# Table of Contents

## Editorial

1. **Project Team and Editors of JOTE**  
   Science Fails. Let’s Publish

6. **Rebecca C. Sindall & Dani J. Barrington**  
   Fail Fast, Fail Forward, Fail Openly: The Need to Share Failures in Development

## Empirical/Reflection

9. **Julie Leboeuf, Stine Linden-Andersen, Jonathan Carriere**  
   Alcohol Cues and their Effects on Sexually Aggressive Thoughts

21. **Ruud Abma**  
   Experiment and Fail: a Comment on “Alcohol Cues and their Effects on Sexually Aggressive Thoughts”

27. **Juliane Traxler, Roxane V. Philips, Andreas von Leupoldt, Johan W. S. Vlaeyen**  
   Trial and Error (-Related Negativity): An Odyssey of Integrating Different Experimental Paradigms

39. **Maarten Derksen**  
   Reflection on “Trial and Error (-Related Negativity)”
<table>
<thead>
<tr>
<th>Page</th>
<th>Author</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>43</td>
<td>Kristel De Groot</td>
<td>Burst Beliefs – Methodological Problems in the Balloon Analogue Risk Task and Implications for Its Use</td>
</tr>
<tr>
<td>53</td>
<td>Nicole Nelson</td>
<td>Towards an Expanded Understanding of Publication Bias</td>
</tr>
<tr>
<td>59</td>
<td>Veronika Chelpygina</td>
<td>CrowdDetective: Wisdom of the Crowds for Detecting Abnormalities in Medical Scans</td>
</tr>
<tr>
<td>72</td>
<td>René H. F. P. Bekkers</td>
<td>Global Giving</td>
</tr>
</tbody>
</table>
Science Fails. Let’s Publish

Sean Devine¹, Max Bautista Perpinya², Valentine Delrue³,⁴, Stefan Gaillard²,⁵, Thomas F. K. Jorna², Martijn van der Meer⁵,⁶, Lottricia Millett⁷, Chelsea Pozzebon⁸, Jobke Visser⁹

¹Department of Psychology, McGill University, Montreal, Canada
²Descartes Centre for the History and Philosophy of the Sciences and the Humanities, Utrecht University, Utrecht, the Netherlands
³Department of History, Ghent University, Ghent, Belgium
⁴Department of Philosophy and Cultural Heritage at Ca’ Foscari, Venice, Italy
⁵Open Science Community Utrecht, Utrecht University, Utrecht, the Netherlands
⁶Tilburg School of Humanities and Digital Sciences, Tilburg University, the Netherlands
⁷School of English, Drama and Creative Studies, University of Birmingham, Birmingham, United Kingdom
⁸Faculty of Medicine and Health Sciences, McGill University, Montreal, Canada
⁹Department of Philosophy & Religious Studies, Utrecht University, Utrecht, the Netherlands

†Sean Devine is the lead author. All other authors contributed equally.

Correspondence: JOTE Team, Europaplein 812, 3526WR Utrecht, the Netherlands
Email: info@jtrialerror.com

A Journal of Trial and Error

With the publication of our first issue, we are pleased to formally introduce the Journal of Trial and Error (JOTE). Our goal in this editorial article is to (i) outline the problem JOTE aims to solve, (ii) detail our solution, and (iii) state our goals as a journal. As readers will see, we communicate knowledge rather differently from other platforms. The members of JOTE are unambiguously and unabashedly aware of this departure from the status quo, which we regard as a necessary one in the current academic context. In other respects, however, our mission keeps with an ethos of openness and transparency that has fueled scientific collaboration for centuries. In this regard, our aims are not so different from those formulated in the first issue of Nature, still regarded one of the most prestigious scientific outlets:

A Journal of Trial and Error

What is our problem?

Academia, which we take to refer to all disciplines of inquiry practiced in universities and research centers—not only to science—is radically fragmented nowadays. Our analysis applies primarily to the natural sciences, but we believe some of the challenges we identify apply to the social sciences and the humanities as well. That said, we shall be wary of attempting to find a one-size-fits-all approach to a reform in academia and be attentive to the particularities of each discipline. From here onwards, we will use the term ‘science’ as it is understood in the Anglo-Saxon world: As a term that includes natural sciences, but not humanities. With this in mind, we argue that the way science is interpreted, perceived, and consumed differs substantially from the way it is practiced, theorized, and produced. To characterize this fragmentation, we identify three gaps in science:

I. There is a gap between the image of science and science as practiced, that is, between the science that we see and the science that is done. Publicly, science is thought of as an unambiguous set of operations that gleans truth from chaos. By applying a set of specialized methodologies in a well-delineated process, the “blooming, buzzing confusion” of experience can be quantified, categorized, and systematized. But in practice, science is quite messy. Experiments rarely work on the first try or as expected, results rarely falsify hypotheses directly, and knowledge is gained in small, incremental steps instead of great leaps. This process of fine-tuning is at odds with the public image of science as an enterprise designed for consistent large-scale

Published: 2 December, 2020
jtrialerror.com
There is a gap between what is researched and what is published. A mindset of impact-or-perish dominates academia and contributes to the persistence of positive publication bias. Here, we use the term impact-or-perish instead of the more common publish-or-perish because we believe it better captures the expectations put on academics today: to not only publish, but also to impact, to discover, to amaze, and to flabbergast readers. These institutional pressures widen the space between the research scientists conduct and what is available in the published literature. In the face of failed research — outcomes that do not meet the initial aim of the individual researchers — scientists have two options: not publishing at all or framing the results as productive by engaging in questionable research practices. In line with the self-image of science as suppressing errors and overemphasizing productive big discoveries, this impact-or-perish mindset reinforces the idea that the role of the scientist is to succeed, not to inquire.

There is a gap between research that is popular and research that is replicable. The pressures of big-discovery science and impact-or-perish attitudes push scientists to engage in questionable research practices that result in the publication of big stories that do not replicate. Since the true nature of an effect can only be ascertained in the aggregate, through replicability, the present-day situation of unreplicable experiments is a serious problem. Any individual study, no matter how rigorous, could fail to generalize due to the error inherent in testing a single sample from a broader population. While this ambiguity can never be resolved (we can never be absolutely sure of the magnitude of an effect), we can improve our confidence in a result by aggregating many single studies. For this to work, however, the data available to the meta-scientists doing this must accurately reflect the state of research in a given area. In this respect, recent debate on the replication crisis has (rightfully) focused on the misuse of statistics by scientists, on methodological carelessness, or on theoretical inappropriateness. However, there has been less discussion about what to do when initial hypotheses are not substantiated.

Together these gaps represent a problem that we hope to address with JOTE. Specifically, we want to provide an outlet for research as it is practiced to be shared and discussed. In doing so, we aim to sketch a new scientific image, one that better matches what is actually happening in laboratories and in research institutes around the world. If we can do this, we believe it will improve modern science in many ways. First, it will improve academic integrity and promote honesty in scientific reporting by enabling scientists to publish their otherwise ‘unpublishable’ results in a journal format, still academia’s most important platform for its economy of reputation. Second, by the same means, it will reduce the incalculable number of man-hours lost to “failed” research by reframing what is meant by failure in science. Third, it will greatly improve the magnitude of an effect, we can improve our confidence in a result by aggregating many single studies. For this to work, however, the data available to the meta-scientists doing this must accurately reflect the state of research in a given area. In this respect, recent debate on the replication crisis has (rightfully) focused on the misuse of statistics by scientists, on methodological carelessness, or on theoretical inappropriateness. However, there has been less discussion about what to do when initial hypotheses are not substantiated.

What is trial and error?

