main effect of the *ego threat condition* in the *not debriefed yet* conditions to establish the focal effect that the debriefing procedure would aim to remedy (e.g. lower state self-esteem after threat compared to no threat). If there was evidence for this simple main effect, we analyzed the simple main effect of the *ego threat*

condition in the *debriefed* conditions. Next, we further examined whether the general main effect of *ego threat condition* was significant across both *debriefing status* conditions. Then, we looked for a potential interaction between *ego threat condition* and *debriefing status*. Finally, we examined a potential (unexpected) main effect of the *debriefing status* factor.

If there was no evidence for the simple main effect of the *ego threat condition* in the *not debriefed yet* conditions, this meant that there was no significant influence of the feedback manipulation on the dependent variable. In this case, our central research question referring to the effectiveness of the debriefing could not be examined: in the absence of an effect of the feedback manipulation, it is impossible to examine whether this non-existent effect persevered after the debriefing. Therefore, we do not report the full set of analyses in these cases and instead move on to the next dependent variable.

All dependent measures are reported in the order in which participants completed them.

State self-esteem. In the *not debriefed yet* conditions, participants reported lower levels of performance state self-esteem after *threat* (M = 3.16, SD = 0.65) than after *no threat* (M = 3.54, SD = 0.61), $t(85) = 2.16$, $p = .034$, $d = 0.61$. This establishes the detrimental effect of ego threat on state performance self-esteem. The planned contrast in the debriefed conditions revealed that the effect persevered after the debriefing (M_{Threat} = 3.19, SD = 0.53; M_{NoThreat} = 3.60, SD = 0.58), $t(85) = 2.23$, $p = .028$, $d = 0.73$. The main effect of *ego threat condition* was significant (M_{Threat} = 3.18, SD = 0.59; M_{NoThreat} = 3.57, SD = 0.59), $F(1, 85) = 9.62$, $p = .003$, $\eta_p^2 = .10$. There was no interaction between *ego threat condition* and *debriefing status* $F(1, 85) = 0.01$, $p = .933$, $\eta_p^2 < .01$. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 3.37, SD = 0.65; M_{Deb} = 3.43, SD = 0.59), $F(1, 85) = 0.13$, $p = .723$, $\eta_p^2 < .01$, as expected (Figure 1a). Taken together, ego threat led to lower performance state self-esteem, no matter whether participants had already received a debriefing or not.

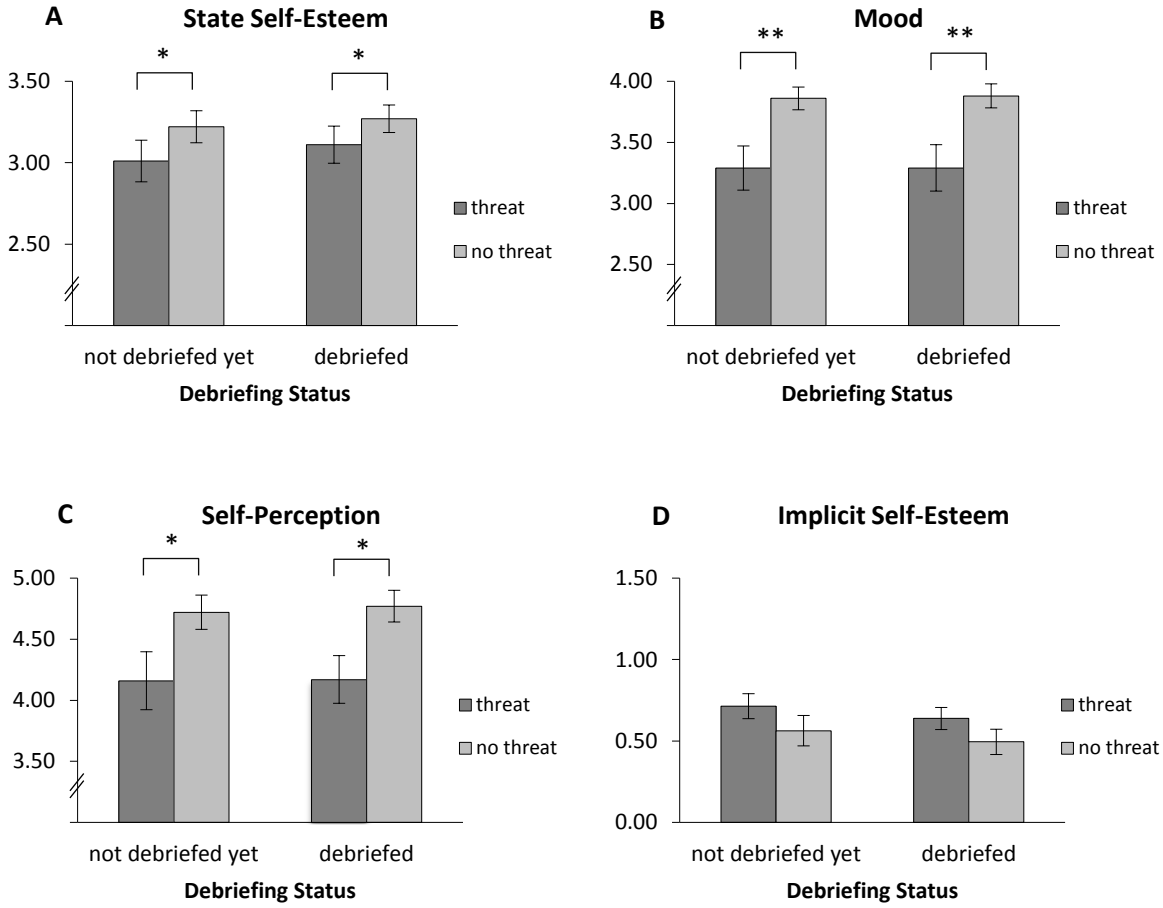
Mood. As expected, in the *not debriefed yet* conditions, participants reported lower

mood after *threat* (M = 3.29, SD = 0.81) than after *no threat* (M = 3.86, SD = 0.47), $t(84)$ = 2.97, p = .004, d = 0.89. This establishes that the ego threat manipulation successfully manipulated mood. The simple contrast in the *debriefed* conditions revealed that the effect persevered after the debriefing (M_{Threat} = 3.29, SD = 0.79; $M_{NoThreat}$ = 3.88, SD = 0.51), $t(84)$ = 3.00, p = .004, d = 0.93. The main effect of *ego threat condition* was significant (M_{Threat} = 3.29, SD = 0.79; $M_{NoThreat}$ = 3.87, SD = 0.48), F(l, 84) = 17.80, p < .001, η_p^2 = .18. There was no interaction between *ego threat condition* and *debriefing status* F(l, 84) = 0.01, p = .919, η_p^2 < .01. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 3.61, SD = 0.70; M_{Deb} = 3.65, SD = 0.69), F(l, 84) = 0.01, p = .925, η_p^2 < .01 (Figure 1b). Taken together, ego threat led to decreased mood, and the debriefing intended to remedy this effect was unsuccessful in doing so.

Self-perception. In the *not debriefed yet* conditions, participants showed less favorable self-perceptions after *threat* (M_{Threat} = 4.16, SD = 1.03) than after *no threat* ($M_{NoThreat}$ = 4.72, SD = 0.69), $t(84)$ = 2.30, p = .024, d = 0.66. Thus, the ego threat manipulation successfully manipulated self-perception. The planned contrast of *ego threat condition* in the *debriefed* conditions revealed that the effect persevered after the debriefing ($M_{NoThreat}$ = 4.78, SD = 0.67; M_{Threat} = 4.17, SD = 0.83), $t(84)$ = 2.46, p = .016, d = 0.83. The main effect of *ego threat condition* was significant ($M_{NoThreat}$ = 4.75, SD = 0.67; M_{Threat} = 4.17, SD = 0.92), F(l, 84) = 11.30, p = .001, η_p^2 = .12. There was no interaction between *ego threat condition* and *debriefing status* F(l, 84) = 0.02, p = .898, η_p^2 < .01. Finally, there was no main effect of debriefing status (M_{NoDeb} = 4.53, SD = 0.79; M_{Deb} = 4.48, SD = 0.89), F(l, 84) = 0.03, p = .857, η_p^2 < .01 (Figure 1c). Taken together, ego threat led to less favorable self-perceptions, and the debriefing was unsuccessful in changing this effect.

Implicit self-esteem. In the *not debriefed yet* conditions, participants showed similar levels of implicit self-esteem after *threat* (M_{Threat} = 0.71, SD = 0.35) as after *no threat*

($M_{NoThreat}$ = 0.56, *SD* = 0.47), *t*(85) = 1.29, *p* = .199, *d* = 0.36 (Figure 1d). Thus, the feedback manipulation had no significant effect on implicit self-esteem. Therefore, we stopped the analysis of this dependent variable at this point.



Figures 1a–1d. State self-esteem, mood, self-perception, and implicit self-esteem as a function of *ego threat condition* and *debriefing status*. Participants who had received negative feedback showed lower levels of mood and state self-esteem, as well as less favorable self-perceptions than participants who had received no feedback, no matter whether they had been debriefed at the time of the collection of the respective dependent variable or not. Error bars indicate +/- 1 SEM. **p* < .05, ***p* < .01.

Discussion

Participants who had received negative performance feedback on an ostensible IQ test showed impaired subjective well-being as indicated by lower state self-esteem and mood, as well as less favorable self-perceptions. A Revised Outcome Debriefing failed to eliminate

these aversive effects. Debriefed participants who had learned that they in fact had not worked on an IQ test and that they had purposefully received particularly difficult word puzzles were equally affected by the ego threat manipulation as participants who had not been debriefed yet.

Study 2

Possibly, the ineffectiveness of the debriefing to remedy the aversive effects caused by the ego threat manipulation in Study 1 was due to an unknown peculiarity of this particular ego threat manipulation. We did not have a specific hypothesis about this, but we considered it appropriate to remove possible doubts concerning this matter. Therefore, Study 2 tested whether the effects obtained in Study 1 would generalize to a different ego threat manipulation. Participants received *negative social feedback*: two ostensible other participants conveyed that they found the participant not particularly likable and physically attractive (Horton & Sedikides, 2009). Previous work established that this kind of negative social feedback constitutes a serious challenge to usually positive self-images and a threat to participants' need to belong (Horton & Sedikides, 2009). The threat manipulation is similar to several other ego threat manipulations that rely on similar negative social feedback (i.e. Beer, Chester, & Hughes, 2013; Leary, Haupt, Strausser, & Chokel, 1998; Somerville, Heatherton, & Kelley, 2006; Swann, Hixon, Stein-Seroussi, & Gilbert, 1990).

Study 1 followed previous research in that it focused on the short-term effects of the ego threat manipulation measured in the minutes after the threat occurred. Study 2 additionally examined effects of the ego threat manipulation on preoccupation with the study long after all participants had been debriefed and left the laboratory.

Method

Participants and design. The ego threat manipulations used in the present research were previously employed by Heatherton & Vohs (2000; see also Vohs and Heatherton 2004) as well was Horton and Sedikides (2009). A German sample in the mid-2010s likely

considerably differs from US-American samples in the 2000s. We therefore followed other researchers (e.g., Carr et al., 2017) and based our sample size analysis on effect sizes from our own work (i.e., Study 1). This approach also had the advantage that we could estimate the required sample size based on the same dependent variables that Study 2 would employ. A power analysis using G*Power (Faul et al., 2007) revealed a minimal sample size of $N = 79$ for at least 80% power ($\alpha = .05$, $f = 0.32$). We therefore recruited a minimum of 80 participants (20 per condition; see Simmons et al., 2011), but as many as possible until the end of the semester.

Eighty-one students (47% females; mean age $M = 23.49$, $SD = 3.26$) of various disciplines from a German university (homogenous sample with respect to ethnicity) participated in exchange for €4 and sweets. They were randomly assigned to one of four conditions of a 2 (ego threat condition: yes vs. no) × 2 (debriefing status: debriefed vs. not debriefed yet) design and run in individual sessions that lasted about 30 minutes. Dependent variables were various measures indicative of subjective well-being (see below for details). To prevent the inclusion of participants who could become suspicious of the hypothesis, no psychology students were tested. No participants were excluded from the analyses.

Procedure. After giving informed consent, participants were told that the study would examine people's behavior in an online dating situation. For this purpose, they would have to record a brief video in which they would introduce themselves to two ostensible other participants. Participants then recorded their video, which was allegedly sent to the other participants via a video chat service. Next, they saw the videos of the other two participants and rated them on a standardized rating sheet. Participants then received the rating sheets referring to their previously recorded video that were allegedly completed by the other two participants. Afterwards, participants indicated their demographics. Then, half of the participants continued by completing the dependent variables (mood, state self-esteem, self-perception, implicit self-esteem). They then received the debriefing. The other half of the

participants received the debriefing *before* completing the dependent measures. In several pilot runs of the study, we measured the time that participants spent reading the debriefing information. In the *not debriefed yet* conditions, we included a loading screen that lasted for the average of these measurements (45 seconds). This ensured that for all participants the same amount of time passed between the manipulation and the completion of the dependent variables. After debriefing and completion of the dependent measures, participants were asked to fill out an online follow-up questionnaire later on the same day. Then they were thanked and dismissed. Later on the same day, participants received the link to the online questionnaire via email. Participants entered a personalized code that they had previously used in the lab session (e.g., first letter of the father's first name and similar information). This allowed us to match the data from the lab session and the follow-up for each participant while participants still remained anonymous.

Experimental manipulations.

Ego threat. To manipulate ego threat, we used a video recording and exchange task similar to the one used by Horton and Sedikides (2009). Participants were informed that there were two other participants in adjoining rooms and that the experimenter would regularly have to leave the participant's room to check on them. They were also told that the study was about people's behavior in an online dating situation. To this end, each participant was to record a video briefly introducing her/himself (stating her/his name, age, major, hobbies, side jobs and stays abroad). These videos would later be shared among participants and each participant would subsequently anonymously evaluate the other two based on the respective video. In reality, there were no other participants. The experimenter sent the videos of "the other participants" using a video chat program (ooVoo) in which she was logged in with two different accounts at the same time. The videos participants saw had been recorded beforehand by two confederates, one male, one female, so that every participant saw a video of a potential romantic partner regardless of their gender and sexual orientation. This

increased the manipulation's efficacy (and thereby statistical power). Additionally, both confederates had been rated previously by 22 independent raters on 7-point scales as being intelligent (female: *M* = 5.32, *SD* = 0.65; male: *M* = 5.32, *SD* = 1.17), attractive (female: *M* = 5.86, *SD* = 1.32; male: *M* = 4.32, SD = 1.69) and likeable (female: *M* = 6.09, *SD* = 0.87; male: *M* = 5.48, *SD* = 1.29) and acting naturally in their videos (female: *M* = 5.59, *SD* = 1.33; male: *M* = 5.07, *SD* = 1.83).

After watching the videos, participants rated their "partners" on 5 different dimensions. Following Horton and Sedikides (2009) we used 7-point rating scales regarding partner's intelligence, likeability, attractiveness and appeal as a coworker. We additionally included an item regarding the desire to meet the partner socially (see Appendix for a list of all items). After completing the ratings, participants gave the rating sheets to the experimenter who then left the room and returned a few minutes later with the rating sheets that the ostensible partners had allegedly filled out in response to watching the video recorded by the actual participant. Those rating sheets had been filled out by the experimenter beforehand.

Half of the participants received negative ratings on all dimensions (mean score across the five ratings: *M* = 1.9, *SD* = 0.74). The other half of participants received positive ratings on all dimensions (mean score across the five ratings: *M* = 6.3, *SD* = 0.67; see Appendix for a list of the exact ratings).

Debriefing status at the time of the DV measurement. In the *not debriefed yet* conditions, participants completed the dependent variables before the debriefing, following previous research (Horton & Sedikides, 2009). In the *debriefed* conditions, participants were first debriefed and then completed the dependent variables.

Debriefing. The debriefing was presented to participants on the computer screen. They read that in contrast to what they had learned earlier (1) there were no other participants, (2) the videos of the alleged other two participants had been recorded beforehand by two confederates, (3) no one actually rated the participant, (4) the ratings they had received

concerning their recorded video had been created beforehand, and that (5) they had been randomly assigned to a condition where they either received very negative (*threat* conditions) or positive (*no threat* conditions) feedback in response to the video. In order to continue with the study, participants had to confirm the following sentence by marking a checkbox: "I fully understand that there were no other participants and that therefore I have not received a real negative (*threat* conditions)/positive (*no threat* conditions) rating". All participants confirmed this.

Dependent variables.

Mood. Mood was assessed in the same way as in Study 1. After reverse-coding items indicating negative mood, internal consistency was high (α = .91).

State self-esteem. State self-esteem was assessed in the same way as in Study 1. In Study 2, participants in the ego threat condition received negative feedback concerning their intelligence, attractiveness and likeability. We therefore expected effects on all three subscales (performance, social and appearance) and analyzed the complete state self-esteem scale. Internal consistency across all three subscales was good (α = .89).

Self-perception. Self-perception was assessed in the same way as in Study 1 with the following exceptions: We used 11 focal items addressing the qualities threatened by the false feedback: well-educated, interesting, likeable, unkempt, socially competent, popular with own sex, attractive, loveable/kind[3], pleasing, popular with opposite sex, intelligent. The five distractor items addressing qualities irrelevant to the false feedback were: impatient, jealous, religious, faithful, technophilic. Internal consistency for focal items was good (α = .87).

Implicit self-esteem. Implicit self-esteem was assessed and analyzed in the same way as in Study 1 with the following exceptions: Evaluative stimuli were six positive and six negative words addressing qualities threatened by the false feedback (positive: smart,

[3] We used the German word "liebenswürdig" which means both "loveable" and "kind" at the same time.

intelligent, pretty, popular, loveable, beautiful; negative: stupid, ugly, hated, undesirable, repulsive, hard of understanding). The "self" and "other" categories consisted of six generic stimuli (self: me, own, self; other: not-me, someone else's, other) and six idiographic stimuli. To create the idiographic stimuli of the "self" category, participants indicated their initials, the day and month of their birthday, and a city they have a strong connection to (e.g., their hometown). For the "other" category, participants indicated initials, a birthday and a city they do not have any connection to. The mean error rate was 8.27%.

Follow-up. Participants answered 10 items referring to thoughts and feelings about the false feedback they had received to their video self-presentation. Sample items include "How often did thoughts about the feedback you received pop into your mind?" and "Since the experiment, how often did you think about whether others find you (un)attractive?" (5-point rating scale, 1 = *not at all*, 5 = *very much/often*; α = .83; see Appendix for a list of all items).

Demographic variables. Participants indicated their age, gender, major, and relationship status (response options "single", "short-term relationship", "long-term relationship"). In addition, sexual orientation was assessed using the Kinsey-Scale (Kinsey, Pomeroy, & Martin, 1948), a 7-point rating scale ranging from *exclusively heterosexual* to *exclusively homosexual*. We added an eighth option: "I'd rather not say".

Results

Analytic strategy. The analytic strategy was the same as in Study 1. An exception to the general procedure was the follow-up questionnaire. At the time of the follow-up questionnaire, all participants had already been debriefed. We therefore collapsed across the *debriefing status* conditions for this dependent variable.

Mood. As expected, in the *not debriefed yet* condition that follows the standard procedure in previous research (Horton & Sedikides, 2009), participants reported lower mood after *threat* (M_{Threat} = 3.00, SD = 0.91) than after *no threat* (M_{NoThreat} = 4.03, SD = 0.52), $t(30.15) = 4.41$, $p < .001$, $d = 1.39$. This establishes that the ego threat manipulation

successfully manipulated mood. The simple contrast of *ego threat condition* in the *debriefed* conditions revealed that the effect of the ego threat manipulation on mood partially persevered after the debriefing ($M_{NoThreat}$ = 3.70, *SD* = 0.63; M_{Threat} = 3.33, *SD* = 0.60), *t*(39.00) = 1.97, *p* = .056, *d* = 0.60. The main effect of *ego threat condition* was highly significant across both *debriefing status* conditions ($M_{NoThreat}$ = 3.86, *SD* = 0.59; M_{Threat} = 3.16, *SD* = 0.78), *F*(l, 77) = 21.77, *p* < .001, η_p^2 = .22. The interaction between *ego threat condition* and *debriefing status* was also significant, *F*(l, 77) = 4.72, *p* = .033, η_p^2 = .06, indicating that the effect of the ego threat manipulation on mood was stronger in the *not debriefed yet* condition than in the *debriefed* condition. Follow-up analyses revealed a marginally significant effect of *debriefing status* in the *no threat* condition, *t*(38.29) = 1.82, *p* = .076, *d* = 0.57, but not in the *threat* condition, *t*(32.89) = 1.32, *p* = .186, *d* = 0.42. After receiving positive feedback, participants tended to feel less positive when they were debriefed before completing the dependent measures. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 3.51, *SD* = 0.90; M_{Deb} = 3.52, *SD* = 0.64), *F*(l, 77) < 0.01, *p* = .992, η_p^2 < .01 (Figure 2a). Taken together, negative feedback led to decreased mood, and the debriefing was unsuccessful in eliminating this effect.

State self-esteem. In the *not debriefed yet* condition, participants reported lower state self-esteem after *threat* (M_{Threat} = 3.22, *SD* = 0.70) than after *no threat* ($M_{NoThreat}$ = 3.63, *SD* = 0.57), *t*(74) = 1.90, *p* = .062, *d* = 0.64. This establishes the detrimental effect of *threat* relative to *no threat* on state self-esteem. The planned contrast of *ego threat condition* in the *debriefed* conditions revealed that the effect of the ego threat manipulation on state self-esteem persevered after the debriefing ($M_{NoThreat}$ = 3.60, *SD* = 0.67; M_{Threat} = 3.20, *SD* = 0.66), *t*(74) = 1.95, *p* = .055, *d* = 0.60. The main effect of *ego threat condition* was significant across both *debriefing status* conditions ($M_{NoThreat}$ = 3.61, *SD* = 0.62; M_{Threat} = 3.21, *SD* = 0.67), *F*(l, 74) = 7.39, *p* = .008, η_p^2 = .09. There was no interaction between *ego threat condition* and *debriefing status F*(l, 74) < 0.01, *p* = .976, η_p^2 < .01. Finally, there was no main effect of

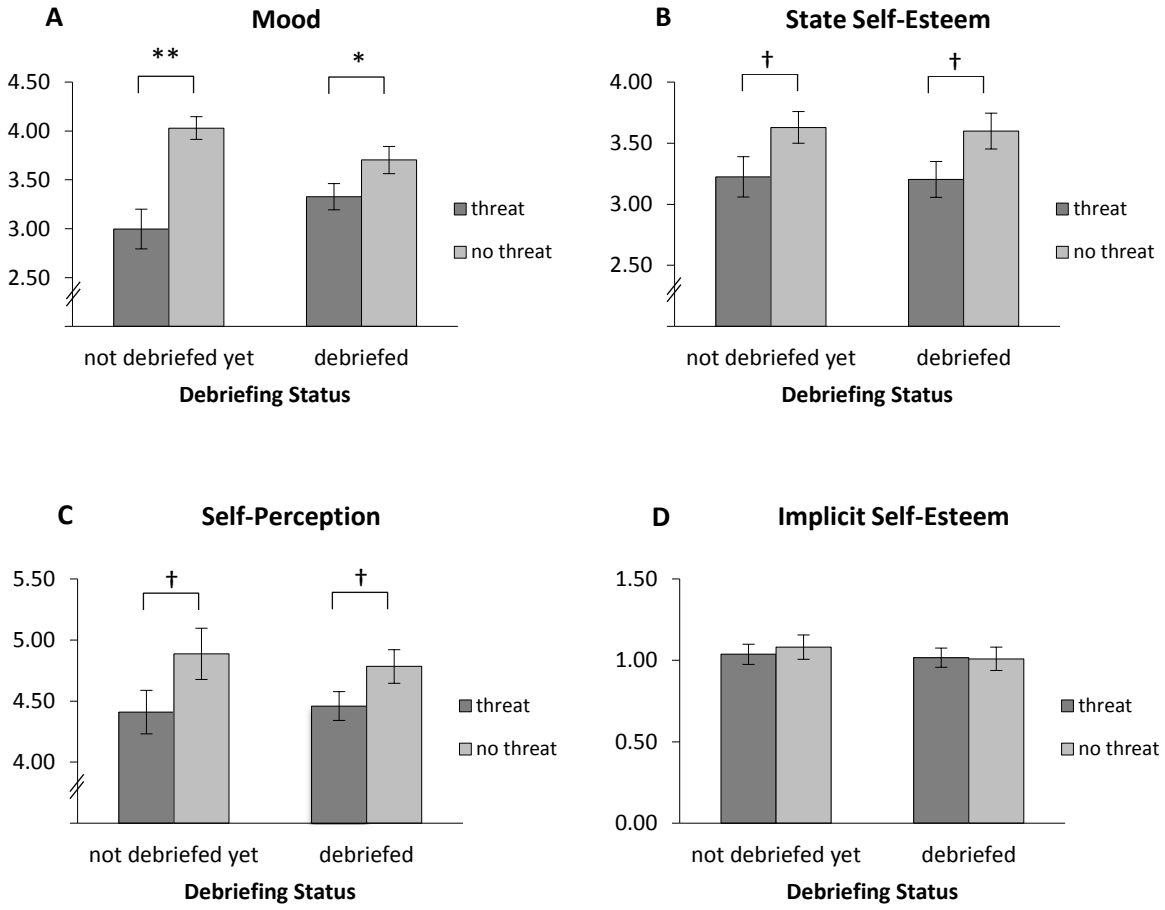
debriefing status (M_{NoDeb} = 3.43, *SD* = 0.66; M_{Deb} = 3.41, *SD* = 0.69), *F*(l, 74) = 0.03, *p* = .871, $\eta_p^2 < .01$ (Figure 2b). Taken together, negative feedback led to lower self-esteem, no matter whether participants had already received a debriefing or not.