We interpret ‘trial and error’ as the elementary process in knowledge production. In other words, we argue that every instance of scientific inquiry relies on a commitment to the process of trial and error. Put simply, ‘trial’ refers to trying something new, exploring and/or testing a new hypothesis or methodology, or in some other form pushing against the boundaries of what is known. ‘Error’ refers to obtaining some kind of unexpected result from those trials, be that negative, ambiguous, or going against some long-held belief. ‘Trial and error’ is an inherent and fundamentally collaborative mechanism of the scientific process, whereby scientists share knowledge of both successes and failures to inform future endeavors. To know what is, we must know what is not.

When is trial and error productive?

We differentiate between two equally useful types of errors: methodological errors and conceptual flaws. Methodological errors refer to procedural, observational, statistical, and other broadly practical errors. Reflecting on and correcting for these errors helps drive improvement in the understanding and application of scientific techniques. Therefore, sharing these results helps increase technical rigor in a given scientific field.

Conceptual flaws refer to patterns or pieces of our understanding of particular phenomena which seem to conflict with experimental observation. Reflecting on and exposing these flaws, especially in our own research, allows other researchers to update their belief in certain theories. In other cases, it could lead to a reworking of entire frameworks. Thus, not only hits (positive results), but also misses (negative results) are key to scientific progress.

How does trial and error relate to various disciplines?

In any empirical inquiry, trial and error are integral and fairly intuitive concepts to understand. However, at JOTE, we believe that all academic disciplines are indispensable in constructively reflecting on trial and error. To this end, we highlight two ways in which interdisciplinarity plays a central role at JOTE. First, scholars from other fields can offer different perspectives on negative results. Due to their inherently ambiguous status, negative results naturally lend themselves to discussion, philosophizing, and problematizing. They are anchors in contextualizing science. As such, one of our key objectives is to encourage the discussion of negative results between all branches of academia to reap the full benefits of reflection on failure (see How do we publish?). Second, we recognize the unique nature of error within the humanities. A culture of success has likewise incentivized humanities scholars to evaluate their work mainly in terms of originality and novelty. Much like in the sciences then, this type of academic economy promotes the spread of grand conclusions that are usually only supported by partial accounts. However, the way the humanities relate to these practices is different, such that humanities scholars are strongly trained to self-reflect about their own limitations and failures. For us, the challenge so far has been to provide a space for the reflections to be formalized and shared not only with other humanities scholars, but also with researchers across disciplinary borders. Therefore, while being aware of (and fascinated by) methodological differences between the humanities and the sciences, we certainly want to provide a platform to highlight, learn...
from, and reflect on errors in the humanities. Lastly, JOTE aims to highlight a ubiquitous but underrepresented form of academic failure: rejected grant applications. Indeed, as an institutional consequence of the highly competitive grant system in academia, many extraordinary proposals are left unfunded. Across all disciplines, a significant amount of research potential is wasted as funding rejections, despite re-applications, too often end in abandoning the project or having to leave academia altogether.

**Why do we want to publish trial and error?**

In addition to the arguments we have given above in favor of trial and error, we want to publish trial and error for (at least) the following four reasons. First, we want to provide an openly accessible platform that reconstructs the self-image of researchers who have most affected by the impact-or-perish mentality (graduate students, postdocs, early-career researchers, etc.). For these reasons, by simply publishing trial and error, a process we regard as integral to the scientific process, we hope to improve scientific and academic culture on multiple fronts.

**How do we publish?**

To accomplish the goals listed above, we have developed a unique publication style for JOTE. In this section, we will describe the types of articles we publish, how these articles fit within JOTE’s overall mission, and how our editorial process works. In addition, we will point readers to articles of each kind that appear in this first issue.

**Empirical articles**

Empirical articles describe conceptual developments based on empirical and experimental studies which have produced null, unexpected, negative, or mixed results. These articles may contain incomplete findings or lack a closed, cohesive narrative. We are aware that authorship still is academic coinage in a severe and competitive economy of reputation. Therefore, by providing an indexable publication for null or inconclusive results, we hope to reduce institutional pressures to communicate positive results. Furthermore, we hope that having such an outlet will discourage the wide array of questionable research practices that researchers may engage in to make their results statistically significant and suitable for publication elsewhere, such as p-hacking, selective reporting, or even falsifying data. Finally, publications in this category will help improve meta-analyses by showcasing null results that would otherwise not be accessible to scholars. We further encourage these efforts by highly the open sharing of data for all submitted empirical articles. See Leboeuf et al. (2020) and Traxler et al. (2020) in this issue for two examples of such empirical articles.

**Reflection articles**

By their very nature, negative results are difficult to interpret. In light of this, JOTE invites other researchers to reflect on and contextualize the issues raised in the original empirical study in a reflection article that accompanies each empirical article. JOTE welcomes reflections that frame the original results in the context of existing literature, comments on the techniques used, and/or discussions about the epistemological and ontological implications of the original study’s findings. Critically, we invite authors from the humanities and other disciplines to contribute to the journal in the form of reflection articles. By doing so, we aim to foster crosstalk between disciplines and create a dialogue about the broad issues surrounding failure in science. See Abma (2020) in this issue for an example of a reflection article on Leboeuf et al. (2020), or Derksen (2020)’s reflection on the piece by Traxler et al. (2020).

**Meta-research articles**

Thirdly, JOTE publishes meta-research articles. These articles pertain to “research on research” studies. Here, we welcome reports of methodological challenges, suggestions, or technical flaws that carry relevant information (i.e., advice on the dos and don’ts) for the field to which they belong, as well as studies that deal with the processes of trial and error at a systemic level, such as on publication bias or replication. We hope to connect discussions on the institutional, social, cultural, and political context of knowledge production with reflections on methodological and conceptual challenges. This category of articles is the broadest and aims to highlight the important work done in the meta-science community and to promote meta-analytic thought across the sciences. See De Groot (2020) and Nelson (2020) for two examples of the heterogeneous category of meta-research articles.

**Rejected Grant Applications**

Lastly, JOTE publishes rejected grant applications together with their anonymous reviewers’ comments. We believe there are at least three reasons why doing this is valuable. Firstly, grant applications are valuable in and of themselves. Proposals are overviews of the state of the field in which the author(s) has/have identified a gap in knowledge. Secondly, publishing rejected proposals is valuable to shape the future of science. These proposals contain metadata that catalogue the ideas, hypotheses, and theoretical perspectives that circulate within an academic field. As such, by providing these proposals with a platform, we formalize a body of knowledge that would otherwise be completely inaccessible to other researchers. Thirdly, failed grant applications are historically valuable. Scientific ideas rarely develop linearly. Instead, they are molded by a complex network of competing hypotheses and theoretical viewpoints. Thus, as a source for the initial, pre-experimental formulations of these hypotheses, research proposals provide historians of science with a valuable means of tracking the history and development of a scientific idea. Similarly, they provide metadata on the trends and biases in the process of awarding grants itself, which in turn could be used to understand what ideas are ‘in the air’ at a given point in history. For two examples of rejected grant applications in this issue, see Bekkers (2020) and Cheplygina (2020).
JOTE’s editorial pipeline
To emphasize our commitment to transparency on all fronts, we believe it is important to describe our publishing process. JOTE focuses on failed research, but this does not mean we publish sloppy science. All submitted articles are peer-reviewed by experts from the article’s field and according to its disciplinary conventions. Throughout this process, we are committed to principles of transparency, fairness, and Open Science. As such, all peer-reviews are double-blind and available post-publication.

Following submission of an article, our editors contact potential reviewers to ensure that the article adheres to the standards of its field. Once a minimum of two reviewers agree that it does, the first of up to three review rounds begins. During this process, editors contact potential authors for a reflection piece. After reflection, pieces undergo editorial review and are finalized, authors of the original empirical paper have the opportunity to respond to the reflection piece. This process can be repeated until all parties involved are satisfied with the quality of the pieces. Once both the author(s) of the empirical article and those of the reflection piece are satisfied, the pieces are ready for publication. The articles as well as the peer-reviews and back-and-forth discussion between the authors are prepared for publication and receive a Digital Object Identifier (DOI) for database indexing. In the case of meta-research articles, the process is the same as for empirical articles, but there is no reflection piece associated with the article.

Science is not static, but publications, especially PDFs, unfortunately are. Although peer-reviewers are traditionally perceived as gatekeepers of scientific quality and trustworthiness of the publication, we believe it is naive to assume that two other experts in addition to the original author(s) can speak on behalf of the whole academic community. In line with our emphasis on trial and error, we do not see peer-reviewed publications as final objects, and we believe that an article is never truly finished, not even after it has been reviewed, proofread, type-set, and copy-edited. We therefore use the interactive online platform ‘PubPub’ to enable readers to become active participants in the publication process by allowing and encouraging comments on the online version of every article. We advocate for discussion rather than retraction where possible.

Fees and Funding
We firmly believe that academic knowledge should be accessible to the community in its broadest sense. For this reason, all our articles are published in Open Access. That said, it should be noted that just like knowledge production, which costs time, effort, materials, and money, publishing this knowledge also comes at substantial costs. JOTE is published by JOTE Publishers, and therefore we can control our own processing costs. Thanks to the relatively low costs of the open-source licenses and platforms we use, we are able to opt for the diamond route of Open Access: We publish and release all content for free, for everyone. This does not mean that there are no costs or that they are negligible. Rather, we can support these costs with generous funding from the Descartes Centre for the History and Philosophy of the Sciences and the Humanities at Utrecht University. We are grateful for their support and their commitment to developing innovative, social, and ethical venues for academic publishing. We would also like to emphasize that our decision to remove all article-processing and publication fees is only made possible by the fact that a dozen (under)graduate students have, over two years, consistently dedicated several hours a week to making JOTE a reality. Publishing, even in an online format, has material costs that should not be forgotten.

Who is our audience?
Considering our unorthodox approach to scientific publishing, our readers might very well ask themselves: ‘Who is this journal for, really?’ To preempt this question, we distinguish between two types of audiences we hope to accommodate with our publications.

On the one hand, we hope that JOTE may serve as a repository for null and mixed statistical results. In this sense, we believe our content appeals specifically to other researchers, meta-analysts, and scientific reviewers. To them, we offer content that enriches their estimates of effects and that deepens their understanding of contextual variability as well as the real rate of replicability of scientific phenomena. On the other hand, we also challenge philosophers of science, sociologists, anthropologists, and humanities scholars to join the discussion about science in practice with actual researchers. In this regard, we appeal to humanities scholars who wish to gain a better understanding of what scientists are actually doing. As we have argued, a survey of mainstream science alone paints a biased picture of the true state of current research. We are confident that there is far more failure in research than is currently reported.

In this sense, we hope that our empirical results and the associated reflection articles both help paint a clearer picture of the scientific process as it is practiced for scholars and scientists alike.

It is unfortunately common for the two groups mentioned above to be thought of as starkly divided. Though we recognize the different interests of these groups, we hope that we can also contribute to the dissolution of this divide by creating a space for critical dialogue between scholars from all backgrounds.

What is our ultimate goal?
Throughout this introductory piece, we have outlined JOTE’s proximal goals. We seek to provide an outlet for academic failure, foster interdisciplinary discussion about failure in science, inform meta-analyses, and close the gap between what is researched and what is published. Above and beyond these short-term goals, JOTE also has loftier aspirations. We discuss these long-term goals in this section.

Broadly speaking, our objective is to rethink success and failure in research. We wish to formalize a body of knowledge that has only been shared at water coolers in academic hallways so far. We want to ask, answer, and problematize the question of “What went wrong?” In doing so, we hope to normalize failure in the sciences such that all too common discussions of positive publication bias, file drawer problems, questionable research practices and/or p-hacking are rendered moot.

That said, our ultimate goal is to make JOTE obsolete. JOTE can only exist in a scientific climate in which trial and error is systematically ignored - it would not make sense to have a journal of trial and success. Therefore, our goal is to change the scientific landscape, highlight the importance of failure in science, uproot the way science is published, and to ultimately become unnecessary.
Conclusion
With the publication of this first issue, we introduce the Journal of Trial and Error, a journal dedicated to highlighting the importance of trial and error in scientific practice and scholarship. To address what we view as key problems in science today we implement a unique publication process that promotes scientific integrity, honesty, openness, and interdisciplinarity. Our goal in doing so is to change the scientific landscape, foster good-quality meta-analytic thought, reduce impact-or-perish pressures in academia, and revise the way science is published. If we succeed, all the better. If we fail, we hope that our failure can at least be instructive to others.

We thank the readers for taking an interest in this first issue of the Journal of Trial and Error.

References

License
Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
Failure is an integral part of learning: the adage “If at first you don’t succeed, try, try, try again” was originally coined to encourage children to persevere with their school work (Palmer, 1840). In the same way that we expect people to fail during the learning process as they gain new skills, theoretically we accept that failure is part of the scientific process – the quickest way to disprove a hypothesis is for an experiment, which tests that hypothesis, to fail to give the expected result.

In principle then, academia should embrace failure: It is a community built on the twin pillars of education and research using the scientific method, with failure playing a vital role in both. In reality, though, those who pursue research careers are often people who are used to being the “smart” ones, those who did not experience much failure at school. In a competitive environment with failure playing a vital role in both. In reality, though, those who pursue research careers often do not talk about failure. This reluctance to discuss ‘failed’ research causes problems for research and for researchers. If we do not openly discuss research failures, we increase the likelihood that other research will repeat the same study and we learned something, even though we produced a result other than the one we were expecting. However, the researcher who wants to report that result faces pressures from multiple angles. First, interventions that do not achieve their aim are not necessarily zero harm and the researcher may be unwilling to publicly acknowledge the detrimental impacts of their work on the lives of purported ‘beneficiaries’ (Barrington et al., 2017). For those who are willing to make that step, there may be additional pressures from funders and project partners. Those who funded, developed, or implemented the intervention may have a vested interest in it succeeding, and disincentives for admitting it did not work. This could be the technology developer who stands to lose funding for the roll-out of a product, local political leaders who had staked their re-election on the success of a trial, or a funder who does not want to admit that funds were spent on the ‘wrong solutions’. Even if the researcher manages to convince all parties of the benefits of sharing their findings, they may still face barriers to publishing. Despite having done everything ‘right’, journals (particularly those with the highest impact) often prefer not to publish null results because those are not generally considered sufficiently novel. With all of this in mind, it is no surprise that for many, the effort put into writing up failures and dealing with the fallout is not worth the limited rewards.

This reluctance to discuss ‘failed’ research causes problems for research and problems for researchers. If we do not openly discuss research failures, we increase the likelihood that other research will repeat the same study and discover the same null results. Repeating studies costs time, money, and in some disciplines, runs the risk of costing lives. For researchers, particularly those early in their research careers, not discussing failures gives a false view of how science works, namely that every piece of research will be accompanied by a ‘eureka!’ moment. This in turn heightens the feeling of ‘imposter syndrome’ among researchers – the fear that they are not good enough to be an academic or a researcher, and that it is only a matter of time before others discover that they are a fraud because their hypotheses are often proven wrong.

With that in mind, it takes a degree of bravery to talk about failures in WASH. There are countless examples of interventions that turned out to be inappropriate, such as the PlayPump, a roundabout that was connected to a water pump so that children’s play could be harnessed to deliver clean water to
communities across Africa (http://www.playpumps.co.za/). The shortcomings of this solution are well-documented (Borland, 2011). The joy of roundabouts is that they spin freely, so the PlayPump was not as much fun to use as you might think, requiring constant effort to keep it moving. This meant that in many places, children did not play on the roundabout for as long as was needed to ensure an adequate water supply for the community. Instead, the task of pushing the roundabout fell to women collecting water, when to them a hand pump would have been much preferred and required considerably less physical effort. Often (and in a surprisingly common theme for the sector) communities were not consulted on the water problems they faced before the PlayPump was selected as a solution, so even when sufficient children played on the roundabout to pump water, the PlayPump did not address issues of water scarcity or water quality. The PlayPumps also raised ethical questions about the blurring of the line between work and play. Many of the installed PlayPumps have since been abandoned or retrofitted with hand pumps, and the original backers of the technology have admitted that their extensive plans for roll-out were overly ambitious. Although exact numbers are hard to find, it is generally accepted that ‘millions of dollars’ were spent on PlayPumps, each of which cost as much as four conventional wells with hand pumps (Chambers, 2009).