Self-perception. In the *not debriefed yet* condition, participants showed less favorable self-perceptions after *threat* (M_{Threat} = 4.41, *SD* = 0.80) than after *no threat* ($M_{NoThreat}$ = 4.89, *SD* = 0.94), *t*(37.05) = 1.74, *p* = .090, *d* = 0.55. Thus, the ego threat manipulation successfully manipulated self-perception. The planned contrast of *ego threat condition* in the *debriefed* conditions revealed that the effect of the ego threat manipulation on self-perception tended to persevere after the debriefing ($M_{NoThreat}$ = 4.78, *SD* = 0.63; M_{Threat} = 4.46, *SD* = 0.51), *t*(37.54) = 1.79, *p* = .082, *d* = 0.55. The main effect of *ego threat condition* was significant across both *debriefing status* conditions ($M_{NoThreat}$ = 4.83, *SD* = 0.79; M_{Threat} = 4.43, *SD* = 0.67), *F*(l, 76) = 5.90, *p* = .017, η_p^2 = .07. There was no interaction between *ego threat condition* and *debriefing status F*(l, 76) = 0.22, *p* = .642, $\eta_p^2 < .01$. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 4.65, *SD* = 0.89; M_{Deb} = 4.63, *SD* = 0.59), *F*(l, 76) = 0.03, *p* = .873, $\eta_p^2 < .01$ (Figure 2c). Taken together, negative feedback led to less favorable self-perceptions, and the debriefing was unsuccessful in changing that.

Implicit self-esteem. In the *not debriefed yet* condition, participants showed similar levels of implicit self-esteem after *threat* (M_{Threat} = 1.04, *SD* = 0.29) as after *no threat* ($M_{NoThreat}$ = 1.08, *SD* = 0.33), *t*(77) = 0.45, *p* = .652, *d* = 0.13 (Figure 2d). Thus, the ego threat manipulation had no significant effect on implicit self-esteem. Therefore, we stopped the analysis of this dependent variable at this point.

Follow-up. Fifty-six participants completed the online follow-up questionnaire (69% retention rate) on average 20.99 hours after they had left the lab (*Md* = 11.80 hours, *SD* = 26.81). All participants had been debriefed when they filled out the follow-up questionnaire. Therefore, only the effect of the *ego threat condition* could be meaningfully analyzed. A one-way analysis of variance (ANOVA) revealed that after *threat* compared to *no threat*

participants reported thinking about the manipulation more often and feeling affected by it in a more negative way ($M_{NoThreat}$ = 2.49, *SD* = 0.58; M_{Threat} = 3.03, *SD* = 0.79), $F(1,55) = 8.25$, $p = .006$, $\eta_p^2 = .13$



Figures 2a–2d. Mood, state self-esteem, self-perception, and implicit self-esteem as a function of *ego threat condition* and *debriefing status*. Participants who had received negative feedback showed lower levels of mood and state self-esteem, as well as less favorable self-perceptions than participants who had received positive feedback, no matter whether they had been debriefed at the time of the collection of the respective dependent variable or not. Error bars indicate +/- 1 SEM. †$p < .10$,*$p < .05$, **$p < .01$.

Discussion

Negative social feedback led to lower mood, state self-esteem and self-perception as compared to positive social feedback. Similar to Study 1, a thorough Revised Outcome Debriefing procedure failed to eliminate these effects. Participants learned that in fact there

were no other participants in the study and that the feedback they had received was false. However, this did little to ameliorate the aversive effects of the ego threat manipulation. Additionally, participants in the *threat* conditions continued to have more negative thoughts about the experiment several hours after leaving the laboratory even though all participants had already been thoroughly debriefed at that point. Taken together, the debriefing failed to undo the psychological harm caused by the ego threat manipulation both in the short-term during the laboratory session and also more enduringly during the day after participating in the study.

Study 3

In Studies 1 and 2, participants indicated their mood, state self-esteem and self-perception via self-report questionnaires. The critical reader may be concerned whether the results are (at least partly) driven by demand effects—participants may have sensed what the researchers "wanted to see". We consider this possibility rather unlikely for two reasons. First, if participants tried to "cooperatively" respond in a way they thought researchers would want them to respond they should have done so equally for both the *ego threat condition* and the *debriefing status* conditions. That is, after *threat*, participants should have reported better well-being when the debriefing preceded the assessment of the dependent variables than in the conditions in which the dependent variables preceded the debriefing. This was not the case. We deem it unlikely that participants sensed demand characteristics of the ego threat manipulation, but not the debriefing.

Second, in Study 2, effects persevered in the follow-up several hours after the lab session. Any demand characteristic of the ego threat manipulation is likely to have fainted hours after the manipulation when responding from outside the lab via the Internet. Nevertheless, differences between the ego threat conditions were remarkably strong at follow-up.

These arguments notwithstanding, Study 3 focused on dependent variables that are

even less likely to be influenced by demand characteristics: actual behavior and ratings of other people's behavior. We used the same social rejection ego threat manipulation as in Study 2 (Horton & Sedikides, 2009), and measured aggressive behavior and hostile attributions. Both aggressive behavior and hostile attributions were affected by ego threat in previous research (DeWall, Twenge, Gitter, & Baumeister, 2009; Twenge, Baumeister, Tice, & Stucke, 2001).

Method

Participants and design. Study 3 realized a combination of ego threat manipulation and dependent variables that has – to our knowledge – never been realized before. It was therefore most adequate to base an expected effect size on all dependent variables used in all previous studies (Studies 1 and 2, weighted by sample size). A power analysis using G*Power (Faul et al., 2007) revealed a minimal sample size of $N = 75$ for at least 80% power ($\alpha = .05$, $f = 0.33$). We therefore recruited a minimum of 80 participants (20 per condition; see Simmons et al., 2011), but as many as possible until the end of the semester.

Ninety-three students (58% females; mean age $M = 24.53$, $SD = 4.61$) of various disciplines from a German university (homogenous sample with respect to ethnicity) participated in exchange for €5 and an energy drink. They were randomly assigned to one of four conditions of a 2 (ego threat condition: yes vs. no) × 2 (debriefing status: debriefed vs. not debriefed yet) design and run in individual sessions that lasted about 35 minutes. Dependent variables were aggressive behavior and hostile attributions. To prevent the inclusion of participants who could become suspicious of the hypothesis, no psychology students were tested. No participants were excluded from the analyses.

Procedure. The ego threat manipulation was the same as in Study 2. After the ego threat manipulation, participants in the *not debriefed yet* conditions were told that it was necessary to take a break before they could continue with the study. They were asked to help the experimenter with the preparation of another study during this waiting period. The

preparation consisted of the allocation of hot sauce, allegedly for another participant to consume in an ostensible taste test. After the allocation of hot sauce, participants were told that they could continue with the main study and were asked to fill out questionnaires on the computer. Participants then filled out a hostile attribution questionnaire as well as demographics. They then received the same debriefing as in Study 2.

Participants in the *debriefed* conditions received the debriefing *before* allocating hot sauce and filling out the questionnaires. After the debriefing and completion of the dependent measures, participants were asked to fill out an online follow-up questionnaire later on the same day. Then they were thanked and dismissed. Later on the same day, participants received the link to the online follow-up questionnaire via email. To start the questionnaire, participants had to enter a password that we had sent them. These passwords were personalized and conveyed information about each participant's experimental condition from the lab session before. We did not retrieve any other information from the passwords, including any information about participants' identity. Therefore, anonymity was guaranteed for all participants.

Debriefing. The debriefing was the same as in Study 2.

Experimental manipulations. The ego threat manipulation as well as the debriefing status manipulation were the same as in Study 2.

Dependent variables.

Aggressive behavior. The *Hot Sauce Paradigm* (Lieberman, Solomon, Greenberg, & McGregor, 1999) was used to measure aggressive behavior. On an information sheet, participants read about a taste experiment that was allegedly about to take place in the laboratory on the same day. Participants were asked to choose an amount of hot sauce as well as the intensity of the sauce for another ostensible participant (ranging from 1 = *least hot* to 4 = *most hot*). They were informed that they were not allowed to mix sauces and that the other participant had to drink the whole amount of hot sauce that they would choose. For the

allocation of hot sauce, participants were provided with a paper cup with a lid and four plastic squeeze bottles with hot sauce that had numbers from one to four on them (referring to the hotness of the respective sauce).

After the lab session, when participants had left the lab, the experimenter weighed the cup prepared by the participants as well as the four bottles with hot sauce in them to determine the weight difference (each bottle has been weighed before participants came into the lab) and therefore how many grams of hot sauce participants allocated and which sauce they had chosen.

Following previous research, aggressive behavior was defined as the sum of the standardized hotness rating of the selected sauce and the weight in grams (see Barlett, Branch, Rodeheffer, & Harris, 2009).

Hostile attributions. To measure hostile attributions, we used the Social Information Processing-Attribution and Emotional Response Questionnaire (SIP-AEQ; Coccaro, Noblett, & McCloskey, 2009). The SIP-AEQ consists of eight vignettes which describe aversive and socially ambiguous actions conducted by one or more persons (see Appendix for an example). For each vignette, participants were asked to rate the likelihood of four different attributions of the actors' intents on a 4-point rating scale (0 = *not at all likely*, 3 = *very likely*). The four different attributions reflect benign attribution, instrumental attribution and hostile attribution which consists of direct hostile attribution and indirect hostile attribution. Internal consistency for hostile attribution items was good (α = .83). For each of the attribution subscales, we calculated the mean across all items of the respective subscale for every participant.

Follow-up. Participants answered 12 items referring to feelings about the false feedback they had received, about whether the experiment had affected them negatively, about the experimenter and about how they liked the study (7-point rating scale, 1 = *not at all*, 7 = *very much*; α = .76, see Appendix for a list of all items).

Results

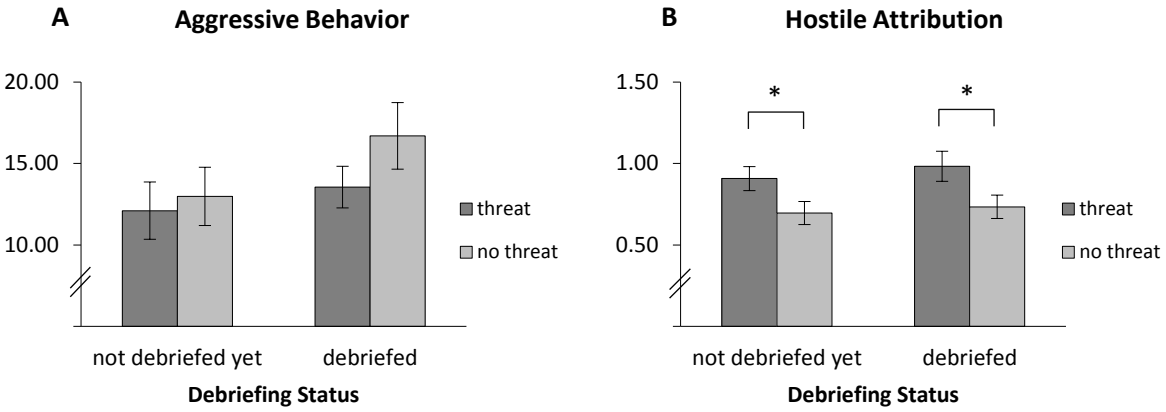
Analytic strategy. The analytic strategy was the same as in Studies 1 and 2.

Aggressive behavior. In the *not debriefed yet* condition, participants showed similar levels of aggressive behavior after *threat* (M_{Threat} = 12.10, *SD* = 8.43) as after *no threat* ($M_{NoThreat}$ = 12.98, *SD* = 8.57), *t*(89) = 0.35, *p* = .725, *d* = 0.10 (Figure 3a). Thus, the ego threat manipulation had no significant effect on aggressive behavior. Therefore, we stopped the analysis of this dependent variable at this point.

Hostile attribution. In the *not debriefed yet* condition, participants showed more hostile attributions after *threat* (M_{Threat} = 0.91, *SD* = 0.36) than after *no threat* ($M_{NoThreat}$ = 0.70, *SD* = 0.34), *t*(43.87) = 2.07, *p* = .044, *d* = 0.60. Thus, the ego threat manipulation successfully manipulated the hostile attribution bias. The planned contrast of *ego threat condition* in the *debriefed* conditions revealed that the effect of the ego threat manipulation on hostile attribution persevered after the debriefing ($M_{NoThreat}$ = 0.73, *SD* = 0.35; M_{Threat} = 0.98, *SD* = 0.44), *t*(42.16) = 2.13, *p* = .039, *d* = 0.63. The main effect of *ego threat condition* was significant across both *debriefing status* conditions ($M_{NoThreat}$ = 0.72, *SD* = 0.34; M_{Threat} = 0.95, *SD* = 0.40), *F*(l, 89) = 8.77, *p* = .004, η_p^2 = .09. There was no interaction *between ego threat condition* and *debriefing status F*(l, 89) = 0.15, *p* = .903, η_p^2 < .01. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 0.80, *SD* = 0.36; M_{Deb} = 0.86, *SD* = 0.42), *F*(l, 89) = 0.74, *p* = .393, η_p^2 < .01 (Figure 3b). Taken together, negative feedback led to more hostile attributions, and the debriefing was unsuccessful in changing that.

Follow-up questionnaire. Eighty-one participants completed the online follow-up questionnaire (89% retention rate) on average 45.93 hours after leaving our lab (*Md* = 24.00 hours, *SD* = 83.67). All participants had been debriefed when they filled out the follow-up questionnaire. Therefore, only the effect of the *ego threat condition* could be meaningfully analyzed. A one-way analysis of variance (ANOVA) revealed that after *threat* compared to *no threat* participants evaluated the experimenter less favorable and reported liking the study less

as well as feeling affected by it in a more negative way, even several hours after taking part in the study ($M_{NoThreat}$ = 5.32, *SD* = 0.84; M_{Threat} = 4.12, *SD* = 3.05), $F(1,81) = 5.88$, $p = .018$, $\eta_p^2 = .07$.



Figures 3a–3b. Aggressive behavior and hostile attribution as a function of *ego threat condition* and *debriefing status*. Participants who had received negative feedback showed higher levels of hostile attribution than participants who had received positive feedback, no matter whether they had been debriefed at the time of the collection of this dependent variable or not. Error bars indicate +/- 1 SEM. *$p < .05$.

Discussion

Ego threat through negative social feedback led to more hostile attributions, but not more aggressive behavior. A thorough debriefing procedure did not eliminate the effects on hostile attributions. Additionally, following the negative compared to positive social feedback participants continued to have more negative thoughts about the experiment almost two days after leaving the lab. Taken together, the debriefing failed to undo the effects caused by the social rejection both in the short-term during the laboratory session and also more enduringly during the days after participating in the study. These findings add to Studies 1 and 2 by showing that perseverance of negative effects caused by an ego threat manipulation is not limited to subjective well-being, but can generalize to other negative consequences such as hostile attributions.

There was no effect of the manipulation on aggressive behavior as indicated by hot sauce allocation, which contrasts with previous research about hot sauce allocation after social rejection (Wesselmann, Butler, Williams, & Pickett, 2010) and aggressive behavior after social rejection in general (Twenge et al., 2001; Twenge & Campbell, 2003). One possible reason for this finding is that the negative feedback manipulation made participants feel sad and become wary (hence the effect on hostile attributions), but did not make them angry. This hypothesis is supported by Study 2 that employed the same manipulation and assessed mood as a dependent variable: looking at the items of the mood scale separately, we found that participants who had received negative feedback did not differ on their reported anger from participants who had received positive feedback ($M_{\text{NoThreat}} = 1.24$, $SD = 0.58$; $M_{\text{Threat}} = 1.45$, $SD = 0.90$), $F(1,79) = 1.49$, $p = .225$, $\eta_p^2 = .02$).

Study 4

Studies 1–3 conceptually replicated that ego-threatening manipulations can impair subjective well-being and increase hostile attributions. Three studies consistently showed that these effects persevered after a Revised Outcome Debriefing, arguably the most effective debriefing procedure (McFarland et al., 2007). Moreover, even several hours after participants had left the laboratory, detrimental effects of the ego threat manipulation persevered.

In light of these findings, Study 4 investigated two additional means to re-establish participants' pre-study conditions:

First, instead of a written debriefing we included a standardized verbal debriefing, presented by a carefully trained experimenter. One of the benefits of a verbally presented debriefing is that participants get the opportunity to ask questions during the debriefing procedure. Another benefit is that the experimenter, too, can ask participants whether they fully understood what they have been told during the debriefing procedure.

Second, in an additional condition we investigated the effectiveness of yet another intervention to remedy the adverse effects of the ego threat manipulation: a self-affirmation

procedure. Self-affirmation is a method to restore a person's positive self-image after it has been threatened. By asking a person to focus on important aspects of their self-image that differ from the one that has been threatened, the person's overall self-adequacy is reinforced. A large literature suggests that self-affirmation reduces rumination and increases various aspects of well-being including mood and self-esteem (e.g., Klein et al., 2001; Koole et al., 1999). For example, in one study participants had experienced a threat to their self-image when they had to work on an ostensible IQ test that was impossible to solve. When they engaged in a self-affirmation task (filling out a subscale concerned with a value important to them) after being threatened, participants subsequently reported better mood and showed higher implicit self-esteem than participants who had not worked on a self-affirming task (Koole et al., 1999).

In sum, we expected ego threat to impair subjective well-being and investigated the effectiveness of three different means to counteract these effects: (a) written debriefing, (b) verbal debriefing, and (c) a self-affirmation procedure.

Method

Participants and design. The expected effect size was based on the mean of all in previous studies observed effect sizes (weighted by sample size) of the dependent variables employed in Study 4. A power analysis using G*Power (Faul et al., 2007) revealed a minimal sample size of $N = 85$ for at least 80% power ($\alpha = .05$, $f = 0.39$). We therefore recruited a minimum of 100 participants (20 per condition; see Simmons et al., 2011), but as many as possible until the end of the semester.

One hundred and fifty-seven students (60% females; mean age $M = 23.67$, $SD = 4.10$) of various disciplines from a German university (homogenous sample with respect to ethnicity) participated in exchange for €5 and an energy drink. They were randomly assigned to one of five conditions (no threat not debriefed yet vs. threat not debriefed yet vs. threat written debriefing vs. threat

verbal debriefing vs. threat self-affirmation)[4] in a one-factorial design and run in individual sessions that lasted about 35 minutes. Dependent variables were mood, state self-esteem, and self-perception. Part of the study was an online follow-up. To prevent the inclusion of participants who could become suspicious of the hypothesis, no psychology students were tested. Thirteen participants were excluded from the study because they had participated in another false feedback study in the same laboratory a few days earlier. The final sample consisted of 144 participants (58% females; mean age $M = 23.66$, $SD = 4.09$).

Procedure. After giving informed consent, participants experienced the same social rejection manipulation as in Studies 2 and 3 in which they were told that two ostensible other participants found them (un)intelligent, (un)likeable and (un)attractive and would (not) want to work with them or meet them in person.

After the ego threat manipulation, participants indicated their demographics (age, gender, and study major). Then, participants in the *no threat not debriefed yet* condition and participants in the *threat not debriefed yet* condition wrote an essay as a filler task, completed the dependent variables (state self-esteem, mood, self-perception) and received the same written debriefing as in Studies 2 and 3. Participants in the *threat written debriefing* condition and in the *threat verbal debriefing* condition received the debriefing (written or verbal, respectively) before writing an essay as a filler task and completed the dependent variables subsequently. Participants in the *threat self-affirmation* condition received a written debriefing before writing a self-affirmation essay and then completed the dependent variables.

At the end of the study, participants were asked to fill out an online follow-up questionnaire later on the same day, were thanked and dismissed. Later on the same day,

[4] From Study 4 on, we skipped the *no threat debriefed* condition. In Studies 1–3 we established that there was perseverance of the negative effects caused by negative feedback even after a debriefing procedure. This difference can also be shown through a comparison between the *threat debriefed* condition with the *threat not debriefed yet* condition or the *no threat not debriefed yet* condition, respectively. For economic reasons we therefore continued without the *no threat debriefed* condition.

participants received the link to the online questionnaire via email. To start the questionnaire, participants had to enter an individual password that we had sent them. These passwords were saved in the final data set and conveyed information about each participant's experimental condition from the lab session. We did not retrieve any other information from the passwords. Therefore, anonymity was guaranteed for all participants.

Ego threat manipulation. The ego threat manipulation was the same social rejection manipulation in an online dating context as in Studies 2 and 3.

Debriefing. The written debriefing was the same as in Studies 2 and 3. The wording of the verbal debriefing was identical to the written debriefing but presented in person by the experimenter who had been carefully trained to convey the debriefing in a standardized manner. At the end of the debriefing, the experimenter asked a standardized question to learn whether each participant had fully understood the content of the debriefing.

Self-affirmation. Participants in the *threat self-affirmation* condition wrote a self-affirming essay for eight minutes about how good they felt about themselves when they had helped other people in the past (for the instruction see Appendix). In past research, writing about one's own prosocial behavior (Lekes, Guilbault, Philippe, & Houle, 2014; Weinstein & Ryan, 2010) and about being proud of oneself (Blanton, Pelham, DeHart, & Carvallo, 2001; Klein et al., 2001) has raised subjective well-being.

Participants in the four conditions without a self-affirmation intervention were asked to write an eight-minute description of the shoes they were wearing that day as a filler task (for the complete instructions, see Appendix). Describing one's shoes in great detail has been used as a non-affirming, emotionally neutral control task in previous self-affirmation studies (Burton & King, 2009; King & Miner, 2000).

Dependent variables.

Mood. Mood was assessed in the same way as in Studies 1–3 with 12 items taken from the PANAS-X (Watson, & Clark, 1999). After reverse-coding items indicating negative

mood, internal consistency was good (α = .87).

State self-esteem. State self-esteem was assessed in the same way as in Studies 1–3 with the German version (Rudolph et al., 2009) of the State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991). Internal consistency for was good (α = .85).

Self-perception. Self-perception was assessed in the same way as in Studies 2 and 3 with a modification of the How I See Myself questionnaire (HSM; Taylor, Lerner, Sherman, Sage, & McDowell, 2003). Internal consistency for the focal items was good (α = .86).

Follow-up. Participants answered the same 12 items as in Study 3. The items referred to feelings about the false feedback participants received, about whether the experiment affected them negatively, perceptions about the experimenter and about how they liked the study. Internal consistency was good (α = .80).