In some respects, wasting money is the least of concerns in the international development sector. In WASH, there are documented cases of sanitation behavior change interventions causing harm, with those who defecate in the open being punished by others throwing stones at them (Devine, 2009) or dumping feces in their kitchens (Chatterjee, 2011). It is of little surprise that some of these interventions result in negative impacts on physical health and social cohesion (Barrington et al., 2017). In these scenarios, the people that we are trying to help end up in a significantly worse situation than before we stepped in.

One of the challenges with WASH interventions is that water and sanitation form part of complex systems that include people, technology, and the environment. Too frequently, interventions focus on only one aspect of that system, leading to unintended consequences elsewhere. There are recorded cases where individuals have died or been injured from falling into pit latrines, because the vital role of maintenance has been neglected or overlooked. In 2014, a five-year-old boy drowned in a school pit latrine in South Africa that collapsed under him during his first week of school. In 2018, the incident was repeated with a five-year-old girl drowning in another school pit latrine. This led to the government insisting on the eradication of pit latrine toilets from South African schools, but with little detail of what would replace them in water scarce areas where waterborne sanitation was not an option (Fihlani, 2018). Though the government understood that this was a catastrophic failure of the system, they were uncertain how to quickly reshape the system to ensure a safer service in the future.

Clearly, talking about failure has its advantages. In WASH research and practice, repeating failed interventions can lead not only to money spent on inappropriate implementation, but can also harm the intended beneficiaries, in the very worst cases resulting in death. How can we morally justify not discussing our failed ideas if sharing them can save money and lives?

Sharing negative outcomes of research helps to ensure that mistakes are not repeated. Anyone who has undertaken health and safety training knows that for every death that happens, there are numerous injuries and many more ‘near miss’ events. The reason that health and safety officers are so insistent on ‘near miss’ reporting is that this provides valuable data to understand how to prevent more serious incidents from occurring. Research, particularly in fields where our work impacts on people’s everyday lives, needs to take more heed of that approach. Talking about failure should be a vital aspect of research transparency.

There are global campaigns to push for all medical trials to be registered in advance so that negative results are captured (Goldacre & Gray, 2016). Proponents point out that this allows researchers to look at all the evidence relating to a drug, rather than only the pieces of information that support the pharmaceutical industry’s desired results. Perhaps international development needs to take a similar approach.

Convincing researchers to share failures is only the first hurdle, however. Funders are powerful players in any research partnership, and whilst some funders do share projects that do not turn out as planned, it is not common. One of the advocates for funders to share negative outcomes as well as positive outcomes, so that lessons can be learned and carried forward, is Elhra. In 2017, they funded Action Against Hunger Spain and the London School for Hygiene and Tropical Medicine to test if Moringa leaves could be used as an alternative to soap in Ghana, following successful laboratory trials. These new tests showed that, rather than acting as a disinfectant, when mixed with water the leaves provided the perfect breeding ground for E. coli. These results are easily accessible on the Elhra website (Action Against Hunger Spain, 2018). Whilst Elhra offers one positive example of how funders can share negative outcomes, there are many more organizations with much larger research budgets that informally prefer negative results to be quietly left out of reporting. As such, there is still more to be done in identifying convincing arguments for funders to openly share failures to promote learning.

Yet, as stated above, the bias associated with which results are published and which are not does not lie solely with the researchers and project partners. Journals want to publish novel research and failure is too often not seen as novel. This culture of success is damaging to research and is highly unscientific given that the scientific method explicitly makes room for failure. The drive towards discussing failures needs to come from journals as well, not just from researchers. Open Science pushes for all publicly funded research to be made publicly available, and that should hold true even when research gives null results or leads to failure. If this were the case, researchers would not be pushed to dredge data for “significant results”; publishing null results would be accepted practice. Data dredging is seen as one of the main causes of non-reproducible results and so opening up academic publishing to failures and null results would support the move towards more reproducible and more transparent science.

If we know what fails, we have a better chance of working out why so that we can succeed in future. Failure, particularly in complex systems such as those encountered in international development, is a vital part of the puzzle for achieving impactful research; the consequences of attempting to solve the puzzle without all the pieces can be extraordinarily high. With this understanding of the importance of publishing failures, we hope that this will one day be the norm. In the meantime, we encourage all researchers to find ways to
talk more openly about failure, within projects and to a wider audience. As the founders of ‘WASH Failures’, we have used game show style events at conferences to give people space to talk about failures in their work and are carrying out research to understand how organizational culture can impact how failure is communicated in water and sanitation organizations working in sub-Saharan Africa. We need more structured ways to report on failures – at conferences, in academic writing, in reports to funders, and with the general public. We were pleased to see that the Journal of Trial and Error is one of those nascent outlets for failure. We certainly hope that it will not be the last.

References

License
Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
Alcohol Cues and their Effects on Sexually Aggressive Thoughts
Julie Leboeuf1, Stine Linden-Andersen1, Jonathan Carriere1

1Department of Psychology, Bishop’s University, Sherbrooke, Quebec, J1M 1Z7, Canada

Correspondence: Julie Leboeuf, Department of Psychology, Bishop’s University, Sherbrooke, Quebec, J1M 1Z7, Canada

Email: jleboeuf16@ubishops.ca

Companion Piece: This paper has an accompanying reflection piece, Abma (2020). Read the abstract at the end of this article.

Abstract
Alcohol and its effects on aggression have been the subject of many discussions and research papers. Despite this fact, there is still a debate surrounding what it is exactly about alcohol that causes aggression. The current study sought to replicate the past finding by Bartholow & Heinz (2006), that alcohol cues without consumption increase the accessibility of aggressive thoughts, which can then influence aggressive behaviors. In the present study, participants had to complete a lexical decision task that was set up to assess whether aggressive words were detected faster in the presence of alcohol-related pictures compared to neutral pictures. The results of this study did not replicate the expected finding as only a main effect of word type was found in which participants detected neutral words faster than aggressive words. Furthermore, the study aimed to assess the role of gender stereotype acceptance levels in this association, but due to faulty design considerations, such analyses were not possible. The results are discussed in terms of the limitations of the study, and propositions for future directions are addressed.

KEYWORDS
alcohol, automatic cognition, gender stereotypes, lexical decision task, sexual aggression

Take Home Message
The present study failed to replicate the finding by Bartholow and Heinz (2006) that alcohol cues without consumption increase the accessibility of aggressive thoughts. Although the authors of the 2006 study mostly focused on aggressive thoughts of a physical nature, this study aimed to expand the results to sexually aggressive thoughts. Furthermore, the potential role of gender stereotype endorsement in the association between alcohol cues and aggressive thoughts was explored. The failure to replicate can be partly explained by a flawed sample: the small sample size reduced statistical power, and the different composition of the sample compared to the original study impacted the ability to recreate the results. Furthermore, some important methodological limitations concerning the choices of stimuli, particularly with regards to sexually aggressive words and neutral pictures, seriously reduced the quality of the material. Overall, poor design considerations most likely account for the failure to replicate. Alternatively, it is possible that sexually aggressive thoughts are processed differently than physically aggressive thoughts, although more research needs to be done to determine whether this distinction truly exists.

Introduction
According to the World Health Organization (2018), alcohol abuse results in three million deaths worldwide every year. Wells et al. (2000) found that of all the physical aggression incidents reported by the participants, 67.9% happened while someone involved in the confrontation had been drinking. Similarly, severe alcohol intoxication is believed to play a role in almost half of the sexual aggression cases worldwide (Testa, 2002). Furthermore, Field et al. (2004) found that among all the factors under investigation, what appeared to be the strongest predictor of intimate partner violence was the expectation of aggressive behavior following alcohol consumption. Although the pharmacological effects of alcohol have been well researched (Chermack and Taylor, 1995; Giancola, 2000; Heinz et al., 2011), there is less information available with regards to other factors, such as cognition and expectancies, that can lead to aggression in alcohol-related contexts. Bartholow and Heinz (2006) as well as Subra et al. (2010) researched whether simple exposure to alcohol-related cues unconsciously increases the availability of aggressive thoughts, thus increasing the possibility of subsequent aggressive behaviors. In both studies, the researchers found that participants made faster lexical decisions when aggression-related words were paired with alcohol-related pictures compared
with neutral primes. The first goal of the present study was to replicate the finding that alcohol cues increase the accessibility of aggressive thoughts and expand them to consider specific types of aggressive thoughts, such as those of a sexual nature. The second aim was to assess the potential role of gender stereotypes in the association between alcohol-cues and aggressive thoughts.

Before exploring more thoroughly the body of research pertaining to automatic aggressive cognition associated with a sheer exposure to alcohol cues, it is worthwhile to mention that similar studies have examined how other stimuli can generate aggressive thoughts. In 1998, Anderson, Benjamin Jr and Bartholow reported that simple identification of weapon primes was linked to an increase in aggressive thoughts. In this study, participants named aggressive words faster compared with nonaggressive words when presented with a weapon prime. The authors argued that this increase resulted from the weapon stimuli automatically priming aggression-related thoughts (Anderson et al., 1998). More specifically, Anderson and colleagues referred to the semantic network model of memory to support their claim. This theory posits that words or concepts that are similar in meaning or that repeatedly co-occur are activated simultaneously in the semantic memory and therefore develop strong associations. This model goes as far as to propose that this increase in aggressive thoughts subsequently increases the likelihood that these thoughts will affect behavior (Bartholow and Heinz, 2006). From this model, it would be hypothesized that two concepts associated together, such as weapons and aggression, could develop such a strong association that this association would then increase the probability of someone behaving aggressively.

While weapons have been reported to be linked with an increase in aggression-related thoughts, other elements from the environment that could similarly influence levels of aggressive thoughts, such as alcohol, should also be considered. Although it is generally accepted that alcohol increases aggression, there is still a debate as to what precisely causes or explains this increase (Bartholow and Heinz, 2006), and it is usually best explained through a combination of multiple theories and viewpoints (see Heinz et al. (2011) for a comprehensive review). Some leading theories include the physiological disinhibition hypothesis, the indirect cause hypothesis, and the expectancy hypothesis (Bushman, 2002). Briefly stated, the physiological disinhibition hypothesis states that alcohol intake increases the levels of aggression by anesthetizing the part of our brain that usually keeps our aggressive impulses under control, making people more likely to express aggressive behaviors (Bushman, 2002). Following a similar line of reasoning, the indirect cause explanation proposes that alcohol consumption might increase aggression tendencies by producing changes at the cognitive and emotional levels, such as by affecting intellectual functioning and reducing self-awareness, which then increases the likelihood of aggressive acts being committed (Bushman, 2002). As for the expectancy hypothesis, it holds that alcohol is linked with aggression because people expect it to be that way (Bushman, 2002). This presumed effect/expectancy hypothesis suggests that people tend to associate aggression and alcohol, even if only unconsciously, which potentially accounts for one of the ways in which alcohol intake is linked with aggression (Bartholow and Heinz, 2006). The problem with this hypothesis is that evidence supporting it mostly comes from placebo designs in which participants were led to believe they had consumed alcohol. Thus, it is unclear whether a belief that alcohol consumption has occurred is necessary for this unconscious association to be activated, or whether the presence of alcohol cues alone can increase the accessibility of aggressive thoughts (Bartholow and Heinz, 2006).

To address this methodological limitation, Bartholow and Heinz (2006) conducted a study in which they examined the extent to which alcohol cues without consumption or belief that alcohol had been consumed (i.e., placebo effect) could increase the accessibility of aggressive thoughts. They tested 121 undergraduate students and had them participate in a lexical decision task. The participants were first primed with a stimulus, were then shown a string of letters, and had to decide whether the letter string presented to them was a legitimate English word. The priming stimuli could either be alcohol-related pictures, weapon-related pictures, or neutral images, and the letter strings could either represent aggression-related words, neutral words, or nonwords. In this experiment, the neutral images consisted of plants. The weapon pictures were included to create a reference point by allowing comparison with the results showing a link between weapon exposure and aggressive thoughts obtained by Anderson et al. (1998). Results showed that participants identified aggression-related words faster when exposed to aggression-related pictures compared with neutral images, suggesting that alcohol and aggression and linked together in semantic memory.

In 2010, Subra and colleagues replicated the study by Bartholow and Heinz (2006) with a different population as well as a slight methodological alteration (Subra et al., 2010). First, their study was conducted with a French-speaking population as opposed to an English-speaking one, and the sample was not restricted to undergraduate students. Subra and colleagues used the same stimuli as those used by the authors of the 2006 study, with the exception that they selected different images for the neutral condition. They criticized the ones used previously by Bartholow and Heinz (2006) on the basis that their nature (i.e., plants) was not neutral since nature scenes have been shown to decrease aggressive thoughts (Kuo and Sullivan, 2001, Kuo and Sullivan, 2001). Instead, they used photos displaying non-alcoholic beverages for their neutral condition. Similarly to the study under replication, participants made faster lexical decisions about aggression-related words when exposed to alcohol or weapon primes compared with neutral primes, again supporting the idea that alcohol and aggression might hold strong associations in semantic memory.

The two studies by Bartholow and Heinz (2006) and Subra et al. (2010) suggest that exposure to alcohol cues without consumption is linked with an increase in aggression-related thoughts. Interestingly, the target words that were used for the aggression category were mainly of a physical nature (e.g., punch, assault, murder) and did not specifically assess sexual violence. It would be worth investigating whether this increase in thoughts of an aggressive nature also extends to sexual violence, especially considering that the World Health Organization (n.d.) listed the use of alcohol or drugs as a factor that increases the risk of men committing rape. Additionally, past research has shown that alcohol priming without consumption can increase sexual expectancies. Friedman et al. (2005) reported that men who were exposed to subliminal alcohol-related words rated women as being more sexually attractive, and that this effect was more precisely caused by the sexual expectancies associated with alcohol intake. Importantly, the belief that alcohol has been consumed has been linked with an increase in sexual arousal. More precisely, Abbey et al. (2001) reported...
that men who were led to believe they had consumed alcohol, regardless of whether this was actually the case, experienced more sexual arousal than men in the control group. This becomes problematic when considering the finding by another group of researchers, that induced sexual arousal increased participants’ expectations of their own sexual aggressiveness (Loewenstein et al., 1997). Overall, alcohol cues seem to be linked with increases in both sexual expectancies and sexual arousal, the latter being further associated with elevated expectations of sexual aggressiveness. Therefore, it is of paramount importance to better understand the relationship between alcohol cues and sexually aggressive thoughts.