Results

Analytic strategy. For every dependent variable, we first examined the predicted effect of *threat* relative to *no threat* in the *not debriefed yet* conditions to establish the focal effect that the debriefing and/or the self-affirmation procedure would aim to remedy. If there was evidence for this effect, we compared the *threat written debriefing*, the *threat verbal debriefing* and the *threat self-affirmation* condition to the *threat not debriefed yet* condition to examine whether these procedures significantly counteracted the effect of the ego threat manipulation. If there was no significant difference, the respective intervention was not able to appreciably reduce the effect of the ego threat.

If there was a difference between the *threat not debriefed yet* condition and any of the conditions aiming to ameliorate the effect of the ego threat, this indicated that the intervention either partially or completely eliminated the negative effects caused by the ego threat manipulation. In this case, we compared the respective intervention condition to the *no threat not debriefed yet* condition to examine whether there still was a difference between participants who had received negative feedback plus an intervention and participants who

had received positive feedback.

If, in the first step, there was no significant effect of the ego threat manipulation on the dependent variable our central research question referring to the effectiveness of different debriefing and self-affirmation procedures could not be examined. Therefore, we do not report the full set of analyses in these cases and instead move on to the next dependent variable.

An exception to this general procedure was the follow-up questionnaire. At the time of the follow-up questionnaire, all participants had already been debriefed. We therefore collapsed across all *threat* conditions for this dependent variable.

Mood. As expected, in the *not debriefed yet* conditions, participants reported lower mood after *threat* (M = 3.30, SD = 0.72) than after *no threat* (M = 3.93, SD = 0.51), $t(139) = 3.75$, $p < .001$, $d = 1.01$. This establishes that the ego threat manipulation successfully manipulated mood. The comparison between the *threat written debriefing* condition and the *threat not debriefed yet* condition revealed no significant difference (M_{WritDeb} = 3.50, SD = 0.67), $t(139) = 1.21$, $p = .226$, $d = 0.29$. Thus, the written debriefing was not effective in restoring participants' well-being.

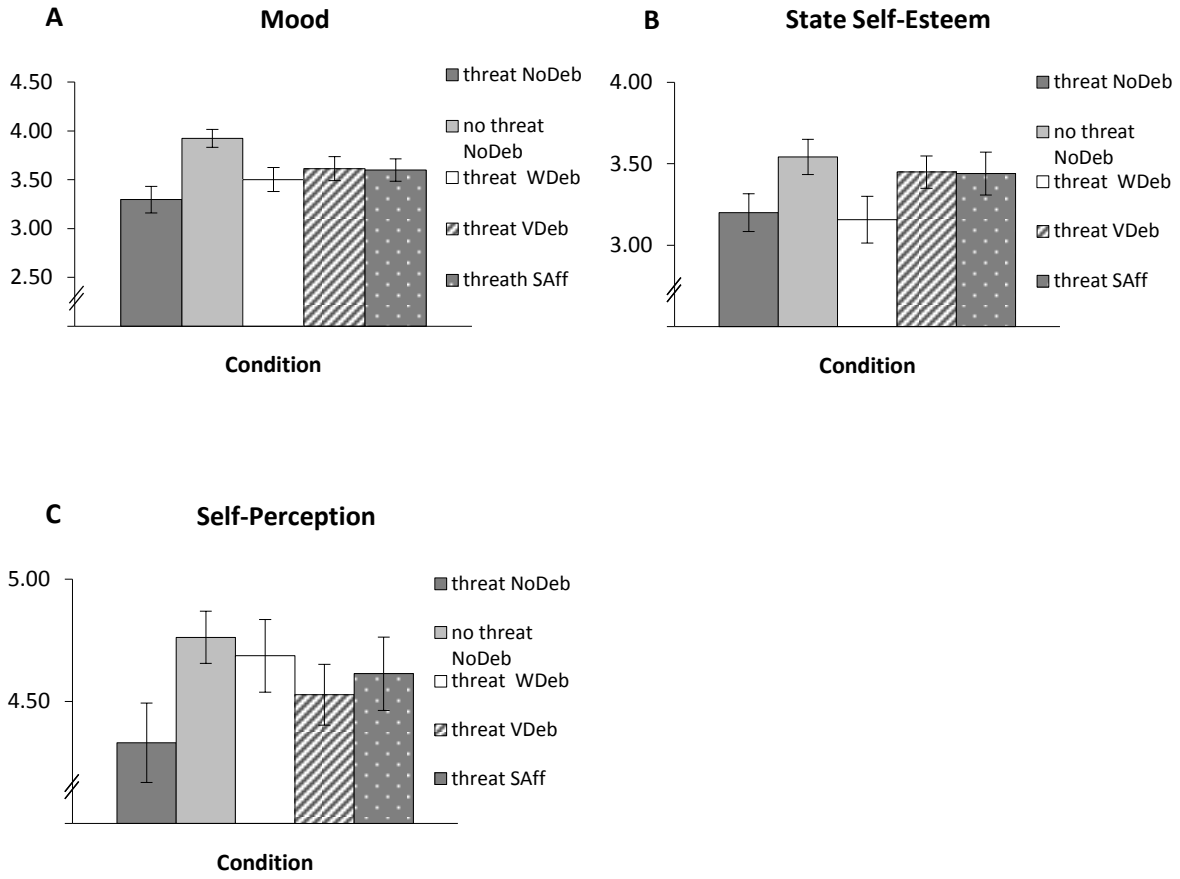
Receiving a verbal debriefing (*threat verbal debriefing* condition) or writing a self-affirming essay (*threat self-affirmation* condition) led to marginally better mood compared to the *threat not debriefed yet* condition (M_{VerbDeb} = 3.61, SD = 0.67; $M_{\text{Self-Aff}}$ = 3.60, SD = 0.60), both ts(139) < 1.89, both ps > .061, both ds = 0.45. Both the *Threat verbal debriefing* as well as the *threat self-affirmation* condition still differed marginally from the *no threat not debriefed yet* condition, both ts(139) < 1.93, both ps > .055, both ds < 0.60 (Figure 4a). Thus, the verbal debriefing and the self-affirmation intervention partially reduced, but did not completely eliminate the detrimental effects of the ego threat manipulation on participants' mood.

State self-esteem. As expected, in the *not debriefed yet* conditions, participants

reported lower levels of state self-esteem after *threat* (M = 3.20, SD = 0.62) than after *no threat* (M = 3.54, SD = 0.60), $t(139) = 2.02$, $p = .045$, $d = 0.56$. This establishes that the ego threat manipulation successfully manipulated state self-esteem. None of the intervention conditions (*threat written debriefing*, *threat verbal debriefing*, *threat self-affirmation*) differed significantly from the *threat not debriefed yet* condition, suggesting that none of them was effective in restoring participants' state self-esteem after threat (M_{WritDeb} = 3.16, SD = 0.77; M_{VerbDeb} = 3.45, SD = 0.54; $M_{\text{Self-Aff}}$ = 3.44, SD = 0.68), all ts(139) < 1.47, all ps > .144, all ds < 0.43 (Figure 4b).

Self-perception. As expected, in the *not debriefed yet* conditions, participants showed less favorable self-perceptions after *threat* (M = 4.33, SD = 0.84) than after *no threat* (M = 4.76, SD = 0.57), $t(137) = 2.18$, $p = .031$, $d = 0.60$. This establishes that the ego threat manipulation successfully manipulated self-perception. None of the intervention conditions (*threat written debriefing*, *threat verbal debriefing*, *threat self-affirmation*) differed significantly from the *threat not debriefed yet* condition, suggesting that none of them was effective in restoring participants' self-perceptions after *threat* (M_{WritDeb} = 4.69, SD = 0.80; M_{VerbDeb} = 4.53, SD = 0.68; $M_{\text{Self-Aff}}$ = 4.61, SD = 0.79), all ts(137) < 1.62, all ps > .111, all ds < 0.44 (Figure 4c).

Follow-up questionnaire. One hundred twenty-two participants completed the online follow-up questionnaire (85% retention rate) on average 46.73 hours after leaving the lab (Md = 26.10 hours, SD = 43.67). All participants had been debriefed when they filled out the follow-up questionnaire. Therefore, only the effect of the ego threat manipulation could be meaningfully analyzed. A one-way analysis of variance (ANOVA) revealed that participants who had experienced *threat* compared to *no threat* evaluated the experimenter less favorable, reported liking the study less, and reported feeling affected by it in a more negative way, even almost two days after taking part in the study (M_{NoThreat} = 5.47, SD = 0.73; $M_{\text{EgoThreat}}$ = 4.48, SD = 0.79), $F(1,120) = 31.90$, $p < .001$, $\eta_p^2 = .29$.



Figures 4a–4c. Mood, state self-esteem, and self-perception as a function of five conditions: *threat not debriefed yet* (NoDeb) vs. *no threat not debriefed yet* vs. *threat written debriefing* (WDeb) vs. *threat verbal debriefing* (VDeb) vs. *threat self-affirmation* (SAff). Participants who had received negative feedback showed lower levels of mood and state self-esteem, as well as less favorable self-perceptions than participants who had received positive feedback, no matter whether they had received a written debriefing or a verbal debriefing, or had written a self-affirmation essay before the collection of the respective dependent variable. Error bars indicate +/- 1 SEM.

Discussion

Ego threat by negative social feedback decreased mood and state self-esteem and led to less favorable self-perceptions compared to positive social feedback. Neither a written debriefing, a verbal debriefing, nor a written debriefing plus writing a self-affirming essay eliminated the aversive effects of the ego threat, although a verbal debriefing and a written debriefing plus self-affirmation reduced effects on mood. Additionally, following the ego-threatening feedback participants continued to have more negative thoughts about the

experiment almost two days later even though all participants had been thoroughly debriefed.

Study 5

Study 5 provides a conceptual replication of Study 4: it investigated written debriefing, verbal debriefing, and self-affirmation as interventions against the perseverance of aversive effects caused by a different ego-threatening manipulation. To ensure that the effects obtained in Study 4 would generalize to a different ego threat manipulation, in Study 5 similar to Study 1, participants received *indirect failure feedback*. They were led to believe that they performed poorly in an ostensible IQ Test supposedly predictive of future academic success.

Method

Participants and design. The expected effect size was based on the mean of all in previous studies observed effect sizes (weighted by sample size) of the dependent variables employed in Study 5. A power analysis with G*Power (Faul et al., 2007) revealed a minimal sample size of $N = 90$ for at least 80% power ($\alpha = .05$, $f = 0.38$). We therefore recruited a minimum of 100 participants (20 per condition; see Simmons et al., 2011), but as many as possible until the end of the semester.

One hundred twenty-three students (59% females; mean age $M = 24.53$, $SD = 4.61$) of various disciplines from a German university (homogenous sample with respect to ethnicity) participated in exchange for €4 and an energy drink. They were randomly assigned to one of five conditions (no threat not debriefed yet vs. threat not debriefed yet vs. threat written debriefing vs. threat verbal debriefing vs. threat self-affirmation) in a one-factorial design and run in individual sessions that lasted about 25 minutes. Dependent variables were mood, state self-esteem and self-perception. Part of the study was an online follow-up. To prevent the inclusion of participants who could become suspicious of the hypothesis, no psychology students were tested. No participants were excluded from the analyses.

Procedure. After giving informed consent, participants experienced the same ego-threatening failure feedback manipulation as in Study 1 in which they were either asked to

complete an ostensible IQ test that had in fact very difficult items and was described as being a strong indicator of future academic achievement (*threat* conditions) or they were led to believe that they participated in a pilot study and were asked to test several items' difficulties that were in fact moderately low (*no threat* condition). The procedure after the experimental manipulation was the same as in Study 4.

Debriefing. The written debriefing was the same as in Study 1. The wording of the verbal debriefing was identical to the written debriefing but presented in person by the experimenter who had been carefully trained to convey the debriefing in a standardized manner. At the end of the debriefing, the experimenter asked a standardized question to learn whether each participant had fully understood the content of the debriefing.

Ego threat manipulation. Manipulation of ego threat was the same indirect failure feedback manipulation as in Study 1.

Self-affirmation. The self-affirmation intervention was the same as in Study 4.

Dependent variables.

State self-esteem. State self-esteem was assessed in the same way as in Studies 1–4 with the German version (Rudolph et al., 2009) of the State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991). As in Study 1, we only analyzed the performance state self-esteem subscale. Internal consistency for this subscale was good (α = .87).

Mood. Mood was assessed in the same way as in Studies 1–4 with 12 items taken from the PANAS-X (Watson, & Clark, 1999). After reverse-coding items indicating negative mood, internal consistency was good (α = .87).

Self-perception. Self-perception was assessed in the same way as in Study 1 with a modification of the How I See Myself questionnaire (HSM; Taylor, Lerner, Sherman, Sage, & McDowell, 2003). Internal consistency for the focal items was high (α = .90).

Follow-up. Participants answered the same 12 items as in Studies 3 and 4. The items referred to feelings about the false feedback participants received, about whether the

experiment affected them negatively, about the experimenter and about how they liked the study. Internal consistency was good (α = .83).

Results

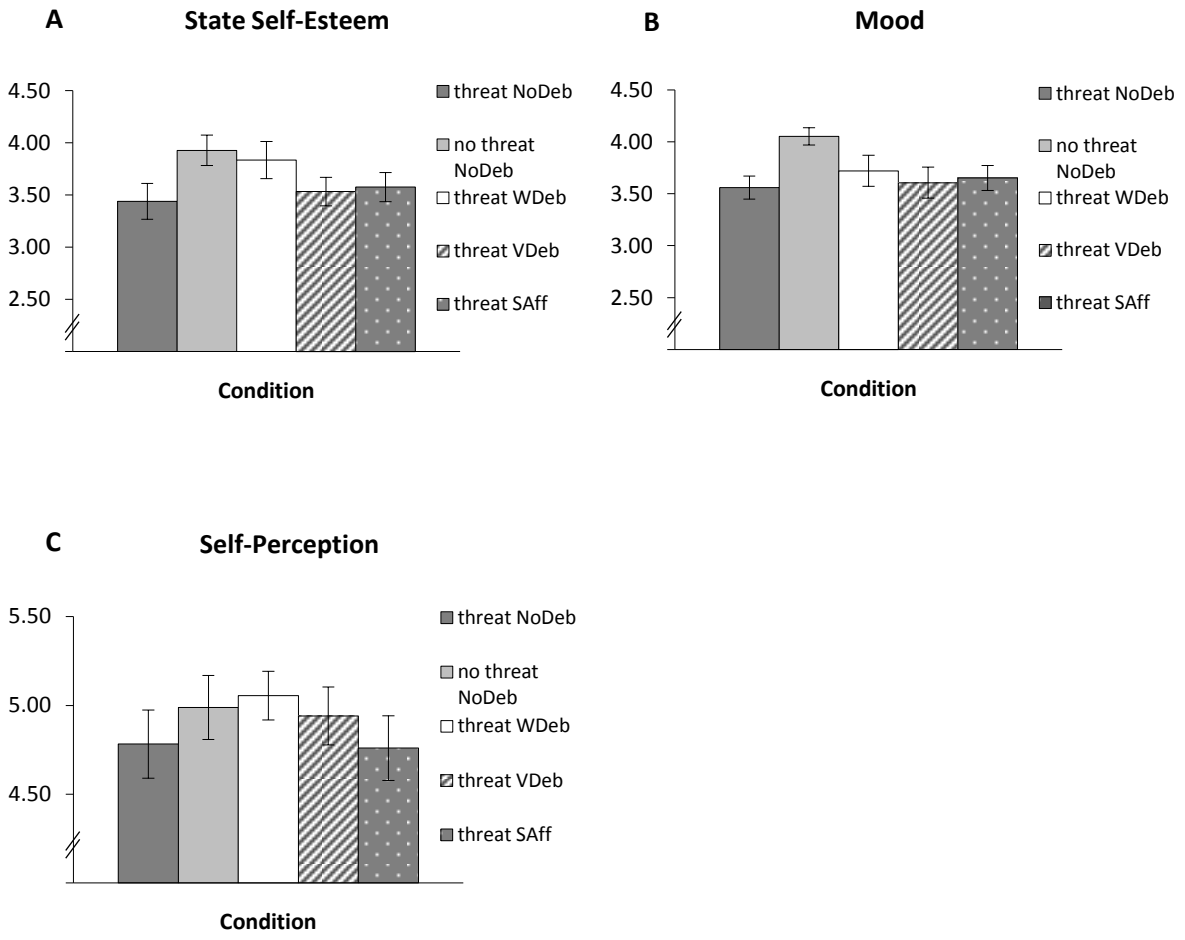
Analytic strategy. The analytic strategy was the same as in Study 4.

State self-esteem. As expected, in the *not debriefed yet* conditions, participants reported lower performance state self-esteem after *threat* (M = 3.44, SD = 0.86) than after *no threat* (M = 3.93, SD = 0.73), $t(117) = 2.26$, $p = .026$, $d = 0.61$. This establishes that the ego threat manipulation successfully manipulated state self-esteem. None of the intervention conditions (*threat written debriefing*, *threat verbal debriefing*, *threat self-affirmation*) differed significantly from the *threat not debriefed yet* condition, suggesting that none of them was effective in restoring participants' state self-esteem after ego threat (M_{WritDeb} = 3.83, SD = 0.85; M_{VerbDeb} = 3.53, SD = 0.67; $M_{\text{Self-Aff}}$ = 3.58, SD = 0.70), all ts(117) < 1.60, all ps > .117, all ds < 0.46 (Figure 5a).

Mood. As expected, in the *not debriefed yet* conditions, participants reported lower mood after *threat* (M = 3.56, SD = 0.55) than after *no threat* (M = 4.05, SD = 0.42), $t(117) = 2.85$, $p = .005$, $d = 1.00$. This establishes that the ego threat manipulation successfully manipulated mood. None of the intervention conditions (*threat written debriefing*, *threat verbal debriefing*, *threat self-affirmation*) differed significantly from the *threat not debriefed yet* condition, suggesting that none of them was effective in restoring participants' mood after ego threat (M_{WritDeb} = 3.72, SD = 0.72; M_{VerbDeb} = 3.61, SD = 0.73; $M_{\text{Self-Aff}}$ = 3.65, SD = 0.60), all ts(117) < 0.91, all ps > .364, all ds < 0.25 (Figure 5b).

Self-perception. Surprisingly, in the *not debriefed yet* conditions, participants showed similar self-perceptions after *threat* (M = 4.78, SD = 0.96) as after *no threat* (M = 4.99, SD = 0.90), $t(117) = 0.85$, $p = .397$, $d = 0.23$ (Figure 5c). Thus, the ego threat manipulation had no significant effect on self-perception. Therefore, we stopped the analysis of this dependent variable at this point.

Follow-up questionnaire. One hundred eighteen participants completed the online follow-up questionnaire (96% retention rate) on average 43.10 hours (*Md* = 27.89 hours, *SD* = 44.48) after they had left the lab. All participants had been debriefed when they filled out the follow-up questionnaire. Therefore, only the effect of the ego threat manipulation could be meaningfully analyzed. A one-way analysis of variance (ANOVA) revealed that participants who had experienced *threat* compared to *no threat* evaluated the experimenter less favorable and reported liking the study less as well as feeling affected by it in a more negative way, even almost two days after taking part in the study (M_{NoThreat} = 5.63, *SD* = 0.65; $M_{\text{EgoThreat}}$ = 4.48, *SD* = 0.84), $F(1,116) = 39.40$, $p < .001$, $\eta_p^2 = .41$.



Figures 5a–5c. State self-esteem, mood, and self-perception as a function of five conditions: *threat not debriefed yet* (NoDeb) vs. *no threat not debriefed yet* vs. *threat written debriefing* (WDeb) vs. *threat verbal debriefing* (VDeb) vs. *threat self-affirmation* (SAff). Participants who had received negative feedback showed lower levels of mood and state self-esteem than

participants who had received no feedback, no matter whether they had received a written debriefing or a verbal debriefing, or had written a self-affirmation essay before the collection of the respective dependent variable. Error bars indicate +/- 1 SEM.

Discussion

Ego threat by indirect failure feedback decreased mood and state self-esteem compared to no threat. Neither a written debriefing, a verbal debriefing, nor a written debriefing plus writing a self-affirming essay eliminated the aversive effects of the ego threat manipulation. Additionally, following the ego-threatening feedback participants continued to have more negative thoughts about the experiment almost two days later even though all participants had been thoroughly debriefed.

Internal Meta-Analytical Summary

Four studies consistently showed that negative compared to positive or neutral feedback led to lower well-being, no matter whether or not participants had received a debriefing before their well-being was measured. One additional study (Study 3) showed that threat increased hostile attributions independent of whether participant had been debriefed or not. To substantiate the ego threat effect and the (in)effectiveness of debriefings, we conducted internal meta-analyses across all five studies. Internal meta-analyses are recommended because they allow for more robust and cogent conclusions than individual studies (Braver, Thoemmes, & Rosenthal, 2014; Goh, Hall, & Rosenthal, 2016; Maner, 2014). In particular, due to increased statistical power, internal meta-analyses deliver more reliable effect size estimates and are better able to detect small effects that individual studies may have been unable to reliably detect. In turn, they allow for more confidence in null effects or negligible effects that are based on cumulative evidence.

Lab Session

Shared dependent variables across the included studies (Study 1, 2, 4, and 5) were mood, state self-esteem and self-perception. Shared experimental conditions were *no threat*

not debriefed yet, *threat not debriefed yet* and *threat (written) debriefed.* We therefore meta-analytically estimated the following three effect sizes: (1) the basic ego threat effect between the *threat not debriefed yet* condition versus the *no threat not debriefed yet* condition. (2) The effect of the *threat (written) debriefed* condition versus the *no threat not debriefed yet* condition.[5] (3) The *threat (written) debriefed* versus the *threat not debriefed yet* condition. The latter comparison constitutes a less conservative estimate of the effect of the debriefing because meta-analyses tend to detect smaller effects than individual studies. Hence, the likelihood of finding an effect of the debriefing is greater.

Follow-Up

All studies that included an online follow-up, i.e. all studies except for Study 1, were included in the meta-analysis of the follow-up effects. The shared dependent variable across the included studies was the follow-up questionnaire. Since all participants had been debriefed when they completed the questionnaire, only the effect of the ego threat manipulation (*threat* vs. *no threat*) could be meaningfully analyzed. We meta-analytically estimated the effect size of this comparison.

Method

We conducted random effects meta-analyses. The three effect sizes per study for the lab session were combined into one study effect size using the procedure described in Borenstein, Hedges, Higgins, & Rothstein (2009). This procedure weighs the effect size variances by the respective correlation between the outcome variables. Correlations were corrected for attenuation to estimate the correlations if the variables could have been measured without error (Spearman, 1904). In addition, we ran all analyses separately for each of the dependent variables. For the follow-up effect, we ran one analysis for the single

[5] A comparison of the *threat debriefed* with the *no threat debriefed* condition as reported for Studies 1 and 2 was not possible because the *no threat debriefed* condition was not realized in Studies 4 and 5.

dependent variable used in all included studies.

Results

Lab session. (1) For the combined study effect sizes of all included dependent variables, the basic ego threat effect was highly significant with a medium-to-large effect size, $z = 6.25$, $p < .001$, $d = 0.65$, 95% CI [0.45, 0.86]. A similar picture emerged in separate analyses for each of the three dependent variables, all zs > 3.38, all ps < .001, all ds > 0.50. Thus, across four studies, participants indicated lower well-being after having received negative compared to positive or no feedback.[6]

(2) The effect of the ego threat persevered after a debriefing: Well-being was lower for participants who had received negative feedback and a debriefing compared to participants who had received positive/no feedback, $z = 3.85$, $p < .001$, $d = 0.55$, 95% CI [0.27, 0.83]. The effect was also significant for each of the dependent variables analyzed separately, all zs > 2.08, all ps < .037, all ds > 0.34. Hence, the debriefing was ineffective in restoring participants' well-being compared to control conditions.

(3) A comparison between participants who had received negative feedback and had versus had not been debriefed revealed a small but significant effect of the debriefing, $z = 2.09$, $p = .036$, $d = 0.22$, 95% CI [0.01, 0.42]. None of the meta-analytic effects was significant when dependent variables were analyzed separately, all zs < 1.70, all ps > .089, all

[6] For full disclosure: We ran one additional study conducted as part of a student project ($N = 80$) using the indirect failure task used in Studies 1 and 5. For unknown reasons, we found no effect of the false feedback on the three dependent variables in this study. Therefore, our main research question about the effectiveness of debriefings could not be examined in this study due to the lack of manipulation success: There was no ego threat effect that a debriefing could have helped to eliminate. Hence we did not report this study in the current article. When including this study in the meta-analysis of the basic ego threat effect, the effect was still highly significant, $z = 5.43$, $p < .001$, $d = 0.58$. As Lakens & Etz (2017) point out, non-significant findings within a series of studies with significant findings are not only to be expected, but can even provide stronger evidence for a true effect. Apart from the student project mentioned here, there are no unreported studies in our file drawer.

*d*s < 0.25. Hence, the debriefing was mildly effective in improving participants' well-being compared to *threat* conditions without debriefing.

Follow-up. The ego threat effect was highly significant with a large effect size, $z = 4.61$, $p < .001$, $d = 1.00$, 95% CI [0.58, 1.42]. Thus, participants indicated adverse long term effects after *threat* compared to *no threat* control conditions several hours or even days after the end of the lab session.

General Discussion

Participants experienced two different kinds of ego threat: In Studies 1 and 5, participants received indirect failure feedback regarding their performance in a test presumably indicative of intellectual abilities. In Studies 2–4, participants received negative social feedback when two ostensible other participants conveyed that they found the participant neither intelligent, nor particularly likable, or physically attractive. In all five studies, receiving false negative compared to positive or no feedback led to lower mood and state self-esteem (Studies 1, 2, 4, & 5) and more hostile attributions (Study 3), respectively. Furthermore, negative compared to positive or no feedback led to less favorable self-perceptions (Studies 1, 2, & 4). Neither a written debriefing (Studies 1–5), nor a verbal debriefing (Studies 4 & 5), nor a written debriefing plus writing a self-affirming essay (Studies 4 & 5) were able to fully undo the detrimental effects caused by the ego threat manipulations. An internal meta-analysis confirmed these findings and revealed that receiving a debriefing only had a small positive effect on participants' well-being. Moreover, negative thoughts and feelings regarding the manipulations persevered the debriefing and persisted for one day and beyond (Studies 2–5). The fact that these findings occurred across two different ego threat manipulations suggests that these findings were not specific to one particular manipulation, but may generalize to similar other manipulations. As our studies were conducted at a German university, the findings might not be generalizable to a non-Western, non-academic population. However, the academic, undergraduate student sample of the

present research is highly representative for the populations used in previous research on ego threat. Furthermore, it is likely that the majority of future research on the topic of ego threat will rely on academic student samples as well. Indeed our population seems comparable to the ones used in previous research as demonstrated by the fact that we were able to consistently replicate previous findings on the effects of ego threat that were originally obtained in various labs in the US and Europe. Members of non-Western cultures might react to ego threat in a different way because of different social conventions but we would expect that receiving believable negative feedback about core aspects of one's self (such as intelligence or likeability) is universally threatening for people, regardless of their cultural background.

The goal of the present research was to examine the effectiveness of post-experimental debriefings in restoring participants' well-being in the short term. The debriefing procedures used here clearly fell short of this goal – despite deliberate efforts to maximize their effectiveness: as a starting point, we employed the Revised Outcome Debriefing, the most effective debriefing method in prior research (McFarland et al., 2007). To further improve the effectiveness, we extended this procedure in two ways: In one condition, participants received the debriefing verbally from a carefully trained experimenter. In another condition, participants received a written debriefing and additionally wrote a self-affirmation essay – an intervention that has repeatedly been shown to positively affect various aspects of well-being in prior research (Klein, Blier, & Janze, 2001; Koole, Smeets, van Knippenberg, & Dijksterhuis, 1999).

To ensure that participants understood the respective debriefing, they explicitly confirmed their full comprehension of the debriefing's content in the written as well as the verbal condition. Indeed, the small meta-analytic effect of the debriefing corroborates that participants understood the debriefing. However, the main result of the present studies is that the various debriefing procedures were largely unsuccessful in restoring participants' well-being. They thereby failed to meet the APA's guideline to avoid psychological harm caused

by participating in research studies. These results thus have important implications for psychological researchers given the field's omnipresent reliance on the (alleged) effectiveness of debriefings to undo adverse effects caused by ego threat and other manipulations.

What are these substantive implications of the present findings? On the positive side, in light of the ongoing debate about the replicability of psychological research our results suggest that ego threat reliably impairs well-being in the short-term with healthy effect sizes.

Concerning the ineffectiveness of the various debriefing procedures to undo these effects, we want to stress that we do not mean to insinuate that researchers should straightaway stop using ego threat manipulations in their work. Important research questions are dependent on the ability to temporarily threaten participants' ego. Neither do we mean to accuse researchers of having neglected their duty to seriously care for their participants' well-being. The starting point for the present research simply was the pronounced imbalance between the field's enormous trust in debriefings' effectiveness to restore participants' well-being on the one hand and the striking scarcity of empirical evidence for this assumption on the other hand. The present research contributes considerable evidence that our trust in debriefings may have been premature and too credulous. We believe that the present findings mandate to engage in reflection and discussion about the way that we, as a field, want to do research. We hope that one contribution of the present research is to stimulate such discussion and raise important questions that may not have easy answers, but belong on our field's agenda:

(1) Which aftereffects of participating in psychological research do we, as a discipline, regard as acceptable versus problematic? Should we regard a reduced well-being when leaving the laboratory or an emotional preoccupation during the following day as minimal risk or rather as troublesome? Yes, the debriefings failed to re-establish pre-study conditions and to fully undo psychological harm caused by the experiences during the study. Then again, what constitutes harm is not clearly defined. Should any level of discomfort – e.g., lowered

mood, impaired self-esteem – be considered a form of harm? Or do we set a higher threshold and demand that harm only begins with clinically relevant symptoms? Are some negative effects acceptable to inflict on participants who freely consented to participating in a research study?

(2) What are effective remedies for psychological harm caused by research procedures? In the present research, neither a (written or verbal) Revised Outcome Debriefing, nor a debriefing plus self-affirmation was successful in undoing the adverse effects caused by ego threat. Future work should examine the effectiveness of further procedures. For example, while the self-affirming task used in the present research did not prove to be effective in restoring participants' well-being, other self-affirming tasks might be. In previous research, value affirmations as opposed to attribute affirmations have often only been shown to affect either mood *or* self-esteem (cf. Galinsky, Stone, & Cooper, 2000; Schmeichel & Martens, 2005; Sherman & Kim, 2005; Shrira & Martin, 2005). Since we examined the effects of ego threat on mood *and* self-esteem, we decided against using a value affirmation task. However, given the surprising ineffectiveness of the self-affirmations task used in the present research, future work might reconsider investigating the merit of a value affirmation task as a post-experimental intervention.

Other possible interventions worth investigating include self-compassion and apologies. In previous research, self-compassion has been shown to help people cope with unpleasant self-relevant events (Leary, Tate, Adams, Batts Allen, & Hancock, 2007). A less obvious approach to restore participants' well-being could be the use of apologies as post-experimental interventions: while receiving an apology has usually been investigated as a means to encourage forgiveness (for a meta-analysis see Fehr, Gelfand, & Nag, 2010), it might be reasonable to assume that forgiveness is accompanied by a psychological process that ultimately leads to higher well-being. Therefore, it could be sensible for future work to look into the potential benefits of a heartfelt apology given to participants by the

experimenter.

(3) How long do negative aftereffects caused by an experimental manipulation last, and what can be considered an ethically acceptable time frame? In our studies, participants still reported having negative thoughts and feelings about their research experience one to two days after leaving the lab. We assume that these effects will disappear over time, but we currently have no evidence to suggest how long this may take.

(4) What are the real-life consequences of negative aftereffects caused by research procedures? While it is possible that participants did not experience any appreciable real life consequences, it is also possible that their experiences impacted on their lives as students, friends and partners. For example, how may false feedback about intellectual performance have impacted participants' anxiety, their performance in an exam, or their attention during in a lecture?

(5) Should the failure of debriefing procedures only be viewed as an ethical problem or also as a methodological one regarding the use of limited participant pools? We do not know when former participants of ego threat studies are suitable again for participation in further psychological studies: if they go back to the same lab or a lab in the same building only a few days after participating in an ego threat study, are they still in distress? Do they anticipate that they will have a negative experience again?

Conclusion

Ego threat manipulations caused lowered well-being. Debriefing procedures failed to undo these effects and let participants leave the laboratory with unwanted aftereffects. Negative study-related affect and mental preoccupation with the study persisted for one day and longer. These findings raise questions about research ethics.

Appendix

Supplemental Material

Rating Sheet (used in Studies 2–4)

1. I think this person's intelligence is… (1 = extremely below average, 7 = extremely above average)
2. I find this person… (1 = very unlikeable, 7 = very likeable)
3. I find this person… (1 = very unattractive, 7 = very attractive)
4. I would… (1 = dislike working on a task with this person very much, 7 = like working on a task with this person very much)
5. I would… (1 = dislike meeting this person socially very much, 7 = like meeting this person socially very much)

Ratings Received by the Participants (as used in Studies 2–4)

Negative feedback condition:

intelligence: 2 of 7 (confederate 1) / 3 of 7 (confederate 2)

likeability: 1 of 7/ 2 of 7

attractiveness: 3 of 7 / 2 of 7

working on a task together: 2 of 7 / 1 of 7

meeting each other socially: 1 of 7 / 2 of 7

Positive feedback condition:

intelligence: 6 of 7 (confederate 1) / 6 of 7 (confederate 2)

likeability: 7 of 7 / 6 of 7

attractiveness: 6 of 7 / 7 of 7

working on a task together: 7 of 7 / 5 of 7

meeting each other socially: 6 of 7 / 7 of 7

Items From the Follow-up Questionnaire (as used in Study 2)

1. How much were the things that happened during the experiment in your mind?
2. How often did you think about the feedback you received?
3. How proud did you feel when you thought about the feedback you received?
4. How sad did you feel when you thought about the feedback you received?
5. Since the experiment, how often did you think about the impression you make on other people?
6. How anxious did you feel when you thought about the feedback you received?
7. How happy did you feel when you thought about the feedback you received?
8. Since the experiment, how often did you think about whether others find you (un)attractive?
9. How angry did you feel when you thought about the feedback you received?
10. How often did thoughts about the feedback you received pop into your mind?

Items From the Remote Associates Test (as used in Studies 1 & 5)[7]:

Difficult Items:

1. Zeichen / Wort / Bogen (solution: Frage)
2. Training / Elefanten / Stoerung (solution: Gedächtnis)
3. Stellung / Bau / Glied (solution: Satz)
4. Zeile / Schmuck / Dick (solution: Kopf)

[7] In this task, participants are given three German words and have to find a fourth one that can be combined with each of the given ones to build three meaningful compound nouns. As an example, consider difficult item 8: Fisch (fish) / Rausch (rush) / Kette (chain). The solution is Gold (gold), yielding the compound nouns Goldfisch (goldfish), Goldrausch (gold rush), and Goldkette (gold chain). Due to the structural differences in the two languages, most examples do not yield meaningful English compound nouns, so we only give the translations of the individual words and the compounds if the solution is also meaningful in English.

5. Ruf / Hof / Schiff (solution: Schlacht)
6. Ball / Gas / Gurken (solution: Maske)

   (English: ball / gas / cucumber - solution: mask)
7. Krank / Boden / Igel (solution: See)
8. Fisch / Rausch / Kette (solution: Gold)

   (English: fish / rush / chain - solution: gold)
9. Bund / Bein / General (solution: Schluessel)
10. Kiste / Auto / Laden (solution: Spielzeug)

    (English: box / car / shop - solution: toy)
11. Netz / Zeug / Feuer (solution: Werk)
12. Stück / Beere / Blätter (solution: Wald)

Easy Items:

1. Haus / Tannen / Stamm (solution: Baum)

   (English: house / fir / log - solution: tree)
2. Wasser / Mine / Streuer (solution: Salz)

   (English: water / mine / shaker - solution: salt)
3. Waben / Bienen / Bio (solution: Honig)

   (English: comb / bee / organic - solution: honey)
4. Wasser / Lok / Kessel (solution: Dampf)
5. Abfluss / Metall / Schilf (solution: Rohr)
6. Haus / Liege / Sand (solution: Strand)

   (English: house / chair / sand - solution: beach)
7. Peking / Grütze / Zeitungs (solution: Ente)
8. Onkel / Riesen / Teil (solution: Groß)
9. Lack / Finger / Pilz (solution: Nagel)

   (English: polish / finger /fungus - solution: nail)

10. Maß / Wurm / Arm (solution: Band)
11. Geld / Cocktail / Karaoke (solution: Bar)
12. Schweizer / Kuchen / Hütten (solution: Käse)

    (English: swiss / cake / cottage - solution: cheese)

Items From the Follow-up Questionnaire (as used in Studies 3 & 4)

1. How much would you have liked to participate in the experiment if you had known beforehand what you would experience?
2. How did you like the study?
3. How much would you like to participate in a study about online dating again?
4. How much would you like to participate in this kind of study or a similar one for a second time?
5. How competent did the experimenter appear to you?
6. How likeable did the experimenter appear to you?
7. I feel that the study has negatively affected my well-being.
8. Thinking of the study somehow leaves me with a certain feeling of uneasiness.
9. I felt happy when thinking about the feedback that I was given in response to my video.
10. I felt insecure when thinking about the feedback that I was given in response to my video.
11. I felt proud when thinking about the feedback that I was given in response to my video.
12. I felt sad when thinking about the feedback that I was given in response to my video.

Items From the Follow-up Questionnaire (as used in Study 5)

1. How much would you have liked to participate in the experiment if you had known

beforehand what you would experience?

2. How did you like the study?
3. How much would you like to participate in this study or a similar one again?
4. How much would you like to participate in this kind of study or a similar one for a second time?
5. How competent did the experimenter appear to you?
6. How likeable did the experimenter appear to you?
7. I feel that the study has negatively affected my well-being.
8. Thinking of the study somehow leaves me with a certain feeling of uneasiness.
9. I felt happy when thinking about my performance in the test.
10. I felt insecure when thinking about my performance in the test.
11. I felt proud when thinking about my performance in the test.
12. I felt sad when thinking about my performance in the test.

Example Item From the SIP-AEQ (as used in Study 3) With Attribution That is Reflected by Each Statement

Imagine that you are in a karate class competition and you have to demonstrate your abilities to your instructor. You are matched up to "fight" with someone in class who you do not know well. While you are being evaluated, your karate classmate hits you in a way other than the way you were taught and you are hurt.

Why do you think your classmate hit you in a way other than the way you were taught? Rate the likelihood of each statement on a scale of 0 - 3:

A1. My Karate classmate wanted to physically hurt me. (direct hostile attribution)

A2. My Karate classmate wanted to win the match. (instrumental attribution)

A3. My Karate classmate did it by accident. (benign attribution)

A4. My Karate classmate wanted me to look "bad". (indirect hostile attribution)

Instruction for the Self-affirmation Essay and the Essay From the Control Conditions (as used in Studies 4 & 5)

Self-affirmation Essay

Please remember one or more occasions in which you helped another person and felt good about it. This includes bigger things, like helping a friend, caring for a relative, helping someone move houses, or doing a year of voluntary work but it also refers to smaller things such as helping someone carry a suitcase, giving directions, offering your seat in public transportation, or doing household chores. Because even the smaller things you do for others do matter.

Try to remember as many details as possible, especially about the way helping others made you feel. Describe one or more such events. It is important that you reflect on your feelings.

Control Essay

Please describe the shoes that you are currently wearing as accurately as possible.

Try to be very specific and detail-oriented such that a person that has not seen your shoes has a clear and exact picture of them in mind. This includes describing the shape, the estimated dimensions, the color or pattern, the material, the seams, possible signs of wear, etc.

PART II

When Loneliness Sticks With You

Can a Debriefing Undo the Effects of Ostracism?

Stefanie Miketta & Malte Friese

Saarland University

Manuscript in preparation for submission.

Abstract

Experimentally induced ostracism is a frequently used powerful manipulation that can lead to various undesirable effects including negative mood, low self-esteem, and aggressive behavior. While a temporary reduction of participants' well-being is often sought for research purposes, ethical guidelines require researchers to protect their participants from any psychological or physical harm. Researchers therefore rely on post-experimental debriefings to rectify any prior given false information and restore participants' well-being. However, research suggests that misinformation can persevere even after it has been explicitly corrected. The present research investigates whether a debriefing after experimentally induced ostracism restores participants' subjective well-being. In two studies, participants played the Cyberball game, completed dependent variables indicative of subjective well-being and were debriefed in writing. In two additional conditions, participants were debriefed *before* completing the dependent variables. After having been excluded compared to included in the Cyberball game, participants reported decreased positive and increased negative mood. The debriefing did not change this pattern. The ostracism manipulation affected neither explicit nor implicit self-esteem. A follow-up assessment revealed that unwanted aftereffects of the manipulation persevered for several hours after the experimental session. The present findings question the effectiveness of written debriefings in restoring participants' well-being.

Keywords: Debriefing, Ostracism, Perseverance

Introduction

Being included in social groups is a fundamental need of human beings (Baumeister & Leary, 1995; E. R. Smith, Murphy, & Coats, 1999). It has even been argued that belonging to social groups is crucial for survival (Caporael, 1997; van Beest & Williams, 2006). It should thus not come as a surprise that ostracism – being ignored and excluded by others – can have quite worrisome effects, even if the ostracism experience is only part of an experimental session: after being ostracized in the laboratory, people show more aggressive behavior (Warburton, Williams, & Cairns, 2006; Wesselmann et al., 2010), more negative mood (Gonsalkorale & Williams, 2007; Williams et al., 2000), activation of the dorsal anterior cingulate cortex (indicating that the brain responds to ostracism similar as to physical pain; Eisenberger, Lieberman, & Williams, 2003), lower self-esteem, sense of control, and belonging, as well as increased feelings of meaninglessness (Gonsalkorale & Williams, 2007; Jamieson et al., 2010; Williams et al., 2000) as compared to people who have not been ostracized (for a meta-analysis, see Hartgerink et al., 2015).

The most common procedure to experimentally manipulate ostracism in the laboratory is the so called Cyberball game (Gonsalkorale & Williams, 2007; Williams et al., 2000). In this game, participants play a virtual ball tossing game with two ostensible other participants over the Internet. Unknown to participants, instead of actual others playing with them, the computer runs a pre-programmed script. This script ensures, in the inclusion condition, that participants receive the ball approximately a fair one third of the time while excluded participants receive the ball once or twice at the beginning of the game and then never receive it again. This procedure has been employed in more than 200 published studies (Hartgerink et al., 2015).

The Cyberball game has been developed as a virtual analogue to a real-life ball tossing paradigm and was intended to be less traumatic and debilitating, because the persons engaging in the ostracizing behavior are not physically present (Williams, 2007; Williams et

al., 2000). Taking this into account, the plethora of documented negative consequences of this manipulation outlined above is even more impressive. At the same time, the adverse effects of this manipulation give reason for concern: if such minimal ostracism can cause aggressive behavior and lower participants' mood as well as their self-esteem, it is all the more important to remedy these effects by the time they leave the laboratory.

In their *Ethical Principles of Psychologists and Code of Conduct*, the American Psychological Association (APA; 2002) deals with the question of inflicting harm on participants and sets a clear standard on the topic: "Psychologists take reasonable steps to avoid harming their … research participants" (p. 1065). Starting decades ago, researches have thought of three different strategies on how to take those 'reasonable steps': (1) the abandonment of any harmful manipulations in psychological studies (e.g., Baumrind, 1964), (2) the employment of several precautions such as recruiting only psychologically stable individuals as participants (e.g., Kelman, 1967; Walster et al., 1967), (3) the use of a post-experimental debriefing at the end of an experimental session to counteract negative effects (e.g., Blanck et al., 1992; D. S. Holmes, 1976a, 1976b; S. S. Smith & Richardson, 1983).

In many areas of psychological research including research on ostracism, the first and second approach have not been widely adopted: research on ostracism is thriving and careful a priori screening of participants is rarely conducted (or at least this is not mentioned in the respective publications). The third approach – use of post-experimental debriefings – acknowledges that harm will be done in the short-term, but trusts that a debriefing will be successful in eliminating this harm before participants leave the laboratory. In a debriefing, participants are usually being informed about a potential prior deception. This approach gained widespread acceptance since the early 1970s. The APA (2002) even dedicated a whole section in their *Ethical Principles of Psychologists and Code of Conduct* to the topic of debriefings. In this section, the APA clearly specifies one important intention of a debriefing: "when psychologists become aware that research procedures have harmed a participant, they

take reasonable steps to minimize the harm“ (p. 1070).

The previous section makes clear that the debriefing assumes an important role in fulfilling the aim of minimizing research participants' harm. However, a few researchers criticized the lack of formal evidence for the effectiveness of debriefings and have voiced according doubts and concerns (e.g., Sharpe & Faye, 2009; Tesch, 1977). Additionally, a different line of research showed that beliefs can still persevere when the underlying information has been declared as false (e.g., Anderson et al., 1980; Greitemeyer, 2014; Guenther & Alicke, 2008; Nyhan & Reifler, 2010; Wyer, 2010). In those studies, originally given information about persons, scientific results, or theories about causal relationships persevered after it had been discredited and corrected. Although these studies were not concerned with the effectiveness of debriefings, both contexts bear evident resemblance: they both seek to correct originally given information (e.g., prior deception). Thus, this research raises the possibility that false information given in the context of an experimental manipulation could also persevere after it has been corrected by a post-experimental debriefing, potentially impeding its effectiveness.

Considering the crucial role of debriefings and the doubts about their effectiveness based on research on the perseverance effect, there has been surprisingly little research on the effectiveness of post-experimental debriefings. The few existing studies on the topic come to the conclusion that debriefings are not necessarily effective (McFarland et al., 2007; Ross et al., 1975; Silverman et al., 1970; Walster et al., 1967). As a result, several researchers made efforts to develop better working debriefing procedures. For example, McFarland and colleagues (2007) compared a Standard Outcome Debriefing, which informed participants about the false feedback they had received, with their newly created Revised Outcome Debriefing that additionally informed participants about the fact that the ostensible social perceptiveness test that they took was not real either. This Revised Outcome Debriefing proved to be more effective in eliminating the effects of the manipulation than the Standard

Outcome Debriefing.

The research on the effectiveness of debriefings by McFarland and colleagues suggests, that the Revised Outcome Debriefing is the most potent type of debriefing and able to undo the effects of experimental manipulations. Doubts remain, however, if this conclusion extends to threatening experimental manipulations such as experimental ostracism. In the work by McFarland et al. (2007), participants received false feedback about an ostensible social perceptiveness task in which they had to distinguish fake suicide notes from real ones. In the negative feedback condition, participants then received the feedback that they had performed poorly on the task. Experimental manipulations arguably differ in how emotionally involving, intrusive, harmful, and potentially traumatic they are. Considering the fact that ostracism threatens the fundamental human need to belong and can lead to broad detrimental effects on well-being and psychological adjustment, the manipulation used by McFarland and colleagues seems relatively mild in comparison. After all, most people are unlikely to regard reliably distinguishing real from fake suicide notes a core competence.

The question whether a debriefing procedure is effective in eliminating the effects of potentially harmful manipulations remains unanswered. Some research even suggests that the Cyberball manipulation may have similarly harmful effects when participants are told in advance that they will be interacting with a pre-programmed computer script instead of real persons (Zadro, Williams, & Richardson, 2004). Since a post-experimental debriefing usually provides participants with information about the false nature of the manipulation, the information that participants received in the study by Zadro and colleagues can be thought of as a pre-experimental debriefing. This "debriefing" had no softening effect on the harm caused by the manipulation, rendering the effectiveness of the post-experimental debriefing all the more important.

The importance of an effective debriefing after a Cyberball manipulation is underlined by yet another observation: A recent meta-analysis showed that ostracism effects were not

moderated by the time that had passed since playing the Cyberball game (Hartgerink et al., 2015). This finding suggests that researchers cannot rely on time alone to eliminate the detrimental effects caused by an ostracism manipulation.