Another element that was not assessed in the 2006 and 2010 studies is acceptance of gender stereotypes. In one study, Abbey (2002) reported that traditional gender role beliefs about dating and sexuality could be at play in explaining the occurrence of sexual assaults, irrespective of whether alcohol was involved or not. More precisely, endorsing beliefs such as interpreting a woman’s refusal as an invitation to be convinced, or that forced sex is sometimes acceptable, was a proposed factor that could increase the likelihood of sexual assailants being committed. While recognizing that this is a good first step into assessing the role of gender stereotypes acceptance with regards to sexual violence, more general gender role beliefs should be investigated to truly understand the role that gender stereotypes play in sexual violence occurrences. To this effect, Locke and Mahalik (2005) reported that men who showed a problematic use of alcohol and who conformed to such masculine norms as having power over women and being violent were more inclined to score higher on the Rape Myth Acceptance Scale and to report more sexually aggressive behaviors. The first study mentioned here only looked at gender stereotypes with regards to sexuality and dating (Abbey, 2002), and although the second study by Locke and Mahalik (2005) might seem to assess gender stereotypes more globally, it appears biased towards specific stereotypes with regards to power dynamics between men and women. In other words, the scales used in the above-mentioned studies seem to be assessing sexual harassment beliefs and attitudes more than gender stereotypes. Another problem with the literature on gender stereotype beliefs and its relationship to violence is that the vast majority of the articles available only report on the likelihood of men committing sexually violent behaviors, or on the relationship between men’s level of gender stereotypes and the perpetration of violent acts, thus leaving women out of the picture (Jakupec et al., 2002; Gidycz et al., 2007; Abbey et al., 2003). There is a clear need to investigate the relationship between gender stereotype beliefs and sexual violence as it relates not only to men, but to women as well.

Objective and Hypotheses

The objective of this project is to further the line of research on the effects of alcohol cues on aggressive behaviors by testing this relationship specifically with sexually aggressive words and by taking into account gender stereotype beliefs. More precisely, the aggressive words used by Bartholow and Heinz (2006) will be replaced by aggressive words of a sexual nature. The comparison point will be against the two studies that have investigated this before and for which significant results have been reported, namely the one by Bartholow and Heinz (2006) and the one by Subra et al. (2010). Gender stereotypes acceptance will be evaluated through the German Extended Personal Attributes Questionnaire (Runge et al., 1981).

Following from the findings suggesting that alcohol cues increase both sexual expectancies (Friedman et al., 2005) and aggressive thoughts (Bartholow and Heinz, 2006), it is hypothesized that participants will make faster lexical decisions to aggressive words of a sexual nature when paired with alcohol-related primes compared with neutral primes. A similar effect should also be found for the weapon-related primes. Furthermore, it is suggested that for both men and women this association will likely differ depending on one’s level of gender stereotypes acceptance. Indeed, results are expected to show a non-significant association at low levels of gender stereotypes beliefs, a moderate interaction at medium levels of gender stereotype beliefs, and a dramatically significant interaction at high levels of gender stereotype beliefs. Finally, for replication purposes, it is expected that participants will be slightly more accurate at identifying neutral words compared with aggressive words.

Methods

Participants

Sixty participants took part in this study, but two of them were excluded from the analyses, giving a final sample size of 58. One participant was excluded because the experiment failed before the data could be recorded, and the other was excluded because it was clear from the debriefing session that this participant had not understood the computer task properly. Furthermore, this participant’s accuracy rate was only 68.54% compared with a mean accuracy rate of 95.08% (SD = 4.62) for the sample. On this distribution, a score of 68.54% represents a z-score of -5.74, providing sufficient grounds for excluding the data from the analyses.

From the remaining participants, 49 self-identified as women (84.5%) and nine as men (15.5%). No one expressed a mismatch between sex at birth and gender self-identification, and no one selected the ‘other’ option for their self-identified gender. The age of the participants ranged between 18 years old and 46 years old (M = 21.64, SD = 7.71) and they were all enrolled at Bishop’s University. Please refer to Appendix A for a complete list of the different programs represented in this sample.

Stimuli and Task Questionnaires.

In this experiment two different questionnaires were used: a short demographic questionnaire (see Appendix B) and the German Extended Personal Attributes Questionnaire, a scale evaluating gender stereotype acceptance and beliefs (Runge et al., 1981). This questionnaire includes two subscales both comprising eight items, namely “expressivity” and “instrumentality”, which are intended to measure the degree to which someone can be classified according to stereotypically masculine (i.e., instrumentality subscale) or stereotypically feminine (i.e., expressivity subscale) adjectives. In its original form, this questionnaire was designed to assess self-ascribed masculinity or femininity, but for the purpose of this study, it was modified to assess one’s view and degree of endorsement of gender stereotypes in general (see Appendix C).
FIGURE 1 Alcohol-related picture examples (a, b), weapon-related picture examples (c, d), and neutral picture examples (e, f).

Task
Participants completed a lexical decision task in which they had to decide whether a string of letters presented to them was a legitimate English word. Prime stimuli consisted of 15 photos: five containing alcohol bottles, five portraying weapons, and five showing non-alcoholic beverages (see Figure 1 for sample pictures). Target words were also divided into three categories, each containing 15 words (see Appendix D for the complete list): aggression-related words of a sexual nature (e.g., grope, rape), neutral words (e.g., observe, vanish), and nonword letter strings (e.g., wenct, jork). Each photo was paired with 3 aggressive words, 3 non-aggressive words, and 3 nonword letter strings for a total of 135 trials. For each trial, an image was presented for 300ms, followed by a 200ms interval prior to the showing of the target word, which stayed on the screen until the participants responded or up to 3 seconds. Participants had to indicate by pressing on a key whether the letters formed a legitimate English word or not. An interval of 3 seconds separated each trial.

Procedures
Participants were asked to come to the Psychological Health and Well-Being lab on Bishop’s University campus for one session lasting between 30 minutes and 45 minutes. Following the procedure by Bartholow and Heinz (2006), partial disclosure was used in that participants were told that the goal of the experiment was to measure the speed of language comprehension in the presence of distracting information (i.e., pictures). Once they had agreed to participate in the study, participants were asked to complete two different paper questionnaires, including a short demographic questionnaire and a questionnaire pertaining to gender stereotype acceptance levels, as mentioned earlier. Next, participants were asked to complete the main task of the study, namely the computer-based lexical decision task described above. After completion of the lexical decision task, debriefing took place: participants were informed of the reasons justifying the use of partial disclosure, and a new consent form was presented to them.

Results
Following the procedure used by Bartholow and Heinz (2006), trials on which the participants’ response times (RTs) after the onset of the target word were smaller than 150ms or greater than 1,500ms were deleted and excluded from analyses (2.68% of all trials). Furthermore, response times to nonwords were not included in the analyses because they were only used in the study for methodological reasons and do not have any bearing on the present hypotheses being tested (Bartholow and Heinz, 2006). The data from the demographic questionnaire was not used in the analyses for different reasons. First, since there was no discrepancy between gender self-identification and sex at birth, no comparison was possible in this case. Next, given that there were only nine men in the study, the sample size was not large enough to run analyses based on gender differences. Finally, because every participant was enrolled at Bishop’s University, no comparison could have been done, and field of study was not used either considering that most participants were enrolled in psychology.

Gender Stereotypes Acceptance Levels
It is likely that the original questionnaire and the way in which it should be coded were manipulated to the extent that the results were uninterpretable. For this reason, the results of the German Extended Personal Attributes Questionnaire are not reported. The associated design-related issues are further explored in the Limitations section.1

Response Times
Only the correct-response trials were kept for the response times analyses. That is, the trials where participants misidentified a nonword for a word, and vice-versa, were excluded from the present analyses (5.18% of the remaining trials). The mean response time values did not show any skewness and did not need to be corrected through a log transformation, contrary to what the two groups of authors had previously found (Bartholow and Heinz, 2006; Subra et al., 2010). The RTs were analyzed using a 3 (prime type: weapon-related pictures, alcohol-related pictures, neutral pictures) x 2 (target word type: aggression-related words, neutral words) repeated measures analysis of variance (ANOVA), and the sphericity assumption was not violated in any of the analyses; therefore, the original statistics are reported. Main effects of prime type and target word type are presented in Table 1. The main effect of target word type was significant, $F(1, 55) = 148.589, p < .001$. More precisely, neutral words were detected faster ($M = 600$ms, $SE = 13$ms) than aggressive words ($M = 657$ms, $SE = 1$ To obtain the complete set of results, go to https://osf.io/hw9p/.
Whereas the previous research had found that aggressive words were detected faster than neutral words, all the other main effects were not statistically significant and neither were the interactions. The expected main effect of prime type was not significant \((F(2, 110) = 1.186, p = .309)\), and the interaction between prime type and target word type was also nonsignificant \((F(2, 110) = 0.025, p = .975)\). Those results are contrary to the previous studies where the authors had found both stated analyses to be significant (Bartholow and Heinz, 2006, Subra et al., 2010).