The Present Research

The aim of the present research was to investigate whether a post-experimental debriefing would be effective in eliminating the effects on well-being caused by a Cyberball ostracism manipulation. In two studies, we sought to conceptually replicate the effects of the Cyberball manipulation on indicators of subjective well-being such as mood and self-esteem. We then investigated whether a Revised Outcome Debriefing would eliminate those effects. In Study 2, we additionally investigated potential aftereffects of the ostracism manipulation several hours after participants had left the laboratory. Both studies were approved by the local ethics committee.

Study 1

In Study 1, participants played Cyberball and were either included or excluded by two ostensible other players. We expected participants in the exclusion condition to show signs of impaired subjective well-being as indicated by more negative mood, less positive mood, and lowered state self-esteem. Most importantly, we investigated whether the Revised Outcome Debriefing was able to remedy the effects of the ostracism manipulation.

Method

Participants and design. Eighty-five students (67% females; mean age $M = 22.91$, $SD = 2.96$) of various disciplines from a German university participated in exchange for sweets and a voucher for a hot beverage at a café on campus. They were randomly assigned to one of four conditions in a 2 (ostracism condition: inclusion vs. exclusion) × 2 (debriefing status [at the time of DV measurement]: debriefed vs. not debriefed yet) design. Participants were run in individual sessions that lasted about 20 minutes. No participants were excluded from the study.

Procedure. After giving informed consent, participants were asked to play a virtual ball tossing game on the computer. Afterwards, participants in the *not debriefed yet* conditions completed the dependent variables (positive mood, negative mood, state self-esteem) as well as demographics (age, gender, major). They then received the debriefing. Participants in the *debriefed* conditions received the debriefing before completing the dependent measures. After the debriefing and completion of the dependent measures, participants were thanked, compensated, and dismissed.

Experimental manipulations.

Ostracism. To manipulate ostracism, we used the Cyberball game (Gonsalkorale & Williams, 2007; Jamieson et al., 2010; Williams et al., 2000). Participants played a virtual ball tossing game on the computer with two ostensible other participants from two other German universities. The names of these universities were made up. Participants were told that the ball tossing game was a mental visualization exercise and that they should try to visualize their throws and catches. Participants were further told that during the game, they as well as the two ostensible other players would each be represented by an animated figure, their first name (for the ostensible other players we used one male and one female name) and a specific color on the screen. Participants were instructed to choose freely which player they wished to throw the ball to whenever their animated figure would catch the ball.

The game consisted of 70 trials and lasted for approximately four minutes. In the *exclusion* condition, participants received the ball once from each of the other two ostensible players at the beginning of the game. Then they never received the ball again. In the *inclusion* condition, participants received the ball after every other throw from each of the other two ostensible players which adds up to receiving the ball a third of the time in total.

Debriefing status. In the *not debriefed yet* condition, participants completed the dependent variables before they received the debriefing. In the *debriefed* condition, participants were first debriefed and then completed the dependent variables.

Debriefing. Participants received a Revised Outcome Debriefing that has been shown to be the most effective kind of debriefing in previous research (McFarland et al., 2007). The debriefing was presented on the computer screen. Participants read that in contrast to what they had learned earlier (1) there were no other participants, (2) the two universities did not exist, (3) the ball tossing game consisted of the computer running a pre-programmed script and that (4) they had been randomly assigned to a condition where they either received the ball only twice at the beginning of the game and then never received it again (*exclusion* condition) or where they received the ball regularly (*inclusion* condition). To ensure that participants understood the debriefing, they were asked to explicitly confirm the following sentence in order to continue with the study: "I fully understand that the 'other participants' were the computer running a script and that I have not actually been excluded / playing with someone". All participants confirmed this statement.

Dependent variables. In past research, the Cyberball manipulation reliably affected mood and self-esteem (Hartgerink et al., 2015). Thus, a measure of mood and a measure of self-esteem served as dependent variables.[8] Since in several factor analyses, positive and negative mood have been shown to be two considerably different dimensions (e.g., Crawford & Henry, 2004; Watson, Clark, & Tellegen, 1988), we measured both dimensions.

Positive and negative mood. Participants indicated their current mood on the German version (Krohne, Egloff, Kohlmann, & Tausch, 1996) of the Positive And Negative Mood Scale (PANAS; Watson et al., 1988). Internal consistency was α = .82 for positive and α = .81

[8] Half of the participants received a debriefing rectifying the initially given false information about the ostracism manipulation before completing the dependent variables. This meant that the so-called *need threat* measure – a dependent variable that is often used in Cyberball studies (e.g., Williams et al., 2000) – could not be used in the present research: the questions included in the need threat measure, e.g., "To what extent do you think the other participants value you as a person?" are no longer meaningful questions for participants who have already been informed that there were no "other participants". Therefore, we refrained from using a measure of need threat.

for negative mood.

State self-esteem. To measure state self-esteem, we used the German version (Rudolph et al., 2009) of the State Self-Esteem Scale (SSES; Heatherton & Polivy, 1991; 5-point rating scales, 1 = *strongly disagree*, 5 = *strongly agree*). The SSES consists of 3 subscales (performance, appearance and social self-esteem, five items each). Internal consistency across all three subscales was $\alpha = .85$[9].

Results

Analytic strategy. For each dependent variable, we ran a 2 (ostracism condition: inclusion vs. exclusion) × 2 (debriefing status: debriefed vs. not debriefed yet) ANOVA. In all cases, we started by examining the predicted simple main effect of the *ostracism condition* in the *not debriefed yet* condition (mirroring previous research) to establish the focal effect that the debriefing procedure would aim to remedy. In the case of evidence for this simple main effect, we then analyzed the simple main effect of the *ostracism condition* in the *debriefed* conditions. Next, we investigated whether the general main effect of the *ostracism condition* was significant across both *debriefing status* conditions. Then we examined a potential interaction between *ostracism condition* and *debriefing status*. Finally, we looked for a potential (unexpected) main effect of *debriefing status*.

If there was no evidence for the simple main effect of the *ostracism condition* in the *not debriefed yet* conditions, this meant that ostracism did not significantly influence the dependent variable. In this case, our main research question concerning the effectiveness of the debriefing could not be examined: it is impossible to investigate whether a non-existent effect of the ostracism manipulation persevered after the debriefing. Therefore, we do not report the full set of analyses in these cases and instead move on to the next dependent variable.

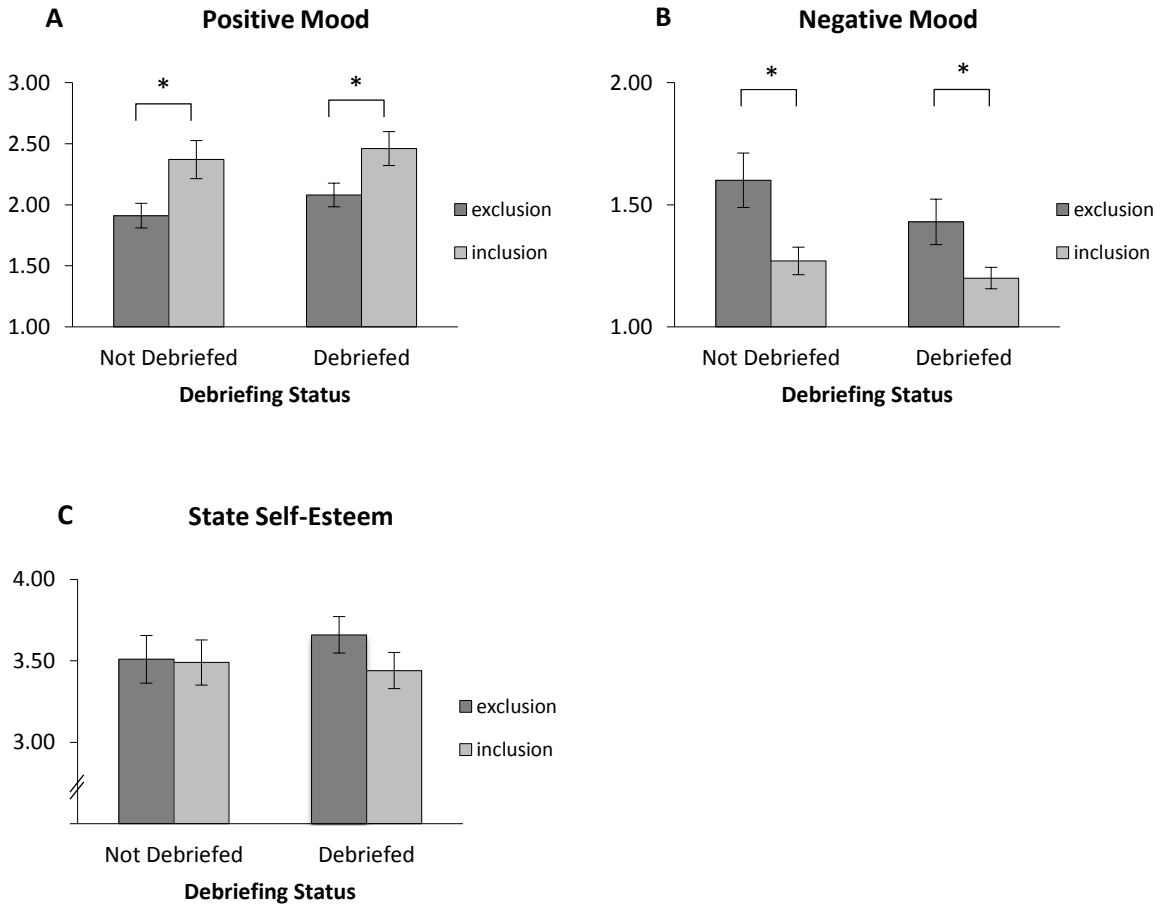
[9] A separate computation of the three subscales of State Self-Esteem led to the same results as the use of the overall score.

Positive mood. In the *not debriefed yet* conditions (mirroring earlier research), participants reported less positive mood after *exclusion* (M = 1.91, SD = .45) than after *inclusion* (M = 2.37, SD = .73), $t(35.29) = -2.46$, $p = .019$, $d = -0.75$. This establishes that the Cyberball manipulation successfully manipulated positive mood, as expected. In the *debriefed conditions*, the simple contrast of the *ostracism condition* revealed that the effect of the Cyberball manipulation on positive mood persevered after the debriefing (M_{Excl} = 2.08, SD = .47; M_{Incl} = 2.46, SD = .62), $t(35.07) = -2.25$, $p = .031$, $d = -0.70$. The main effect of the *ostracism condition* was significant across both *debriefing status* conditions (M_{Excl} = 2.00, SD = .46; M_{Incl} = 2.41, SD = .68), $F(1, 81) = 11.04$, $p = .001$, $\eta_p^2 = .12$. There was no interaction between the *ostracism condition* and *debriefing status* $F(1, 81) = 0.09$, $p = .766$, $\eta_p^2 < .01$. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 2.15, SD = .65; M_{Deb} = 2.25, SD = .58), $F(1, 81) = 1.05$, $p = .308$, $\eta_p^2 = .01$ (Figure 6a). Taken together, exclusion led to decreased positive mood, and the debriefing that was intended to remedy this effect was unsuccessful in doing so.

Negative mood. In the *not debriefed yet* conditions, participants reported more negative mood after *exclusion* (M = 1.60, SD = .49) than after *inclusion* (M = 1.27, SD = .26), $t(28.31) = 2.71$, $p = .011$, $d = 0.85$. This establishes that the Cyberball manipulation successfully manipulated negative mood, as expected. In the *debriefed conditions*, the simple contrast of the *ostracism condition* revealed that the effect of the Cyberball manipulation on negative mood persevered after the debriefing (M_{Excl} = 1.43, SD = .45; M_{Incl} = 1.20, SD = .19), $t(30.99) = 2.25$, $p = .032$, $d = 0.65$. The main effect of the *ostracism condition* was significant across both *debriefing status* conditions (M_{Excl} = 1.51, SD = .47; M_{Incl} = 1.24, SD = .23), $F(1, 81) = 12.29$, $p = .001$, $\eta_p^2 = .13$. There was no interaction between the *ostracism condition* and *debriefing status* $F(1, 81) = 0.43$, $p = .516$, $\eta_p^2 < .01$. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 1.43, SD = .42; M_{Deb} = 1.32, SD = .37), $F(1, 81) = 2.24$, $p = .138$,

η_p^2 = .03 (Figure 6b). Taken together, exclusion led to increased negative mood, and the debriefing that was intended to remedy this effect was unsuccessful in doing so.

State self-esteem. Contrary to expectations, in the *not debriefed yet* conditions, participants showed similar levels of state self-esteem after *exclusion* (M = 3.51, SD = 0.67) as after *inclusion* (M = 3.49, SD = 0.63), $t(81) = 0.14$, $p = .889$, $d = 0.03$ (Figure 6c). Thus, the Cyberball manipulation had no significant effect on state self-esteem. Therefore, we stopped the analysis of this dependent variable at this point.



Figures 6a–6c. Positive and negative mood, as well as state self-esteem as a function of *ostracism condition* and *debriefing status* at the time of DV measurement. Participants who had been excluded showed lower levels of positive mood and higher levels of negative mood than participants who had been included, no matter whether they had been debriefed at the time of the collection of the respective dependent variable or not. There were no significant effects on state self-esteem. Error bars indicate +/- 1 SEM. *$p < .05$.

Discussion

Participants who had been excluded in the Cyberball game showed decreased positive mood and increased negative mood compared to participants who had been included in the game. A Revised Outcome Debriefing failed to eliminate these effects. Debriefed participants who had learned that they in fact had not been excluded because there were no other participants were similarly affected by the Cyberball manipulation as participants who had not been debriefed yet. The debriefing thus failed to undo the psychological harm caused by ostracism.

Contrary to previous research (e.g., Gonsalkorale & Williams, 2007; Jamieson et al., 2010; Williams et al., 2000), the Cyberball manipulation did not reliably affect state self-esteem. However, this finding possibly aligns with the fact that negative effects of social exclusion on self-esteem can be delayed, dependent on rumination, or subject to defensive reactions that protect the self (Blackhart et al., 2009; Kunda, 1990; Tesser, 2000).

Study 2

In Study 2 we sought to replicate the effects of ostracism on mood obtained in Study 1. There was one important modification and one extension compared to Study 1: First, the Cyberball manipulation did not affect participants' state self-esteem in Study 1, despite previous evidence to the contrary (e.g., Gonsalkorale & Williams, 2007; Jamieson et al., 2010; Williams et al., 2000). We therefore used a different, potentially better suited measure of self-esteem in Study 2: an implicit self-esteem task directly tailored to the ostracism manipulation and possibly less susceptible to intentional response editing. Second, Study 2 examined potential aftereffects of the ostracism experience not only after the debriefing, but several hours after participants had left the laboratory.

Method

Participants and design. Ninety-two university students of various disciplines (40% females; mean age $M = 22.52$, $SD = 2.38$) participated in exchange for sweets and a voucher

for a hot beverage at a café on campus. They were randomly assigned to one of four conditions in a 2 (ostracism condition: inclusion vs. exclusion) × 2 (debriefing status [at the time of the DV measurement]: debriefed vs. not debriefed yet) design. Participants were run in individual sessions that lasted about 20 minutes. No participants were excluded from the study.

Procedure. The procedure during the laboratory session was the same as in Study 1. In addition, at the end of the lab session, participants were asked to fill out an online follow-up questionnaire later on the same day. Then they were thanked, compensated, and dismissed. Later on the same day, participants received a personalized link to the online questionnaire via email. The link contained coded information about the participant number and experimental condition, both of which were automatically saved to the data file. This allowed us to match the data from the lab session and the follow-up for each participant. We did not retrieve any information about participants' identity. Therefore, anonymity was guaranteed for all participants.

Experimental manipulations. Ostracism and debriefing status were manipulated the same way as in Study 1.

Debriefing. The debriefing was the same as in Study 1.

Dependent variables.

Mood. Participants indicated their positive and negative mood on 12 items taken from the PANAS-X (Watson & Clark, 1999; 5-point rating scales, 1 = *very slightly or not at all*, 5 = *extremely*). Items for positive mood were: relaxed, at ease, proud, joyful, happy, cheerful. Items for negative mood were: ashamed, irritable, upset, downhearted, angry, sad. Internal consistency was α = .86 for positive and α = .81 for negative mood.

Implicit self-esteem. To measure implicit self-esteem we used a self-esteem-IAT (Greenwald & Farnham, 2000). Participants sorted stimuli that were presented on the computer screen into four categories by pressing one of two response keys. Response

latencies were collected for each response. Participants were instructed to react as fast as possible while committing as few errors as possible. Category labels were "positive", "negative", "self", and "other". Evaluative stimuli were five positive and five negative words addressing qualities potentially affected by the ostracism manipulation (positive: popular, appreciated, liked, popular/liked[10], loveable; negative: lonely, rejected, excluded, unwanted, spurned). The "self" and "other" categories consisted of five generic stimuli each (self: me, self, own, mine, my; other: not-me, you, theirs, they, yours). The IAT consisted of 7 blocks. Blocks 1, 2, and 5 were practice blocks. Blocks 3, 4, 6, and 7 were critical blocks. In blocks 1 and 5, participants sorted stimuli into the "self" and "other" categories. In block 2, participants sorted evaluative stimuli into their respective categories. In blocks 3 and 4, the "self"/"other" categories were each combined with an evaluative category and shared a response key with it. In blocks 6 and 7 the "self"/"other" categories were each combined with the evaluative category that they were not combined with in blocks 3 and 4. It was counterbalanced whether the combination self+positive or the combination self+negative was encountered first. Blocks 1, 2, 3, and 6 contained 20 trials each, blocks 4, 5 and 7 contained 40 trails each. The inter-trial interval was 250 ms. The mean error rate was 6.91%. The D_1 score (Greenwald et al., 2003) was used to determine the IAT effect for implicit self-esteem. A positive score reflects a stronger association between oneself and positive relative to negative concepts, thus indicating a higher implicit self-esteem.

Aftereffects. Participants answered 10 items referring to thoughts and feelings about and consequences of their experience playing Cyberball. Sample items include "Since the experiment, how often did you think about whether others find you (un)likeable?" and "How often did thoughts about the course of the game pop into your mind?" (5-point rating scale, 1 = *not at all*, 5 = *very much/often*; α = .79; see Appendix for a list of all items).

[10] We used the German word "beliebt" which is a synonym to the German words for "popular" and "liked".

Results

Analytic strategy. The analytic strategy was the same as in Study 1, except for the follow-up questionnaire: at the time of the measurement of potential aftereffects, all participants had already been debriefed. We therefore collapsed across the *debriefing status* conditions for this dependent variable.

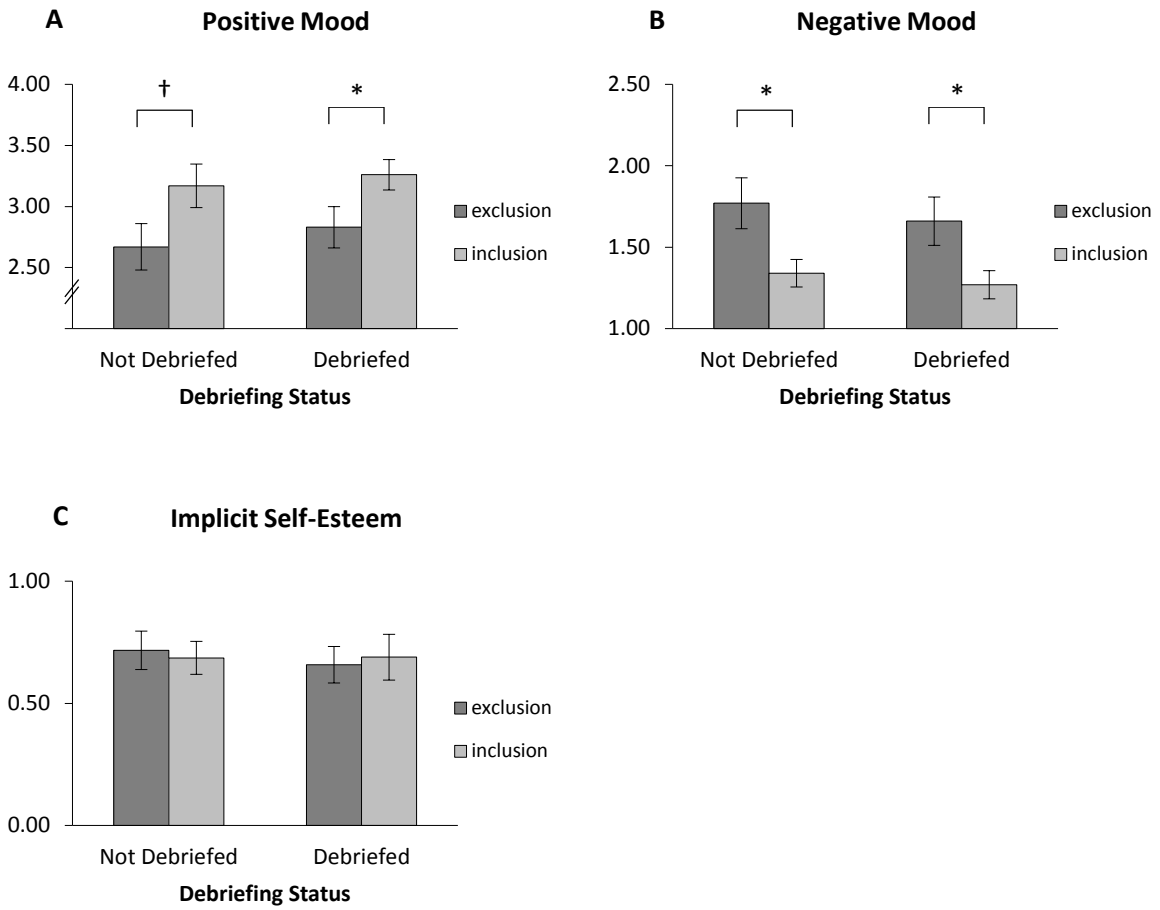
Positive mood. As expected, in the *not debriefed yet* conditions, participants reported less positive mood after *exclusion* (M = 2.67, SD = .95) than after *inclusion* (M = 3.17, SD = .85), $t(44.83) = -1.90$, $p = .064$, $d = -0.55$, although this effect was not quite significant in a two-tailed analysis. Together with the substantial effect size, this establishes that the Cyberball manipulation lowered positive mood, as expected. The simple contrast of the *ostracism condition* in the *debriefed conditions* revealed that the effect of the Cyberball manipulation on positive mood persevered after the debriefing (M_{Excl} = 2.83, SD = .79; M_{Incl} = 3.26, SD = .60), $t(39.15) = -2.077$, $p = .044$, $d = -0.62$. The main effect of the *ostracism condition* was significant across both *debriefing status* conditions (M_{Excl} = 2.74, SD = .87; M_{Incl} = 3.21, SD = .73), $F(1, 88) = 7.66$, $p = .007$, $\eta_p^2 = .08$. There was no interaction between the *ostracism condition* and *debriefing status* $F(1, 88) = 0.04$, $p = .848$, $\eta_p^2 < .01$. Finally, there was no main effect of *debriefing status* (M_{NoDeb} = 2.91, SD = .93; M_{Deb} = 3.05, SD = .72), $F(1, 88) = 0.56$, $p = .456$, $\eta_p^2 = .01$ (Figure 7a). Taken together, exclusion led to decreased positive mood, and the debriefing that was intended to remedy this effect was unsuccessful in doing so.

Negative mood. As expected, in the *not debriefed yet* conditions, participants reported more negative mood after *exclusion* (M = 1.77, SD = .77) than after *inclusion* (M = 1.34, SD = .40), $t(35.33) = 2.37$, $p = .024$, $d = 0.69$. This establishes that the Cyberball manipulation successfully increased negative mood, as expected. The simple contrast of the *ostracism condition* in the *debriefed conditions* revealed that the effect of the Cyberball manipulation on negative mood persevered after the debriefing (M_{Excl} = 1.66, SD = .70; M_{Incl} = 1.27, SD = .41),

$t(33.77) = 2.24$, $p = .032$, $d = 0.68$. The main effect of the *ostracism condition* was significant across both *debriefing status* conditions ($M_{Excl} = 1.71$, $SD = .73$; $M_{Incl} = 1.31$, $SD = .40$), $F(1, 86) = 10.31$, $p = .002$, $\eta_p^2 = .11$. There was no interaction between the *ostracism condition* and *debriefing status* $F(1, 86) = 0.50$, $p = .482$, $\eta_p^2 = .01$. Finally, there was no main effect of *debriefing status* ($M_{NoDeb} = 1.56$, $SD = .65$; $M_{Deb} = 1.47$, $SD = .60$), $F(1, 86) = 0.02$, $p = .890$, $\eta_p^2 < .01$ (Figure 7b). Taken together, exclusion led to increased negative mood, and the debriefing that was intended to remedy this effect was unsuccessful in doing so.

Implicit self-esteem. In the *not debriefed yet* conditions, participants showed similar levels of implicit self-esteem after *exclusion* ($M = 0.72$, $SD = 0.38$) as after *inclusion* ($M = 0.69$, $SD = 0.45$), $t(88) = 0.28$, $p = .783$, $d = 0.07$ (Figure 7c). Thus, the Cyberball manipulation had no significant effect on state self-esteem. Therefore, we stopped the analysis of this dependent variable at this point.

Aftereffects. Fifty-two participants (57% retention rate) completed the online follow-up questionnaire about potential aftereffects of the manipulation on average 41.93 hours after they had left the laboratory ($Md = 10.96$ hours, $SD = 59.83$). All participants had been debriefed when they filled out the follow-up questionnaire. Therefore, only the effect of the *ostracism condition* could be meaningfully analyzed. A one-way analysis of variance (ANOVA) revealed that after *exclusion* compared to *inclusion* participants reported more preoccupation with the study and feeling affected by it in a more negative way ($M_{Excl} = 2.03$, $SD = 0.57$; $M_{Incl} = 1.61$, $SD = 0.48$), $F(1,51) = 8.04$, $p = .007$, $\eta_p^2 = .14$. Separate follow-up analyses of preoccupation with the study and resulting emotions as two different dimensions led to the similar results: after *exclusion* compared to *inclusion* participants tended to *think* about their general likeability as well as about the specific Cyberball study more often ($M_{Excl} = 2.03$, $SD = 0.68$; $M_{Incl} = 1.70$, $SD = 0.56$), $F(1,51) = 3.55$, $p = .065$, $\eta_p^2 = .06$, and also felt *emotionally* affected by it in a more negative way ($M_{Excl} = 2.04$, $SD = 0.62$; $M_{Incl} = 1.51$, $SD = 0.57$), $F(1,51) = 10.09$, $p = .003$, $\eta_p^2 = .16$.



Figures 7a–7c. Positive and negative mood, as well as implicit self-esteem as a function of *ostracism condition* and *debriefing status* at the time of DV measurement. Participants who had been excluded showed lower levels of positive mood and higher levels of negative mood than participants who had been included, no matter whether they had been debriefed at the time of the collection of the respective dependent variable or not. There were no significant effects on implicit self-esteem (as indicated by the IAT D1 score). Error bars indicate +/- 1 SEM. † $p < .10$, *$p < .05$.

Discussion

Participants who had been excluded in the Cyberball game showed less positive and more negative mood compared to participants who had been included in the game. A Revised Outcome Debriefing failed to eliminate these effects: debriefed participants were equally affected by the Cyberball manipulation as participants who had not been debriefed yet. Moreover, participants in the *exclusion* conditions continued to have more negative thoughts and feelings with respect to the experiment several hours after leaving the laboratory even though all participants had already been thoroughly debriefed at that point.

The Cyberball manipulation did not affect participants' implicit self-esteem. This might be due to methodological problems of the construct itself: implicit self-esteem has been criticized for its low convergent validity, its temporal instability, for being susceptible to measurement errors, and for not, in fact, being appreciably correlated to well-being (e.g., Buhrmester, Blanton, & Swann Jr., 2011; Falk, Heine, Takemura, Zhang, & Hsu, 2015; Gawronski, LeBel, & Peters, 2007; Gawronski, Morrison, Phills, & Galdi, 2017).

Overall, the debriefing failed to undo the psychological harm caused by the ostracism manipulation both in the short-term during the laboratory session and also more enduringly during the day(s) after participating in the study.

General Discussion

In two studies, participants were excluded in a virtual ball-tossing game and subsequently indicated their positive and negative mood as well as their self-esteem either before or after receiving the Revised Outcome Debriefing, the most effective debriefing method in prior research (McFarland et al., 2007). Exclusion compared to inclusion led to less positive and more negative mood. The debriefing failed to undo these adverse effects of the ostracism manipulation. Additionally, negative thoughts and feelings regarding the manipulation persevered the debriefing and persisted on average for more than one day (Study 2, $M \approx 42$ h, $Md \approx 11$ h).

Perhaps the present results indicating that a debriefing procedure failed to restore participants' well-being after ostracism should not come as a surprise. Previous research has shown that distress caused by ostracism can be very resilient to moderation by situational factors (Williams, 2007): ostracism can still impair well-being if participants are excluded by a despised outgroup like the KKK (Gonsalkorale & Williams, 2007), if exclusion comes with an incentive (e.g., because inclusion costs or participants' characters are playing with an exploding bomb instead of a ball [van Beest & Williams, 2006; van Beest, Williams, & van Dijk, 2011]), and even if participants know that they are only playing with a pre-programmed

computer script (Zadro et al., 2004). A meta-analysis of 120 Cyberball studies further illustrated the robustness of the ostracism effect: while moderation of the effect can be possible, the relative strength of the ostracism effect itself overpowers moderation, thus leading to negative effects of exclusion even in the presence of a moderator (Hartgerink et al., 2015). Taking this into account, it seems plausible that ostracism still hurts after a thorough debriefing procedure. That notwithstanding, our field displays ubiquitous trust in debriefings' effectiveness to restore participants' well-being. Thus, the present results have substantive implications for psychological researchers and contribute evidence that our trust in debriefings may have been too credulous.

Future Research

Considering the generally impressive resilience of distress caused by ostracism, the question arises how participants' well-being after an ostracism manipulation can be restored after all. Previous research has shown that writing a self-affirming essay about a personally important value might restore participants' subjective well-being after ostracism (Hales, Wesselmann, & Williams, 2016). A different approach for future work could be to create a positive experience for the participants after the ostracism manipulation. Yet, one has to be careful to avoid any deceptive elements in creating that experience. For instance, previous research has indeed shown that after exclusion, receiving messages from ostensible other participants (Rudert, Hales, Greifeneder, & Williams, 2017) or being included in a second round of the Cyberball game (Tang & Richardson, 2013; Zwolinski, 2014) can ameliorate participants' well-being. However, the thereby newly created deception would have to be explained to the participant at the end of the experiment in accordance with the APA's *Code of Conduct* (American Psychological Association, 2002). The selection of a positive post-experimental intervention to counteract the negative effects caused by the ostracism manipulation is thus far from trivial and requires careful consideration.

In addition to that, future work might investigate how long the detrimental aftereffects caused by an ostracism manipulation last and what their real-life consequences are. In Study 2, participants reported feeling negatively affected by the ostracism manipulation several hours after leaving the lab. We assume that this and similar effects will disappear over time, but we currently do not know how long this may take. In this context, it is also unclear what can be considered an ethically acceptable time frame, especially considering the fact that long-lasting aftereffects of ostracism might have a negative impact on participants' real lives as students, friends and partners. For instance, in the present research, participants who had been excluded as compared to included indicated that after their participation in the study, they thought more often about whether other people found them likeable (Study 2). Furthermore, future research about the real-life consequences of experimentally induced ostracism might be particularly important because ostracism is known to negatively affect numerous dimensions such as aggressive behavior (Warburton et al., 2006; Wesselmann et al., 2010), sense of control and belonging, as well as the sense of meaningfulness of life (Gonsalkorale & Williams, 2007; Jamieson et al., 2010; Williams et al., 2000).

It is important for us to underline that with the present research, we do not wish to discourage researchers from using ostracism manipulations in the future. Instead, we hope to stimulate open discussions about several unanswered questions. These questions include not only the aforementioned issue of identifying an effective remedy for adverse experimental effects but also the question of how to interpret the present findings. There might be at least two ways of interpretation: while one could argue that reduced well-being and stressful preoccupation during the day(s) following the experiment qualify as psychological harm and should therefore be seen as problematic, others might argue that these consequences of participating in a study seem rather minor compared to, for example, clinically relevant symptoms of distress and that apart from that, participants freely consented to participating in the study. We hope that the present research encourages open discussions about these and

similar questions about the way that we, as a scientific community, want to conduct research.

Conclusion

Ostracism caused distress, indicated by less positive and more negative mood. A thorough debriefing procedure failed to restore participants' mood and let them leave the laboratory with unwanted aftereffects. Undesired study-related affect and mental preoccupation with the study persisted for more than one day. These findings raise questions about common research practice and encourage future research on finding an effective remedy for adverse experimental effects.

Appendix

Items From the Follow-up Questionnaire (as used in Study 2)

1. How much were the things that happened during the experiment in your mind?
2. How often did you think about the course of the game?
3. How proud did you feel when you thought about the course of the game?
4. How sad did you feel when you thought about the course of the game?
5. Since the experiment, how often did you think about the impression you make on other people?
6. How anxious did you feel when you thought about the course of the game?
7. How happy did you feel when you thought about the course of the game?
8. Since the experiment, how often did you think about whether others find you (un)likeable?
9. How angry did you feel when you thought about the course of the game?
10. How often did thoughts about the course of the game pop into your mind?

PART III

Will a Debriefing Eliminate the Distress Caused by a Trauma Film Paradigm?

Stefanie Miketta, Alexandra Heike Graebener, Tanja Michael, & Malte Friese

Saarland University

Manuscript in preparation for submission.

Abstract

Experimental trauma research often employs the trauma film paradigm where participants watch highly aversive film clips to induce analogue PTSD symptoms. In line with APA requirements, participants' well-being should be fully restored at the end of the study. The most prominent means to eliminate potential harmful effects of participating in research studies is the debriefing. Little is known, however, about the effectiveness of debriefings. The present research investigated whether a debriefing re-established participants' pre-study conditions after participating in study including a trauma film manipulation. Participants watched highly aversive film clips containing depictions of extreme violence. Two days later, participants were debriefed in person by the experimenter. In the present follow-up study, participants indicated their posttraumatic stress at three different time points (T1–T3) as well as subclinical study-related aftereffects of the trauma film manipulation at two different time points (T2 & T3). Measurements took place on the day of the manipulation as well as two days and 20 days after the manipulation, respectively. Participants' level of posttraumatic distress strongly decreased over the course of three weeks, but was not fully eliminated. In addition, participants still reported subclinical study-related aftereffects of the trauma film manipulation after three weeks. These results challenge the effectiveness of debriefings and raise questions about what kinds of aftereffects the research community deems acceptable and appropriate in research employing the trauma film paradigm.

Keywords: Debriefing, Analogue Trauma, Trauma Film Paradigm, Perseverance

Introduction

Posttraumatic stress disorder (PTSD) is a severe mental illness caused by experiencing traumatic events like sexual violence or threatened death (American Psychiatric Association, 2013). PTSD is characterized by distressing symptoms such as intrusive re-experiencing of the traumatic event, flashbacks, negative affect, avoidance of trauma related stimuli, and hypervigilance. The severity of this mental illness highlights the importance of research that aims to find effective treatments for PTSD. To do so, researchers obviously must not induce real traumata in research participants. Instead, researchers rely on an experimental procedure that leads to a less severe analogue trauma: the trauma film paradigm. In this paradigm, healthy research participants view short film clips with highly aversive content like accidents or extreme violence. This procedure is frequently used in trauma research, because it reliably causes subclinical PTSD symptoms, especially intrusive memories, in healthy participants (Bourne et al., 2013; Chou et al., 2014; Clark et al., 2014; Graebener, Michael, et al., 2017; Michael, Holz, & Lass-Hennemann, 2016; Nixon et al., 2007).

The fact that the trauma film paradigm induces an analogue trauma creates, however, an ethical dilemma: On the one hand, research on trauma is crucially important and requires the induction of analogue PTSD symptoms. On the other hand, the American Psychological Association (APA; 2002) mandates that participating in a research study should not lead to noticeable psychological harm.

To avoid psychological harm, the most frequently used measure is a post-experimental debriefing (Baumrind, 1985; Blanck et al., 1992; D. S. Holmes, 1976b; Sharpe & Faye, 2009; S. S. Smith & Richardson, 1983). The APA (2002) defines the intention of a debriefing as minimizing experimentally induced harm: “when psychologists become aware that research procedures have harmed a participant, they take reasonable steps to minimize the harm“ (p. 1070). In studies about posttraumatic stress responses, it is, of course, unrealistic to debrief participants right after the harm has occurred: in order to measure participants’ intrusions, it is

important that participants experience at least an appreciable amount of distress for several days. While this means that researchers have to accept that participants' discomfort continues after they have initially left the lab, researchers certainly agree on the importance of restoring participants' well-being after receiving a debriefing on the last day of the study. Considering the potentially severe effects of posttraumatic symptoms for participants' well-being, it is imperative that debriefing procedures (or any additional measures taken) reliably eradicate participants' distress.

In light of the crucial role that the ability to restore participants' well-being plays in trauma research, the scarcity of literature on whether or not debriefings possess this ability is surprising: while there are a few studies that suggest that debriefings may not be fully effective after manipulations that are used in social psychology (e.g., McFarland et al., 2007; Ross et al., 1975; Silverman et al., 1970; Walster et al., 1967), there are no studies on the effectiveness of debriefings after experimental trauma research.

Very few researchers have addressed the potential perseverance of distressing symptoms after watching a series of aversive film clips. One group of researchers assures that participants' well-being was restored one week after the manipulation: "in our clinical opinion, no participant displayed a significant level of distress at the follow-up session" (E. A. Holmes et al., 2004). The authors leave open, however, what this subjective clinical opinion was based on. Other researchers did not find that distress was fully eliminated at the time of the last measurement (one or two weeks, respectively, after the manipulation; Brewin & Saunders, 2001; M. I. Davies & Clark, 1998).To our knowledge, to this date, there has not been any systematic research on the question of potential perseverance of distressing symptoms and corresponding long-term effects after watching aversive film clips.

Taken together, there is a disturbing discrepancy between the crucial importance of debriefings to effectively eliminate the psychological harm induced in analogue trauma research studies and the scarcity of research that puts this assumed effectiveness to a test. The

present research goes beyond extant research in at least two important ways: First, we investigated the potential perseverance of both clinically relevant distress caused by watching a series of brief aversive film clips and lower-threshold, subclinical study-related well-being. Second, we examined not only whether participants' well-being improved over time, but also whether it was *fully* restored when we collected the dependent variables for the last time. To this end, participants reported their well-being at three different points in time: directly after seeing the aversive film clips, two days later (just before or immediately after receiving a debriefing, depending on the dependent variable), and three weeks after watching the film clips. In the absence of a non-debriefed control group (since omitting a debriefing contradicts the ethical treatment of participants), we investigated the potential perseverance of distress in a conservative way: the effects of the manipulation were subject to both a post-experimental debriefing procedure and the passage of time (three weeks) which by itself likely reduces distress.

The Present Study

Method

Participants and design. Forty-one healthy students (52% females; mean age *M*=23.98, *SD*=3.55) participated in the trauma film study in exchange for €50.[11] We additionally raffled two Amazon gift certificates, worth €20 each, among those who participated in the follow-up. A 1 x 3 repeated measures design was used to examine the effects of the manipulation on clinically relevant posttraumatic stress. A 1 x 2 repeated measures design was used to examine the effects of the manipulation on a measure indicative

[11] The present research is an independent follow-up of a trauma film study (Graebener, Lass-Hennemann, Wilhelm, Ferreira de Sá, & Michael, 2017). In the present paper, the design, methods and results (including demographics) of that follow-up are reported. Specifics of the trauma film study are reported only when they are essential for understanding the procedure of the follow-up. For more detailed information about the trauma film study, see the main report of the study.

of clinically not-relevant study-related aftereffects.

Procedure. After giving informed consent on day 1 of the study, participants watched nine aversive film clips and filled out the measure of posttraumatic stress. On day 3, participants watched one aversive film clip, filled out the measure of posttraumatic stress and received a thorough debriefing in a personal session with the experimenter. Then, they were informed about the follow-up study. After giving informed consent to participate in that follow-up, participants filled out the measure of subclinical study-related aftereffects. Sixteen days after they first came to the lab (i.e. two weeks after they had received the debriefing), participants received a link to an online version of the questionnaire about clinically relevant posttraumatic stress as well as the questionnaire about subclinical study-related aftereffects. After filling out those two questionnaires, participants were thanked and dismissed.

Aversive film clips. Nine aversive film clips with a duration of 16 seconds each were used to induce an analogue trauma. The film clips were generated from commercially available films and depicted scenes of extreme sexual and physical violence (e.g., rape, torture). Most of them had already been shown to successfully induce posttraumatic symptoms in previous studies (Wegerer, Kerschbaum, Blechert, & Wilhelm, 2014; see Appendix for a list of the films).

Debriefing. At the end of day 1, participants were told that the most burdensome part of the study was over and that they would see only one more aversive film clip over the remaining sessions. At the end of day 3, participants were informed that the trauma film study was over and given the monetary compensation. Further, they were invited to talk about the experiment. Finally, contact details of the study supervisor – a trained clinical psychologist – were given to them, together with an encouragement to get in touch in case they felt any kind of uneasiness or distress related to their participation in the study.

Dependent variables.

Clinically relevant posttraumatic stress. To measure clinically relevant posttraumatic

stress, we used the German version (Maercker & Schützwohl, 1998) of the Revised Impact of Event Scale (IES-R; Weiss & Marmar, 1996) with 4-point rating scales (0 = *not at all*, 1 = *rarely,* 3 = *sometimes,* 5 = *frequently*). We report results for the complete IES-R (α_{T1} = .90, α_{T2} = .89,α_{T3} = .79) and separately for the three subscales: Intrusions (7 items, α_{T1} = .82, α_{T2} = .79,α_{T3} = .60), Avoidance (8 items, α_{T1} = .80, α_{T2} = .87,α_{T3} = .71) and Hyperarousal (7 items, α_{T1} = .73, α_{T2} = .75,α_{T3} = .44).

Subclinical study-related aftereffects. Participants indicated their subclinical study-related well-being (2 items, α_{T2} = .78, α_{T3} = .88), their affect when thinking about the study (5 items, α_{T2} = .73, α_{T3} = .78), their affective forecast for the following days (1 item) as well as regret about participating in the study (2 items, α_{T2} = .83, α_{T3} = .90) on 7-point rating scales (see the Appendix for a list of all items).

Results

Clinically relevant posttraumatic stress. The IES-R was assessed on day 1 and day 3 (N = 41). On day 16, participants received an invitation for a final assessment, which they completed on average 20.88 (SD = 4.99) days after the manipulation (N = 23; 56% retention rate).

A repeated measures ANOVA on the IES-R compound score revealed that posttraumatic stress differed depending on how much time had passed since the manipulation, $F(2,44) = 24.85$, $p < .001$, $\eta_p^2 = .53$. Posttraumatic stress decreased from the day of the manipulation until two days later and further decreased until 20 days after the manipulation, all pairwise ts > 3.57, all ps < .001, all d_{av}s > 0.50 (Figure 8; for details about d_{av}, see Lakens, 2013; see Table 1 for descriptive statistics).

Level of intrusions differed across the three measurements, $F(2,44) = 19.92$, $p < .001$, $\eta_p^2 = .48$. Intrusions decreased from the first to the second measurement, and again decreased to the third measurement, all pairwise ts > 1.99, all ps < .053, all d_{av}s > 0.33.

Avoidance of memories and feelings related to the trauma films differed across

measurements, $F(2,44) = 19.87$, $p < .001$, $\eta_p^2 = .48$. Avoidance decreased from the day of the manipulation until two days later, and again until 20 days after the manipulation, all pairwise ts > 3.74, all ps < .001, all d_{av}s > 0.43.

Hyperarousal also differed across measurement occasions, $F(2,44) = 15.92$, $p < .001$, $\eta_p^2 = .42$. Hyperarousal was highest at the day of the manipulation and decreased until two days later and further decreased until 20 days later, all pairwise ts > 3.32, all ps < .002, all d_{av}s > 0.54.

Finally, one-sample *t*-tests revealed that participants' overall level of posttraumatic stress as well as their respective levels of intrusions, avoidance and hyperarousal at all three points in time were significantly different from zero, all ts > 3.29, all ps < .003, all ds > 0.25.

Taken together, participants' levels of posttraumatic stress considerably decreased over the course of almost three weeks. Inspection of the means suggests no clinically relevant strain. However, the overall compound score as well as the subscales remained noticeably above zero (average effect size of d_{av}s > 0.50), indicating unintended aftereffects of the manipulation almost three weeks after the manipulation and more than two weeks after the debriefing.

Subclinical study-related aftereffects. Subclinical study-related aftereffects were assessed on day 3 ($N = 41$) and on average 20.88 ($SD = 4.99$) days after the manipulation ($N = 25$; 61% retention rate). Participants reported improved study-related well-being 20 days after the manipulation compared to two days after the manipulation, $t(24) = 4.19$, $p < .001$, $d_{av} = 0.65$, as well as less negative affect when thinking about the study, $t(18) = 2.97$, $p = .008$, $d_{av} = 0.56$. There were no significant differences in affective forecasting, $t(24) = 0.09$, $p = .930$, $d_{av} = 0.02$, and in regretting having participated in the study, $t(24) = 1.59$, $p = .124$, $d_{av} = 0.22$ (Figure 9; see Table 1 descriptive statistics).

Follow-up analyses of the affect measure revealed that two days after the manipulation, 97.6% of the participants (40 out of 41) reported having thought about the

study during the previous two days. Twenty days after the manipulation, 76% of the participants (19 out of 25) still reported having thought about the study in the previous two weeks. When asked how they felt while thinking about the study, participants reported similar levels of happiness, disgust or calmness, respectively, across measurement occasions, all *t*s(18) < 1.82, all *p*s > .086, all d_{av}s < 0.36. Participants reported lower levels of agitation and shock 20 days after the manipulation compared to two days after the manipulation, both *t*s(18) > 2.80, both *p*s < .012, both d_{av}s > 0.45.

Finally, one-sample *t*-tests revealed that participants' impairment of study-related well-being, negative study-related affect, and regret of having participated in the study in time were significantly different from zero with large effect sizes at both points, all *t*s > 7.16, all *p*s < .001, all *d*s > 1.43.

Taken together, the subclinical study-related aftereffects on well-being decreased over the course of 18 days while participants' level of regret about having participated in the study and their affective forecasting did not change. Some of the effects on affect decreased while others did not change over the course of 18 days. Almost three weeks after the manipulation, three quarters of the participants still reported negative thoughts about the study. The aftereffects on well-being, study-related affect, and regret of having participated in the study remained noticeably above zero with large effect sizes (*d*s > 1.43), indicating unintended aftereffects almost three weeks after the manipulation and more than two weeks after the debriefing.

Table 1

Means and Standard Deviations for Posttraumatic Stress as Well as Study-Related Aftereffects

A) Posttraumatic Stress			
	Time		
	Manipulation Day (T1)	Day 3 (T2)	Day 21 (T3)
IES-R	34.87(21.22)	24.52 (16.39)	8.91 (8.96)
Intrusions	*11.13 (7.59)*	*8.17 (5.70)*	*2.70 (2.27)*
Avoidance	*15.57 (8.97)*	*11.70 (8.34)*	*4.96 (6.18)*
Hyperarousal	*8.17 (6.76)*	*4.65 (5.49)*	*1.26 (1.84)*
B) Study-related Aftereffects			
	Time		
	Manipulation Day (T1)	Day 3 (T2)	Day 21 (T3)
Well-being	-	5.11 (1.66)	5.94 (1.27)
Affect	-	3.13 (1.13)	3.66 (1.12)
Happiness	-	*1.43 (0.93)*	*1.74 (1.41)*
Disgust	-	*4.95 (1.55)*	*4.58 (1.64)*
Calmness	-	*3.03 (1.86)*	*3.95 (1.93)*
Agitation	-	*3.55 (1.63)*	*3.11 (1.41)*
Shock	-	*4.30 (1.74)*	*3.68 (1.60)*
Affective Forecast	-	6.02 (1.35)	5.92 (2.10)
Regret	-	2.41 (1.66)	2.66 (1.86)

Note. Means and standard deviations for Posttraumatic Stress as well as Study-related Aftereffects, including all subscales, on the day of the manipulation, two days after the manipulation, and 20 days after the manipulation.