### Main Effects of Prime Type and Target Type

<table>
<thead>
<tr>
<th>Variable</th>
<th>(M) (ms)</th>
<th>(SE) (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weapon-related pictures</td>
<td>625</td>
<td>13</td>
</tr>
<tr>
<td>Alcohol-related pictures</td>
<td>628</td>
<td>13</td>
</tr>
<tr>
<td>Neutral pictures</td>
<td>633</td>
<td>13</td>
</tr>
<tr>
<td>Sexually aggressive words</td>
<td>657</td>
<td>13</td>
</tr>
<tr>
<td>Neutral words</td>
<td>600</td>
<td>13</td>
</tr>
</tbody>
</table>

**TABLE 1** Main Effects of Prime Type and Target Type

### Accuracy

The analyses performed for the accuracy levels are identical to those performed for reaction times, except that the trials on which the participants made a wrong decision were not excluded. Mean accuracy values did not show any kurtosis and did not need to be corrected through an arcsine transformation, contrary to the main study being replicated (Bartholow and Heinz, 2006). The overall accuracy level was 95.08\% \((SD = 4.62)\), but this variable was further analyzed and broken down through a 3 (prime type: weapon-related pictures, alcohol-related pictures, neutral pictures) x 2 (target word type: aggression-related words, neutral words) repeated measures ANOVA. Here again, the sphericity assumption was not violated. Replicating the results by Bartholow and Heinz (2006) and Subra et al. (2010), the main effect of target word type was significant, \(F(1, 55) = 45.591, p < .001\). More precisely, neutral words were detected with more accuracy \((M = .98, SE = .003)\) than aggressive words \((M = .919, SE = .010)\). All the other main effects were not statistically significant and neither were the interactions, which supports the hypothesis. The main effect of prime type was not significant \((F(2, 110) = 0.101, p = .904)\), and the interaction between prime type and target word type also did not reach the significance level \((F(2, 110) = 0.088, p = .916)\). Therefore, the results obtained in this study are not biased by a speed-accuracy trade-off.

### Discussion

Contrary to what Bartholow and Heinz (2006) and Subra et al. (2010) had previously reported, the results obtained in this study do not support the hypothesis that alcohol cues increase the accessibility of aggressive thoughts. Whereas the previous research had found that aggressive words were detected faster when preceded by alcohol or weapon pictures compared with neutral pictures, and that aggressive words were detected faster than neutral words, none of those results were replicated in the present study. Indeed, there was no significant difference in the speed of detection of aggressive words across the different types of pictures, and neutral words, as opposed to aggressive ones, were recognized faster by the participants. However, the minor finding by both groups of authors that neutral words were detected with more accuracy was replicated in this study. Finally, the unique hypothesis that was added to this study with regards to gender stereotype acceptance levels and their expected influence on response times could not be properly tested due to conceptual flaws. It is possible that major methodological and design-related limitations played a central role in this failure to replicate. Alternatively, on a conceptual level, there is perhaps a fundamental difference between “physically aggressive thoughts” and “sexually aggressive thoughts”, which could be reflected in the types of stimuli that elicit them. The methodological limitations and conceptual alternative will be addressed in turn.

### Limitations

The sample of participants in this study was problematic on many levels. First, when dealing with response times and differences in terms of milliseconds, it usually takes a large sample size to maximize statistical power. A sample size of 58 participants was probably not large enough to optimize statistical power, especially when compared to the 121 participants that were recruited by Bartholow and Heinz (2006). The main author of this study initially aimed at recruiting 120 participants to mirror the number of participants tested in the project under replication, but did not achieve this goal due to time constraints. A stronger conceptual limitation, however, is that no power analysis was done prior to testing to determine the ideal sample size. Therefore, it is difficult to know how many participants would have been needed to optimize the likelihood of detecting an effect. In the future, it would be important to conduct power analyses to determine the optimal sample size, and to ensure that enough participants are recruited in time to meet that standard.

The participants in this study also greatly differed compared with those recruited by Bartholow and Heinz (2006). Indeed, whereas they had recruited 60 men and 61 women, the present study included 49 women and 9 men. This difference in men-women proportion is problematic considering that past studies reporting a link between drinking and aggression have focused on the perpetration of aggressive acts by men, and not women (Jakupcak et al., 2002; Gidycz et al., 2007; Abbey et al., 2003; Locke and Mahalik, 2005). It is therefore plausible that a greater proportion of men in the current study would have led to different results. Thus, further replication attempts should ensure an equal representation of men and women in the sample.

A further limitation of this study pertains to the word choice for the aggressive target word category. Finding salient aggressive words of a sexual nature proved to be challenging for many reasons. One of them is that most of the words that could be found through internet searches were expressions made up of more than one word and could not be used as word length has an impact on reaction times (Bartholow and Heinz, 2006). Additionally, the line between sexually aggressive words and sexual preferences is somewhat blurry (e.g., sodomy, choke). Therefore, some of the words used might not
have been associated with violence for certain participants but instead with pleasure. Finally, the face value of some of the words in the aggressive category was doubtful, the best example being the use of the word ‘prey’. Thus, it is likely that the results of this study might have been negatively impacted by the flawed choices for the aggressive target words. Future studies attempting to understand the relationship between sexual violence and alcohol should revisit the current list of sexually aggressive words. For example, a manipulation check could be done to test the face value of the words before including them in the study.

The images used for the neutral primes and the alcohol-related primes bring another important limitation to the present study. Some of the images that were used contained words in them, such as ‘Corona’ being written on the beer bottles, or again ‘Nesquik’ as could be seen on a chocolate milk bottle. During the debriefing session, some participants reported finding it difficult and confusing to assess the letter strings when they were preceded by a photo on which there was some writing. Ensuring that the priming stimuli do not contain any confusing information such as writing is therefore an important consideration for future replications.

Another point that was brought forward by some participants is that the neutral pictures may have been more effective if they had not been beverages. The argument is that it becomes too obvious that the study is researching the effects of alcohol as it is contrasted with non-alcoholic beverages, despite the fact that weapon pictures are also included. As mentioned earlier, the reason why Subra et al. (2010) chose non-alcoholic beverages pictures was to address the limitation to Bartholow and Heinz’ study (Bartholow and Heinz, 2006). The conclusion from this is that neither plants nor non-alcoholic beverages are an optimal neutral category for this study and the ideal neutral prime has yet to be found. It would be recommended to conduct a norming study prior to the experiment to create a list of possible neutral visual stimuli. It is also possible that the issue does not reside in the nature of the neutral stimuli per se, but instead that not enough neutral stimuli were included, making the goal of the study obvious. Adding filler trials with random sets of pictures and words could help blind the participants to the purpose of the study.

Importantly, all these limitations do not offer a justification as to why neutral words were detected faster than aggressive words, contradicting the results obtained in both previous studies. A plausible explanation has to do with the timing of testing, which occurred close to the end of the academic semester at Bishop’s University. Indeed, ‘conventional wisdom’ among researchers is that the quality of participants most likely declines towards the end of the semester (Ebersole et al., 2016). The timing of testing could likely account not only for the low number of participants, but also for a poor quality of answers that further leads to poor results. However, Ebersole and colleagues (Ebersole et al., 2016) tested this hypothesis and found little evidence in its support. Although participants reported declining effort and attentional levels as well as an increase in stress as the semester unfolded, the researchers only found a weak and negligible effect of time of semester on task performance. Therefore, although it is plausible that testing participants towards the end of the semester had a negative impact on the quality of the results, it cannot fully explain the failure to replicate the main findings obtained by Bartholow and Heinz (2006). Nonetheless, a solution to this problem would be to test participants throughout the semester, and to include a broader sample outside of the university population so that time of semester cannot be a problematic variable.