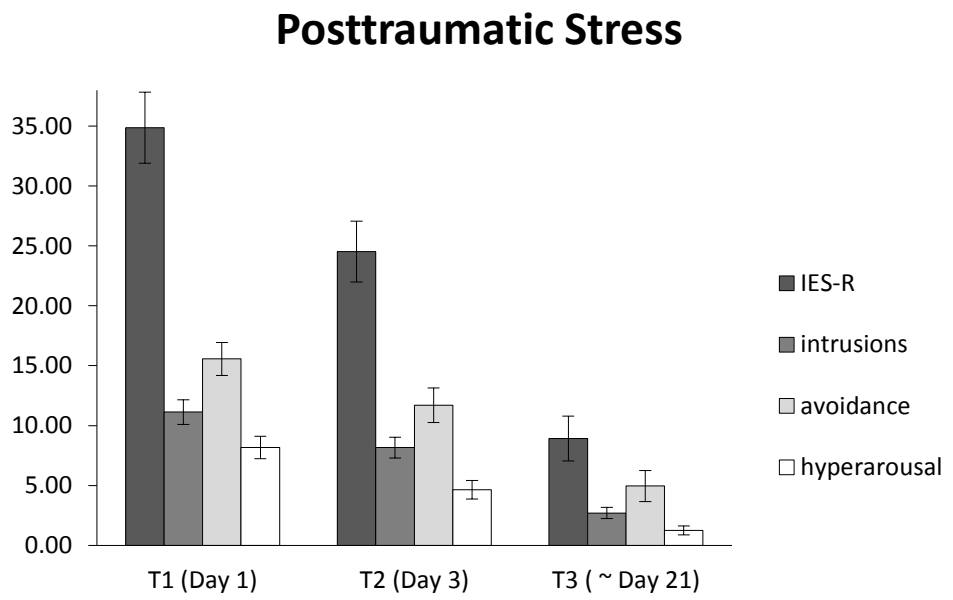


Figure 8. Impact of Event Scale-Revised: Compound score and subscales (intrusions, avoidance, hyperarousal) across measurement occasions. Error bars indicate +/- 1 SEM.

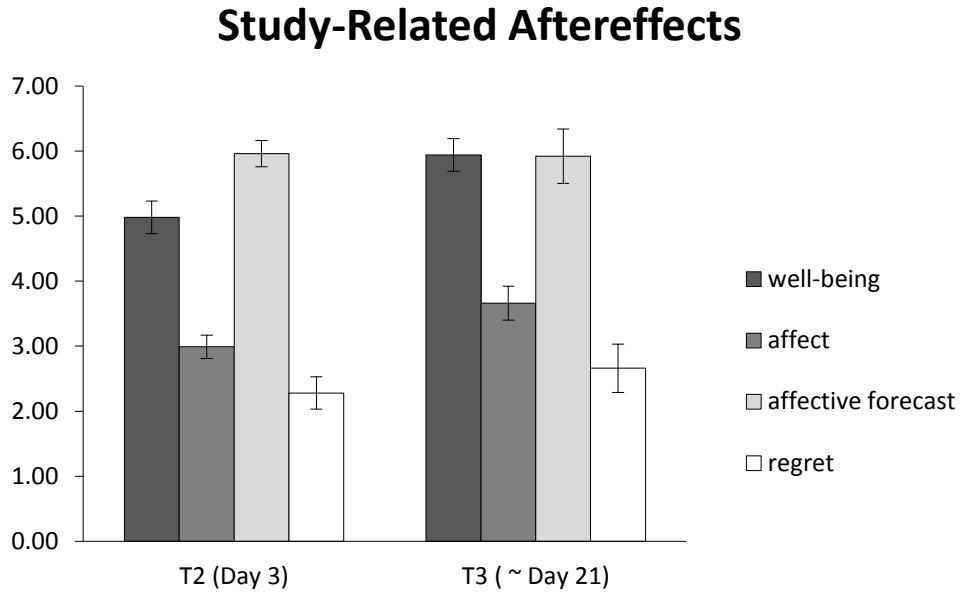


Figure 9. Study-related well-being, study-related affect, affective forecast, and regret about participation in the study at two different points in time. Participants' well-being and affect increased over the course of 18 days while their affective forecast and their regret did not. Error bars indicate +/- 1 SEM.

Discussion

Participants watched highly aversive film clips and were subsequently asked to indicate their levels of clinically relevant posttraumatic stress three different points in time

and their levels of subclinical study-related aftereffects at two different points in time. Measurements took place at day of the manipulation (posttraumatic stress only), two days after the manipulation, and 20 days after the manipulation. A full debriefing of the study was provided to participants two days after the manipulation.

Despite having received a debriefing, watching trauma films led to, first, measurable levels of posttraumatic stress and, second, subclinical study-related aftereffects almost three weeks after the manipulation. Concerning the first, participants' posttraumatic stress decreased considerably while still staying noticeably above zero at the end of the study. Participants' level of posttraumatic stress never reached the threshold for diagnosing a PTSD. However, considering the severity of the symptoms being measured – e.g., intrusions, nightmares, difficulties in concentrating, nervousness, irritability, and dizziness – researchers may see it problematic that participants still experienced *any* of those symptoms three weeks after the manipulation.

Concerning the subclinical study-related aftereffects, these also partially decreased over the course of 18ten days, but also did not reach zero. These aftereffects were measured with items like "the study has negatively affected my well-being" or "the study leaves me with a certain feeling of uneasiness" – items that are not clinically relevant, but that ideally no participant should give an affirmative answer to after three weeks.

Taken together, the joint healing effects of a comprehensive personal debriefing and the passage of almost three weeks of time gone by were not enough to fully undo all of the detrimental effects caused by the trauma film manipulation.

What are the substantial implications of these findings? The APA's (2002) ethical principles state that research participants' harm should be avoided. What constitutes harm, however, is not clearly defined. Should any level of psychological distress be regarded a form of harm? Or should a higher threshold be applied and harm defined as the presence of clinically relevant symptoms? The present research is neither intended nor capable of

answering these questions. Rather, we hope it will stimulate discussion to reflect and ultimately provide consensus about these and other questions:

What kinds of undesired aftereffects on well-being do researchers deem (in)acceptable for participants who gave their free, active and informed consent to participate in a study they knew would include aversive experiences? How long do such aftereffects last and what can be considered an ethically acceptable time frame? In the present study, participants still reported having negative thoughts and feelings about the study three weeks after leaving the lab. We assume that this effect will disappear over time, but we currently have no evidence about how much time this may take. What are effective remedies for negative aftereffects of trauma film manipulations if a debriefing (together with the healing force of three weeks gone by) is not capable of re-establishing pre-study conditions? And, finally, what are real-life consequences of participating in a trauma film study – how do participants' research experiences impact their lives as students, friends and partners?

We want to emphasize that we do not mean to imply that researchers should stop using trauma film manipulations in their work. Important research questions are dependent on researchers' ability to induce an analogue trauma. Neither do we mean to accuse researchers of having disregarded their ethical obligation to care for their participants' psychological well-being. What we do mean, however, is to encourage a reflection and discussion about the outlined questions we think are raised by the present research.

Conclusion

A trauma film manipulation caused psychological distress. A debriefing procedure failed to fully undo this effect and let participants leave the laboratory with aftereffects. Subclinical levels of posttraumatic stress, negative study-related affect, and mental preoccupation with the study persisted for almost three weeks at least. These findings raise questions about research ethics**.**

Appendix

Supplemental Material

Films (With Director & Year) From Which the Aversive Film Clips Shown to Participants Were Taken

- 127 Hours - Danny Boyle (2012)
- Antichrist - Lars von Trier (2009)
- Final Destination - James Wong (2000)
- German Angst - Jörg Buttgereit, Andreas Marschall & Michal Kosakowski (2015)
- Hostel - Eli Roth (2005)
- I spit on your grave 2 - Steven R. Monroe (2013)
- Scar - Jed Weintrob (2007)

Items From the Questionnaire Measuring Study-related Aftereffects (Two Days / Two Weeks After the Manipulation)

1. I feel that the study has negatively affected my well-being
2. Thinking of the study somehow leaves me with a certain feeling of uneasiness
3. During the last two days / weeks, did you (voluntarily or involuntarily) think about the film clips that you watched during the study?[12]
    a. I felt agitated when thinking about the film clips
    b. I felt happy when thinking about the film clips
    c. I felt shocked when thinking about the film clips
    d. I felt calm when thinking about the film clips
    e. I felt disgusted when thinking about the film clips
4. How do you think you will feel during the next two days / weeks?

[12] Items 3a - 3e were only presented if participants answered item 3 with „yes“.

5. Would you have chosen to participate in the study if you had known beforehand how it would make you feel?
6. Would you participate in this study or a similar one for a second time?

Concluding Discussion

In three parts that included eight studies, the effectiveness of post-experimental debriefings and interventions in removing the harm caused by experimental manipulations was investigated. Participants experienced ostracism (Part II), two different kinds of ego threat (Part I) – indirect failure feedback regarding their performance in an alleged IQ test and negative social feedback – and watched traumatic film clips (Part III). In all studies, the experimental manipulations led to detrimental effects, such as negative mood (Parts I & II), lowered self-esteem (Part I), or posttraumatic stress (Part III). Neither the allegedly most effective kind of the debriefing, a Revised Outcome Debriefing (Parts I & II), nor a debriefing combined with a self-affirmation intervention (Part I), nor a debriefing combined with the passage of three weeks of time (Part III) were able to fully remove those adverse effects. An internal meta-analytical summary of the studies including ego threat (Part I) confirmed that the debriefings had only a small *positive* effect on participants' well-being while the ego threat manipulations had a medium-to-large *negative* effect. Negative thoughts and feelings regarding the manipulations overpowered the debriefing interventions and persevered for at least one day (Parts I, II, & III), respectively almost three weeks (Part III) after participants had left the laboratory. The fact that these findings occurred across four different manipulations and three notably different areas of psychological research suggests that they may generalize to other manipulations and other areas of psychological research as well.

The aim of the present research was to investigate the effectiveness of post-experimental debriefings in restoring participants' well-being after potentially harmful experimental manipulations. The debriefing procedures and interventions that were employed in the present research clearly did not meet this goal despite conscious efforts to maximize their effectiveness. Thereby, the use of those debriefing procedures did not fulfill the APA's requirements for researchers to minimize the harm that an experimental manipulation may have caused their participants.

On the one hand, the present findings can be thought of as surprising: Several psychological societies' guidelines advocate the use of post-experimental debriefings to remove any harm that may have occurred and to thereby restore research participants' well-being. This prescription, coupled with the field's enormous trust in debriefings as a remedy, logically implies the effectiveness of post-experimental debriefings – which stands in contrast to the present findings. On the other hand, research on and theories about belief perseverance show that, in general, once given information can hardly be discredited in an effective way. This even applies to information that – in contrast to the manipulations used in the present research – has neither negative valence, nor any personal relevance for participants. Against that background, the present findings do not surprise.

The implications of the present research for the specific context of post-experimental debriefings are twofold:

(1) While the present findings might be unsettling, they provide the psychological research community with an opportunity to engage in open discussions and reflection about how we, as a field, want to conduct research. In particular, this open dialogue could help in answering the question of how we should handle impactful experimental manipulations and post-experimental interventions and how we should define our joint expectations about debriefings' effectiveness.

(2) Furthermore, the present findings show that future research is needed to identify more effective debriefing procedures or other post-experimental interventions. One approach might be to consider the post-experimental debriefing itself to be a second manipulation instead of a mere formality at the end of an experiment. This second manipulation would at least have to match the first, original manipulation in strength and effectiveness. Considering the perseverance effect, one could argue that the second manipulation, i.e. the debriefing, would even have to appreciably exceed the first manipulation's strength in order to be effective. Future research should identify suitable interventions, maybe starting with shedding

more light onto the process behind the perseverance of experimental manipulations' effects. If we, as a field, knew more about the exact mechanism behind the perseverance effect, it would be easier to understand how the effect can be influenced and ameliorated. This could then be of considerable help to advance the development of new, more effective debriefing methods.

More generally, the present research also has implications outside of the specific context of post-experimental debriefings: The experience of receiving negative information that is later retracted is not an experience unique to participating in experimental studies. Instead, this experience is pervasive in people's everyday lives. For instance, most people are familiar with interpersonal conflicts in which hastily spoken words or insults could not simply be taken back but left the recipient with lingering feelings of humiliation and hurt. However, threatening information does not even require interpersonal conflict or malicious intent. Examples for unintentional misinformation include concerns and rumors about health risks of vaccinations, which can ultimately lead to lower immunization rates (Kata, 2010), and fear-driven misinformation about (upcoming) natural disasters, such as earthquakes or tsunamis (Acar & Muraki, 2011). Furthermore, the mere hypothetical possibility of a threat occurring can persevere as well. For instance, research shows that women who had initially been informed of potential breast cancer after a suspicious mammogram experienced increased worrying about the disease even three months after further diagnostic procedures had entirely ruled out the possibility of cancer (Lerman et al., 1991). The question of how to eliminate perseverance of prior given (or implied) information is thus not only relevant for academic discourse about cognitive and affective processes or for the more applied context of ethical treatment of research participants but also for an abundance of everyday situations outside of academia.

Overall, the perseverance effect makes effectively retracting prior given false information a highly challenging task. This affects many areas of everyday life in general as well as post-experimental debriefings after potentially harmful manipulations in particular.

The latter aspect is of great relevance for the scientific psychological community and corresponding questions of research ethics. As demonstrated in the present thesis, adverse effects of experimental manipulations from different areas of psychological research persevered after post-experimental debriefings and interventions. Thus, the question of how to eliminate perseverance after potentially harmful manipulations and resulting ethical and methodological questions about current research practice remain open.

References

Acar, A., & Muraki, Y. (2011). Twitter for crisis communication: Lessons learned from Japan's tsunami disaster. *International Journal of Web Based Communities, 7*, 392–402.

Adair, J. G., Dushenko, T. W., & Lindsay, R. C. L. (1985). Ethical regulations and their impact on research practice. *American Psychologist*, *40*, 59–72.

Adair, J. G., Lindsay, R. C. L., & Carlopio, J. (1983). Social artifact research and ethical regulations: Their impact on the teaching of experimental methods. *Teaching of Psychology*, *10*, 159–162.

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: Author.

American Psychological Association. (1952). A little recent history. *American Psychologist*, *7*, 426–428.

American Psychological Association. (1953). *Ethical standards of psychologists*. Washington, DC: Author.

American Psychological Association. (1973). *Ethical principles in the conduct of research with human participants*. Washington, DC: Author.

American Psychological Association. (1981). Ethical principles of psychologists. *American Psychologist*, *36*, 633–638.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist*, *57*, 1060–1073.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Anderson, C. A., & Kellam, K. L. (1992). Belief perseverance, biased assimilation, and covariation detection: The effects of hypothetical social theories and new data. *Personality and Social Psychology Bulletin*, *18*, 555–565.

Anderson, C. A., Lepper, M. R., & Ross, L. D. (1980). Perseverance of social theories: The role of explanation in the persistence of discredited information. *Journal of Personality and Social Psychology, 39,* 1037–1049.

Anderson, C. A., New, B. L., & Speer, J. R. (1985). Argument availability as a mediator of social theory perseverance. *Social Cognition, 3,* 235–249.

Aronson, E., & Carlsmith, J. M. (1968). Experimentation in social psychology. In G. Lindzey & E. Aronson (Eds.), *The handbook of social psychology: Volume 2 - Research methods* (2nd ed., pp. 1–79). Reading, MA: Addison-Wesley.

Barlett, C., Branch, O., Rodeheffer, C., & Harris, R. (2009). How long do the short-term violent video game effects last? *Aggressive Behavior, 35,* 225–236.

Baron, R. A. (1981). The "costs of deception" revisited: An openly optimistic rejoinder. *IRB: Ethics and Human Research, 3*(1), 8–10.

Bartone, P. T., & Adler, A. B. (1995). *Event-oriented debriefing following military operations: What every leader should know*. Ft. Detrick, Frederick, MD: U.S. Army Medical Research Unit-Europe.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117,* 497–529.

Baumrind, D. (1964). Some thoughts on ethics of research: After reading Milgram's "Behavioral Study of Obedience." *American Psychologist, 19,* 421–423.

Baumrind, D. (1971). Principles of ethical conduct in the treatment of subjects: Reaction to the draft report of the Committee on Ethical Standards in Psychological Research. *American Psychologist, 26,* 887–896.

Baumrind, D. (1972). Reactions to the May 1972 draft report of the ad hoc Committee on Ethical Standards in Psychological Research. *American Psychologist, 27,* 1083–1086.

Baumrind, D. (1985). Research using intentional deception: Ethical issues revisited. *American*

Psychologist, 40, 165–174.

Beer, J. S., Chester, D. S., & Hughes, B. L. (2013). Social threat and cognitive load magnify self-enhancement and attenuate self-deprecation. *Journal of Experimental Social Psychology, 49,* 706–711.

Berg, I. A. (1954). The use of human subjects in psychological research. *American Psychologist, 9,* 108–111.

Birgegard, A., & Sohlberg, S. (2008). Persistent effects of subliminal stimulation: Sex differences and the effectiveness of debriefing. *Scandinavian Journal of Psychology, 49,* 19–29.

Blackhart, G. C., Nelson, B. C., Knowles, M. L., & Baumeister, R. F. (2009). Rejection elicits emotional reactions but neither causes immediate distress nor lowers self-esteem: A meta-analytic review of 192 studies on social exclusion. *Personality and Social Psychology Review, 13,* 269–309.

Blanck, P. D., Bellack, A. S., Rosnow, R. L., Rotheram-Borus, M. J., & Schooler, N. R. (1992). Scientific rewards and conflicts of ethical choices in human subjects research. *American Psychologist, 47,* 959–965.

Blanton, H., Pelham, B. W., DeHart, T., & Carvallo, M. (2001). Overconfidence as dissonance reduction. *Journal of Experimental Social Psychology, 37,* 373–385.

Bobbitt, J. M. (1952). Some arguments for a code of ethics. *American Psychologist, 7,* 428–429.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. Chichester, England: John Wiley & Sons.

Bourne, C., Mackay, C. E., & Holmes, E. A. (2013). The neural basis of flashback formation: The impact of viewing trauma. *Psychological Medicine, 43,* 1521–1532.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014). Continuously cumulating meta-analysis and replicability. *Perspectives on Psychological Science, 9,* 333–342.

Brewin, C. R., & Saunders, J. (2001). The effect of dissociation at encoding on intrusive memories for a stressful film. *British Journal of Medical Psychology*, *74*, 467–472.

British Psychological Society. (2009). *Code of ethics and conduct*. Retrieved from https://www.bps.org.uk/news-and-policy/bps-code-ethics-and-conduct

Brock, T. C., & Becker, L. A. (1966). "Debriefing" and susceptibility to subsequent experimental manipulations. *Journal of Experimental Social Psychology*, *2*, 314–323.

Buhrmester, M. D., Blanton, H., & Swann Jr., W. B. (2011). Implicit self-esteem: Nature, measurement, and a new way forward. *Journal of Personality and Social Psychology*, *100*, 365–385.

Burton, C. M., & King, L. A. (2009). The health benefits of writing about positive experiences: The role of broadened cognition. *Psychology & Health*, *24*, 867–879.

Caporael, L. R. (1997). The evolution of truly social cognition: The core configurations model. *Personality and Social Psychology Review*, *1*, 276–298.

Carr, E. W., Huber, D. E., Pecher, D., Zeelenberg, R., Halberstadt, J., & Winkielman, P. (2017). The ugliness-in-averageness effect: Tempering the warm glow of familiarity. *Journal of Personality and Social Psychology*, *112*, 787–812.

Carroll, J. S. (1978). The effect of imagining an event on expectations for the event: An interpretation in terms of the availability heuristic. *Journal of Experimental Social Psychology*, *14*, 88–96.

Chan, M. S., Jones, C. R., Hall Jamieson, K., & Albarracín, D. (2017). Debunking: A meta-analysis of the psychological efficacy of messages countering misinformation. *Psychological Science*, *28*, 1531–1546.

Choi, I., & Nisbett, R. E. (2000). Cultural psychology of surprise: Holistic theories and recognition of contradiction. *Journal of Personality and Social Psychology*, *79*, 890–905.

Chou, C.-Y., La Marca, R., Steptoe, A., & Brewin, C. R. (2014). Biological responses to

trauma and the development of intrusive memories: An analog study with the trauma film paradigm. *Biological Psychology, 103,* 135–143.

Clark, I. A., Niehaus, K. E., Duff, E. P., Di Simplicio, M. C., Clifford, G. D., Smith, S. M., … Holmes, E. A. (2014). First steps in using machine learning on fMRI data to predict intrusive memories of traumatic film footage. *Behaviour Research and Therapy, 62,* 37–46.

Cobb, M., Nyhan, B., & Reifler, J. (2013). Beliefs don't always persevere: How political figures are punished when positive information about them is discredited. *Political Psychology, 34,* 307–326.

Coccaro, E. F., Noblett, K. L., & McCloskey, M. S. (2009). Attributional and emotional responses to socially ambiguous cues: Validation of a new assessment of social/emotional information processing in healthy adults and impulsive aggressive patients. *Journal of Psychiatric Research, 43,* 915–925.

Cohen, G. L., & Sherman, D. K. (2014). The psychology of change: Self-affirmation and social psychological intervention. *Annual Review of Psychology, 65,* 333–371.

Crawford, J. R., & Henry, J. D. (2004). The Positive and Negative Affect Schedule (PANAS): Construct validity, measurement properties and normative data in a large non-clinical sample. *British Journal of Clinical Psychology, 43,* 245–265.

Davies, M. F. (1982). Self-focused attention and belief perseverance. *Journal of Experimental Social Psychology, 18,* 595–605.

Davies, M. F. (1993). Dogmatism and the persistence of discredited beliefs. *Personality and Social Psychology Bulletin, 19,* 692–699.

Davies, M. F. (1997). Belief persistence after evidential discrediting: The impact of generated versus provided explanations on the likelihood of discredited outcomes. *Journal of Experimental Social Psychology, 33,* 561–578.

Davies, M. I., & Clark, D. M. (1998). Predictors of analogue post-traumatic intrusive

cognitions. *Behavioural and Cognitive Psychotherapy*, *26*, 303–314.

Deutsche Gesellschaft für Psychologie. (2016). *Berufsethische Richtlinien [Professional ethical guidelines]*. Retrieved from https://www.dgps.de/index.php?id=8

DeWall, C. N., Twenge, J. M., Gitter, S. A., & Baumeister, R. F. (2009). It's the thought that counts: The role of hostile cognition in shaping aggressive responses to social exclusion. *Journal of Personality and Social Psychology*, *96*, 45–59.

Dislich, F. X. R., Imhoff, R., Banse, R., Altstötter-Gleich, C., Zinkernagel, A., & Schmitt, M. (2012). Discrepancies between implicit and explicit self-concepts of intelligence predict performance on tests of intelligence. *European Journal of Personality*, *26*, 212–220.

Ecker, U. K., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, *20*, 323–335.

Ecker, U. K., Lewandowsky, S., Fenton, O., & Martin, K. (2014). Do people keep believing because they want to? Preexisting attitudes and the continued influence of misinformation. *Memory & Cognition*, *42*, 292–304.

Ecker, U. K., Lewandowsky, S., Swire, B., & Chang, D. (2011). Correcting false information in memory: Manipulating the strength of misinformation encoding and its retraction. *Psychonomic Bulletin & Review*, *18*, 570–578.

Edwards, K., & Smith, E. E. (1996). A disconfirmation bias in the evaluation of arguments. *Journal of Personality and Social Psychology*, *71*, 5–24.

Eisenberger, N. I., Lieberman, M. D., & Williams, K. D. (2003). Does rejection hurt? An fMRI study of social exclusion. *Science*, *302*, 290–292.

Elms, A. C. (1975). The crisis of confidence in social psychology. *American Psychologist*, *30*, 967–976.

Epstein, Y. M., Suedfeld, P., & Silverstein, S. J. (1973). The experimental contract: Subjects'

expectations of and reactions to some behaviors of experimenters. *American Psychologist, 28,* 212–221.

Falk, C. F., Heine, S. J., Takemura, K., Zhang, C. X., & Hsu, C.-W. (2015). Are implicit self-esteem measures valid for assessing individual and cultural differences? *Journal of Personality, 83,* 56–68.

Farrow, J. M., Lohss, W. E., Farrow, B. J., & Taub, S. I. (1975). Intersubject communication as a contaminating factor in verbal conditioning. *Perceptual and Motor Skills, 40,* 975–982.

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

Fehr, R., Gelfand, M. J., & Nag, M. (2010). The road to forgiveness: A meta-analytic synthesis of its situational and dispositional correlates. *Psychological Bulletin, 136,* 894–914.

Fleming, J., & Arrowood, A. J. (1979). Information processing and the perseverance of discredited self-perceptions. *Personality and Social Psychology Bulletin, 5,* 201–205.

Galinsky, A. D., Stone, J., & Cooper, J. (2000). The reinstatement of dissonance and psychological discomfort following failed affirmations. *European Journal of Social Psychology, 30,* 123–147.

Gawronski, B., LeBel, E. P., & Peters, K. R. (2007). What do implicit measures tell us? Scrutinizing the validity of three common assumptions. *Perspectives on Psychological Science, 2,* 181–193.

Gawronski, B., Morrison, M., Phills, C. E., & Galdi, S. (2017). Temporal stability of implicit and explicit measures: A longitudinal analysis. *Personality and Social Psychology Bulletin, 43,* 300–312.

Gerber, J., & Wheeler, L. (2009). On being rejected: A meta-analysis of experimental

research on rejection. *Perspectives on Psychological Science*, *4*, 468–488.

Gergen, K. J. (1973). The codification of research ethics: Views of a doubting Thomas. *American Psychologist*, *28*, 907–912.

Goh, J. X., Hall, J. A., & Rosenthal, R. (2016). Mini meta-analysis of your own studies: Some arguments on why and a primer on how. *Social and Personality Psychology Compass*, *10*, 535–549.

Gonsalkorale, K., & Williams, K. D. (2007). The KKK won't let me play: Ostracism even by a despised outgroup hurts. *European Journal of Social Psychology*, *37*, 1176–1186.

Graebener, A. H., Lass-Hennemann, J., Wilhelm, F. H., Ferreira de Sá, D. S., & Michael, T. (2017). *Cortisol administration after extinction in a fear conditioning paradigm with traumatic film clips prevents return of fear*. Manuscript submitted for publication.

Graebener, A. H., Michael, T., Holz, E., & Lass-Hennemann, J. (2017). Repeated cortisol administration does not reduce intrusive memories – A double blind placebo controlled experimental study. *European Neuropsychopharmcology*, *27*, 1132–1143.

Greenwald, A. G., & Farnham, S. D. (2000). Using the Implicit Association Test to measure self-esteem and self-concept. *Journal of Personality and Social Psychology*, *79*, 1022–1038.

Greenwald, A. G., Nosek, B. A., & Banaji, M. R. (2003). Understanding and using the Implicit Association Test: I. An improved scoring algorithm. *Journal of Personality and Social Psychology*, *85*, 197–216.

Greitemeyer, T. (2014). Article retracted, but the message lives on. *Psychonomic Bulletin & Review*, *21*, 557–561.

Guenther, C., & Alicke, M. (2008). Self-enhancement and belief perseverance. *Journal of Experimental Social Psychology*, *44*, 706–712.

Hackman, R. C. (1952). Comments on the training of graduate students in ethics in psychology. *American Psychologist*, *7*, 429–430.

Hales, A. H., Wesselmann, E. D., & Williams, K. D. (2016). Prayer, self-affirmation, and distraction improve recovery from short-term ostracism. *Journal of Experimental Social Psychology, 64*, 8–20.

Hall, C. S. (1952). Crooks, codes, and can't. *American Psychologist, 7*, 430–431.

Harris, B. (1988). Key words: A history of debriefing in social psychology. In J. G. Morawski (Ed.), *The Rise of Experimentation in American Psychology* (pp. 188–212). New Haven, CT: Yale University Press.

Hartgerink, C. H. J., van Beest, I., Wicherts, J. M., & Williams, K. D. (2015). The ordinal effects of ostracism: A meta-analysis of 120 Cyberball studies. *PLOS ONE*, *10*, 1–24.

Heatherton, T. F., & Polivy, J. (1991). Development and validation of a scale for measuring state self-esteem. *Journal of Personality and Social Psychology, 60*, 895–910.

Heatherton, T. F., & Vohs, K. D. (2000). Interpersonal evaluations following threats to self: Role of self-esteem. *Journal of Personality and Social Psychology, 78*, 725–736.

Holmes, D. S. (1976a). Debriefing after psychological experiments: I. Effectiveness of postdeception dehoaxing. *The American Psychologist, 31*, 858–867.

Holmes, D. S. (1976b). Debriefing after psychological experiments: II. Effectiveness of postexperimental desensitizing. *American Psychologist*, *31*, 868–875.

Holmes, E. A., Brewin, C. R., & Hennessy, R. G. (2004). Trauma films, information processing, and intrusive memory development. *Journal of Experimental Psychology: General*, *133*, 3–22.

Horka, S. T., & Farrow, B. J. (1970). A methodological note on intersubject communication as a contaminating factor in psychological experiments. *Journal of Experimental Child Psychology, 10*, 363–366.

Horton, R. S., & Sedikides, C. (2009). Narcissistic responding to ego threat: When the status of the evaluator matters. *Journal of Personality*, *77*, 1493–1525.

Howitt, D., & Cramer, D. (2011). *Introduction to research methods* (3rd ed.). Harlow,

England: Pearson Education.

Hughes, E. C. (1952). Psychology: Science and/or profession. *American Psychologist, 7,* 441–443.

Jamieson, J. P., Harkins, S. G., & Williams, K. D. (2010). Need threat can motivate performance after ostracism. *Personality and Social Psychology Bulletin, 36,* 690–702.

Johnson, A. F. (1974). The codification of research ethics: Views of a doubting Thomas: Comment. *American Psychologist, 29,* 470–470.

Johnson, H. M., & Seifert, C. M. (1994). Sources of the continued influence effect: When misinformation in memory affects later inferences. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1420–1436.

Johnson, M. K., Kim, J. K., & Risse, G. (1985). Do alcoholic Korsakoff's syndrome patients acquire affective reactions? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 22–36.

Kata, A. (2010). A postmodern Pandora's box: Anti-vaccination misinformation on the Internet. *Vaccine, 28,* 1709–1716.

Kaufmann, H. (1967). The price of obedience and the price of knowledge. *American Psychologist, 22,* 321–322.

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology, 31,* 457–501.

Kelman, H. C. (1967). Human use of human subjects: The problem of deception in social psychological experiments. *Psychological Bulletin, 67,* 1–11.

Kilham, W., & Mann, L. (1974). Level of destructive obedience as a function of transmitter and executant roles in the Milgram obedience paradigm. *Journal of Personality and Social Psychology, 29,* 696–702.

King, L. A., & Miner, K. N. (2000). Writing about the perceived benefits of traumatic events:

Implications for physical health. *Personality and Social Psychology Bulletin*, *26*, 220–230.

Kinsey, A. C., Pomeroy, W. B., & Martin, C. E. (1948). *Sexual behavior in the human male*. Philadelphia, PA: W.B. Saunders.

Klein, W. M. P., Blier, H. K., & Janze, A. M. (2001). Maintaining positive self-evaluations: Reducing attention to diagnostic but unfavorable social comparison information when general self-regard is salient. *Motivation and Emotion*, *25*, 23–40.

Koole, S. L., Smeets, K., van Knippenberg, A., & Dijksterhuis, A. (1999). The cessation of rumination through self-affirmation. *Journal of Personality and Social Psychology*, *77*, 111–125.

Krohne, H., Egloff, B., Kohlmann, C., & Tausch, A. (1996). Positive and Negative Affect Schedule (PANAS) – German version. *Diagnostica*, *42*, 139–156.

Kruglanski, A. W., & Freund, T. (1983). The freezing and unfreezing of lay-inferences: Effects on impressional primacy, ethnic stereotyping, and numerical anchoring. *Journal of Experimental Social Psychology*, *19*, 448–468.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin*, *108*, 480–498.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12.

Lakens, D., & Etz, A. J. (2017). Too true to be bad: When sets of studies with significant and nonsignificant findings are probably true. *Social Psychological and Personality Science*, *8*, 875–881.

Leary, M. R., Haupt, A. L., Strausser, K. S., & Chokel, J. T. (1998). Calibrating the sociometer: The relationship between interpersonal appraisals and state self-esteem. *Journal of Personality and Social Psychology*, *74*, 1290–1299.

Leary, M. R., Tate, E. B., Adams, C. E., Batts Allen, A., & Hancock, J. (2007). Self-compassion and reactions to unpleasant self-relevant events: The implications of

treating oneself kindly. *Journal of Personality and Social Psychology*, *92*, 887–904.

Leary, M. R., Terry, M. L., Batts Allen, A., & Tate, E. B. (2009). The concept of ego threat in social and personality psychology: Is ego threat a viable scientific construct? *Personality and Social Psychology Review*, *13*, 151–164.

Lekes, N., Guilbault, V., Philippe, F. L., & Houle, I. (2014). Remembering events related to close relationships, self-growth, and helping others: Intrinsic autobiographical memories, need satisfaction, and well-being. *Journal of Research in Personality*, *53*, 103–111.

Lerman, C., Trock, B., Rimer, B. K., Jepson, C., Brody, D., & Boyce, A. (1991). Psychological side effects of breast cancer screening. *Health Psychology*, *10*, 259–267.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*, 106–131.

Lichtenstein, E. (1970). "Please don't talk to anyone about this experiment": Disclosure of deception by debriefed subjects. *Psychological Reports*, *26*, 485–486.

Lieberman, J. D., & Arndt, J. (2000). Understanding the limits of limiting instructions: Social psychological explanations for the failures of instructions to disregard pretrial publicity and other inadmissible evidence. *Psychology, Public Policy, and Law*, *6*, 677–711.

Lieberman, J. D., Solomon, S., Greenberg, J., & McGregor, H. A. (1999). A hot new way to measure aggression: Hot sauce allocation. *Aggressive Behavior*, *25*, 331–348.

Lilienfeld, S. O. (2007). Psychological treatments that cause harm. *Perspectives on Psychological Science*, *2*, 53–70.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology*, *47*, 1231–

1243.

Lord, C. G., Ross, L. D., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology, 37*, 2098–2109.

Maercker, A., & Schützwohl, M. (1998). Erfassung von psychischen Belastungsfolgen: Die Impact of Event Skala-revidierte Version (IES-R) [Measuring psychological consequences of distress: The Impact of Event Scale-revised version (IES-R)]. *Diagnostica, 44*, 130–141.

Maner, J. K. (2014). Let's put our money where our mouth is. If authors are to change their ways, reviewers (and editors) must change with them. *Perspectives on Psychological Science, 9*, 343–351.

Mantell, D. M. (1971). The potential for violence in Germany. *Journal of Social Issues, 27*, 101–112.

Markus, H., & Wurf, E. (1987). The dynamic self-concept: A social psychological perspective. *Annual Review of Psychology, 38*, 299–337.

McFarland, C., Cheam, A., & Buehler, R. (2007). The perseverance effect in the debriefing paradigm: Replication and extension. *Journal of Experimental Social Psychology, 43*, 233–240.

McNamara, J. R., & Woods, K. M. (1977). Ethical considerations in psychological research: A comparative review. *Behavior Therapy, 8*, 703–708.

Mednick, S. A. (1968). The Remote Associates Test. *The Journal of Creative Behavior, 2*, 213–214.

Menges, R. J. (1973). Openness and honesty versus coercion and deception in psychological research. *American Psychologist, 28*, 1030–1034.

Michael, T., Holz, E., & Lass-Hennemann, J. (2016). Analogue PTSD symptoms are best predicted by state rumination. *Journal of Experimental Psychopathology, 8*, 192–213.

Milgram, S. (1963). Behavioral study of obedience. *The Journal of Abnormal and Social Psychology*, *67*, 371–378.

Milgram, S. (1964). Issues in the study of obedience: A reply to Baumrind. *American Psychologist*, *19*, 848–852.

Mills, J. (1976). A procedure for explaining experiments involving deception. *Personality and Social Psychology Bulletin*, *2*, 3–13.

Munro, G. D., & Ditto, P. H. (1997). Biased assimilation, attitude polarization, and affect in reactions to stereotype-relevant scientific information. *Personality and Social Psychology Bulletin*, *23*, 636–653.

Nestler, S. (2010). Belief perseverance: The role of accessible content and accessibility experiences. *Social Psychology*, *41*, 35–41.

Nestor, P. G., & Schutt, R. K. (2014). *Research methods in psychology: Investigating human behavior*. Los Angeles, CA: SAGE Publications.

Nixon, R. D. V., Cain, N., Nehmy, T., & Seymour, M. (2009). The influence of thought suppression and cognitive load on intrusions and memory processes following an analogue stressor. *Behavior Therapy*, *40*, 368–379.

Nixon, R. D. V., Nehmy, T., & Seymour, M. (2007). The effect of cognitive load and hyperarousal on negative intrusive memories. *Behaviour Research and Therapy*, *45*, 2652–2663.

Nyhan, B., & Reifler, J. (2010). When corrections fail: The persistence of political misperceptions. *Political Behavior*, *32*, 303–330.

Perry, L. B., & Abramson, P. R. (1980). Debriefing: A gratuitous procedure? *American Psychologist*, *35*, 298–299.

Resnick, J. H., & Schwartz, T. (1973). Ethical standards as an independent variable in psychological research. *American Psychologist*, *28*, 134–139.

Rich, G. J. (1952). A code of ethics is needed. *American Psychologist*, *7*, 440–441.

Richter, T., Schroeder, S., & Wöhrmann, B. (2009). You don't have to believe everything you read: Background knowledge permits fast and efficient validation of information. *Journal of Personality and Social Psychology, 96*, 538–558.

Ross, L. D., Lepper, M. R., & Hubbard, M. (1975). Perseverance in self-perception and social perception: Biased attributional processes in the debriefing paradigm. *Journal of Personality and Social Psychology, 32*, 880–892.

Ross, L. D., Lepper, M. R., Strack, F., & Steinmetz, J. (1977). Social explanation and social expectation: Effects of real and hypothetical explanations on subjective likelihood. *Journal of Personality and Social Psychology*, *35*, 817–829.

Rudert, S. C., Hales, A. H., Greifeneder, R., & Williams, K. D. (2017). When silence is not golden: Why acknowledgment matters even when being excluded. *Personality and Social Psychology Bulletin*, *43*, 678–692.

Rudolph, A., Schröder-Abé, M., & Schütz, A. (2009). *Development and validation of a German-language version of the State Self-Esteem Scale.* Manuscript in preparation.

Sanford, F. H. (Ed.). (1952). Discussion on ethics [Special section]. *American Psychologist, 7*, 425–455.

Schmeichel, B. J., & Martens, A. (2005). Self-affirmation and mortality salience: Affirming values reduces worldview defense and death-thought accessibility. *Personality and Social Psychology Bulletin*, *31*, 658–667.

Seeman, J. (1969). Deception in psychological research. *American Psychologist, 24*, 1025–1028.

Shanab, M. E., & Yahya, K. A. (1977). A behavioral study of obedience in children. *Journal of Personality and Social Psychology, 35*, 530–536.

Sharpe, D., & Faye, C. (2009). A second look at debriefing practices: Madness in our method? *Ethics & Behavior*, *19*, 432–447.

Sherman, D. K., & Kim, H. S. (2002). Affective perseverance: The resistance of affect to

cognitive invalidation. *Personality and Social Psychology Bulletin*, *28*, 224–237.

Sherman, D. K., & Kim, H. S. (2005). Is there an "I" in "team"? The role of the self in group-serving judgments. *Journal of Personality and Social Psychology, 88*, 108–120.

Shrira, I., & Martin, L. L. (2005). Stereotyping, self-affirmation, and the cerebral hemispheres. *Personality and Social Psychology Bulletin*, *31*, 846–856.

Sieber, J. E. (1983a). Deception in social research II: Evaluating the potential for harm or wrong. *IRB: Ethics & Human Research*, *5*(1), 1–6.

Sieber, J. E. (1983b). Deception in social research III: The nature and limits of debriefing. *IRB: Ethics & Human Research*, *5*(3), 1–4.

Silverman, I., Shulman, A. D., & Wiesenthal, D. L. (1970). Effects of deceiving and debriefing psychological subjects on performance in later experiments. *Journal of Personality and Social Psychology*, *14*, 203–212.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359–1366.

Smith, C. M. (2005). Origin and uses of primum non nocere—Above all, do no harm! *The Journal of Clinical Pharmacology*, *45*, 371–377.

Smith, E. R., Murphy, J., & Coats, S. (1999). Attachment to groups: Theory and measurement. *Journal of Personality and Social Psychology*, *77*, 94–110.

Smith, R. A., & Davis, S. F. (2003). The changing face of research methods. In S. F. Davis (Ed.), *Handbook of research methods in experimental psychology* (pp. 106–126). Oxford, England: Blackwell Publishing.

Smith, S. S., & Richardson, D. (1983). Amelioration of deception and harm in psychological research: The important role of debriefing. *Journal of Personality and Social Psychology*, *44*, 1075–1082.

Somerville, L. H., Heatherton, T. F., & Kelley, W. M. (2006). Anterior cingulate cortex

responds differentially to expectancy violation and social rejection. *Nature Neuroscience*, *9*, 1007–1008.

Sommers, R., & Miller, F. G. (2013). Forgoing debriefing in deceptive research: Is it ever ethical? *Ethics & Behavior*, *23*, 98–116.

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology*, *15*, 72–101.

Swann, W. B., Hixon, J. G., Stein-Seroussi, A., & Gilbert, D. T. (1990). The fleeting gleam of praise: Cognitive processes underlying behavioral reactions to self-relevant feedback. *Journal of Personality and Social Psychology*, *59*, 17–26.

Tang, H. H. Y., & Richardson, R. (2013). Reversing the negative psychological sequelae of exclusion: Inclusion is ameliorative but not protective against the aversive consequences of exclusion. *Emotion*, *13*, 139–150.

Tannenbaum, S. I., & Cerasoli, C. P. (2013). Do team and individual debriefs enhance performance? A meta-analysis. *Human Factors*, *55*, 231–245.

Taylor, S. E., Lerner, J. S., Sherman, D. K., Sage, R. M., & McDowell, N. K. (2003). Portrait of the self-enhancer: Well adjusted and well liked or maladjusted and friendless? *Journal of Personality and Social Psychology*, *84*, 165–176.

Tesch, F. E. (1977). Debriefing research participants: Though this be method there is madness to it. *Journal of Personality and Social Psychology*, *35*, 217–224.

Tesser, A. (2000). On the confluence of self-esteem maintenance mechanisms. *Personality and Social Psychology Review*, *4*, 290–299.

Topolinski, S., & Strack, F. (2010). False fame prevented: Avoiding fluency effects without judgmental correction. *Journal of Personality and Social Psychology*, *98*, 721–733.

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

Twenge, J. M., Baumeister, R. F., Tice, D. M., & Stucke, T. S. (2001). If you can't join them,

beat them: Effects of social exclusion on aggressive behavior. *Journal of Personality and Social Psychology, 81,* 1058–1069.

Twenge, J. M., & Campbell, W. K. (2003). "Isn't it fun to get the respect that we're going to deserve?" Narcissism, social rejection, and aggression. *Personality and Social Psychology Bulletin, 29,* 261–272.

Ullman, D., & Jackson, T. T. (1982). Researchers' ethical conscience: Debriefing from 1960 to 1980. *American Psychologist, 37,* 972–973.

van Beest, I., & Williams, K. D. (2006). When inclusion costs and ostracism pays, ostracism still hurts. *Journal of Personality and Social Psychology, 91,* 918–928.

van Beest, I., Williams, K. D., & van Dijk, E. (2011). Cyberbomb: Effects of being ostracized from a death game. *Group Processes & Intergroup Relations, 14,* 581–596.

Vohs, K. D., & Heatherton, T. F. (2001). Self-esteem and threats to self: Implications for self-construals and interpersonal perceptions. *Journal of Personality and Social Psychology, 81,* 1103–1118.

Vohs, K. D., & Heatherton, T. F. (2004). Ego threat elicits different social comparison processes among high and low self-esteem people: Implications for interpersonal perceptions. *Social Cognition, 22,* 168–191.

Walster, E., Berscheid, E., Abrahams, D., & Aronson, V. (1967). Effectiveness of debriefing following deception experiments. *Journal of Personality and Social Psychology, 6,* 371–380.

Warburton, W. A., Williams, K. D., & Cairns, D. R. (2006). When ostracism leads to aggression: The moderating effects of control deprivation. *Journal of Experimental Social Psychology, 42,* 213–220.

Watson, D., & Clark, L. A. (1999). *The PANAS-X: Manual for the Positive and Negative Affect Schedule - Expanded form*. Iowa City, IA: University of Iowa.

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief

measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology, 54,* 1063–1070.

Weber, S. J., & Cook, T. D. (1972). Subject effects in laboratory research: An examination of subject roles, demand characteristics, and valid inference. *Psychological Bulletin, 77,* 273–295.

Wegerer, M., Kerschbaum, H., Blechert, J., & Wilhelm, F. H. (2014). Low levels of estradiol are associated with elevated conditioned responding during fear extinction and with intrusive memories in daily life. *Neurobiology of Learning and Memory, 116,* 145–154.

Weinstein, N., & Ryan, R. M. (2010). When helping helps: Autonomous motivation for prosocial behavior and its influence on well-being for the helper and recipient. *Journal of Personality and Social Psychology, 98,* 222–244.

Weiss, D. S., & Marmar, C. R. (1996). The Impact of Event Scale - Revised. In J. Wilson & T. M. Keane (Eds.), *Assessing psychological trauma and PTSD* (pp. 399–411). New York, NY: Guilford.

Wesselmann, E. D., Butler, F. A., Williams, K. D., & Pickett, C. L. (2010). Adding injury to insult: Unexpected rejection leads to more aggressive responses. *Aggressive Behavior, 36,* 232–237.

West, S. G., & Gunn, S. P. (1978). Some issues of ethics and social psychology. *American Psychologist, 33,* 30–38.

White, L. A. (1979). Erotica and aggression: The influence of sexual arousal, positive affect, and negative affect on aggressive behavior. *Journal of Personality and Social Psychology, 37,* 591–601.

Williams, K. D. (2007). Ostracism. *Annual Review of Psychology, 58,* 425–452.

Williams, K. D., Cheung, C. K. T., & Choi, W. (2000). Cyberostracism: Effects of being ignored over the Internet. *Journal of Personality and Social Psychology, 79,* 748–762.

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

Wyer, N. A. (2010). You never get a second chance to make a first (implicit) impression: The role of elaboration in the formation and revision of implicit impressions. *Social Cognition, 28*, 1–19.

Zadro, L., Williams, K. D., & Richardson, R. (2004). How low can you go? Ostracism by a computer is sufficient to lower self-reported levels of belonging, control, self-esteem, and meaningful existence. *Journal of Experimental Social Psychology, 40*, 560–567.

Zajonc, R. B. (1980). Feeling and thinking: Preferences need no inferences. *American Psychologist, 35*, 151–175.

Zimbardo, P. G. (1965). The effect of effort and improvisation on self-persuasion produced by role-playing. *Journal of Experimental Social Psychology*, *1*, 103–120.

Zimbardo, P. G., Andersen, S. M., & Kabat, L. G. (1981). Induced hearing deficit generates experimental paranoia. *Science*, *212*, 1529–1531.

Zwolinski, J. (2014). Does inclusion after ostracism influence the persistence of affective distress? *Group Dynamics: Theory, Research, and Practice*, *18*, 282–301.