As was made clear in the introduction, there is a great need to investigate the relationship between the endorsement of gender stereotypes and acts of sexual violence as it relates to both men and women. However, the execution of the hypothesis in the current study was done inadequately, rendering the results uninterpretable. The German Extended Personal Attributes Questionnaire was used because it was the one most closely related to the hypothesis under investigation that was available to the author at that time. However, this questionnaire was originally designed to measure self-ascribed masculinity or femininity, and was thus modified to instead assess a general endorsement of gender stereotypes. What is problematic, is that the way in which this was done was completely arbitrary. In the original form of the questionnaire, participants are presented with 16 items representing masculinity and femininity and have to indicate where they think they fall on the scale with regards to each item (Runge et al., 1981). In this study, participants were asked to indicate to what extent they believe that the 16 characteristics are representative of men in general, and were then asked to fill out the questionnaire a second time, but this time by indicating to what extent they believe that the said characteristics are generally representative of women. The validity of those changes was not tested prior to the main experiment and was not based on any empirical evidence. This manipulation was so great that a new method of coding was necessary, which was also based on arbitrary grounds. A score of 3 (middle of a 5-point semantic differential scale) was given a value of 0, scores of 2 or 4 were given a value of 1, and scores of 0 or 5 (extremes of the scale) were given a value of 2. There is no reason to believe that this coding method was valid. Furthermore, participants were divided into 3 equal groups using the visual binning option in SPSS, reflecting their level of gender stereotypes acceptance: low, medium, high. Again, the decision to divide participants into three equal groups instead of using pre-established cut-offs did not have empirical support. A further problem with this questionnaire was a lack of variability in the answers as most participants tended to ‘sit on the fence’, i.e., selecting a score of 3. Some items on the questionnaire are obvious at face value which may have made the participants answer in a socially desirable way, but there is no way of knowing whether social desirability was at play or if the results truly reflect the participants’ beliefs. Including a social desirability scale or other unrelated questions might have solved this problem, as well as using a 6-point Likert-type scale. Alternatively, the questionnaire could have been done online as to measure reaction time to stereotype-congruent and stereotype-incongruent adjectives. Nonetheless, the bigger question remains as to whether the questionnaire was a valid and reliable measure to begin with. Two major lessons can be learned from this. First, when designing a study, methodological decisions should be empirically-based as opposed to arbitrary. Second, it is critical to ensure the validity of the measures to be used before conducting the main experiment. Future studies looking at the relationship between gender stereotypes and sexual violence would greatly benefit from first designing a sound psychometric questionnaire assessing general levels of gender stereotype endorsement.

View interactive version [here].
Future Directions

Replication attempts should target a larger and more representative sample, much like Subra et al. (2010) did, to ensure that the results not only represent the student body accurately, but that they can also be generalized to the population as a whole. Although the goal of this experiment was to study aggressive behaviors in university students, it is still important to look at the bigger picture and remember that undergraduate samples do not represent the general population. What seems to be most problematic here is the lack of male representation and the overrepresentation of psychology students. The composition of this sample is not representative of the general university population, even less so of the overall population, and although it might not entirely explain the failure to replicate past findings, it is important to keep this issue in mind when designing subsequent research protocols. Understandably, a sample of 22-year-olds who mostly major in psychology might not be optimal when studying abusive relationships with alcohol.

Other than methodological considerations, attention should also be paid to conceptual alternatives. As mentioned earlier, “physically aggressive thoughts” and “sexually aggressive thoughts” might be fundamentally different. To this point, White et al. (2008) found that among over 10 correlates of physical and sexual aggression, the only two that distinguished perpetrators of physical aggression from perpetrators of sexual aggression were motives for sex. More precisely, the participants in the sexual aggression only group scored significantly higher than those in the physical aggression only group on the hedonism and dominance motives for sex (White et al., 2008). However, it is not clear from this study whether the participants had committed their aggressive acts under the influence of alcohol or not. Motive for sex is only one possible factor differentiating sexual aggression from physical aggression: multiple studies have also investigated the predictors of sexual and physical aggression and victimization (Zawacki et al., 2003; James and Young, 2013; Gidycz et al., 2007; Felson and Burchfield, 2004). However, there is no clear pattern emerging from the literature asserting whether there is a fundamental difference between physically and sexually aggressive thoughts, and if so, how it is characterized.

An interesting avenue to resolve this question would be to measure whether there are brain activity differences when participants are processing scenarios involving sexual compared with physical aggression.

Furthermore, future studies would benefit from inquiring about participants’ nationalities to explore whether the patterns of responses differ across countries. This idea emerged during the debriefing sessions as multiple participants reported coming from Europe and being raised with an open-minded attitude towards alcohol, further saying that sexual violence was not necessarily associated with alcohol for them. The study by Subra et al. (2010) took place in France, but it was investigating the link between alcohol and physical violence as opposed to sexual violence. Therefore, it would be important to study participants representing a range of different nationalities and look at whether differences emerge in the pattern of responses when it comes to alcohol and sexual violence. Taking it a step further, it would be even more informative to include both sexually aggressive words and physically aggressive words in the same study to explore more directly the possible differences in the relationships between alcohol priming and various types of aggressive thoughts across nationalities.

Another avenue that would be worth investigating is the relationship between aggressive thoughts and subsequent behavior. The studies by Bartholow and Heinz (2006) and Subra et al. (2010) point to the idea that alcohol-related cues can increase the accessibility of aggressive thoughts. How does this translate into behavior, if it does so at all? As discussed earlier, the semantic network model of memory posits that concepts that are similar in meaning or that co-occur are activated concurrently in the semantic memory and develop strong associations (Anderson et al., 1998). Therefore, when one concept is activated, it produces a spreading activation process through which other related concepts become more accessible, ultimately increasing the likelihood that those concepts will have an impact of subsequent behavior (Bartholow and Heinz, 2006). Bartholow and Heinz (2006) did follow up on their first experiment with a second one, not under replication here, which aimed at investigating the effects of this increase in aggressive thoughts. However, the results only showed that the increase was related to more aggressive interpretations of the behavior of others, and did not assess whether the participants would also be more likely to commit aggressive acts themselves (2010). Other studies have reported a link between the activation of a specific attitude and subsequent behavior. For example, Bargh et al. (1996) found that participants who were primed with an elderly stereotype took more time to walk out of the experiment center than did control participants. The idea was that their attitude towards older individuals (i.e., that they are slow walkers) influenced their behavior. Similarly, other researchers found that priming participants with the stereotype of ‘professor’ enhanced their performance on a general knowledge test, whereas priming the trait ‘stupid’ reduced their performance on said test (Dijksterhuis and Van Knippenberg, 1998). Although these studies have their own limitations, the results suggest that when an attitude toward an object is activated, it can influence behavior. The ultimate goal would be to establish a direct link between the activation of an attitude towards alcohol and aggressive thoughts, and a subsequent increase in aggressive behaviors. How to conduct this kind of experiment in an ethical manner, however, remains to be answered.

Conclusion

This current research project was undertaken to extend the literature by replicating the evidence that simple exposure to alcohol stimuli without actual consumption or belief that consumption has occurred can increase aggressive thoughts. The initial intention of this project was always to complete a replication of the first of two experiments by Bartholow and Heinz (2006) with new data, which implies keeping the same protocol as the original study while collecting new data (NWO, 2019). Dr. Bartholow generously shared the original target word stimuli and a description of the images he employed, allowing highly similar material to be used in this study. However, since different aggressive words were used and a variable (gender stereotypes acceptance levels) was added, it can be argued that the current research project would instead fall under the category of replication with the same research question, which is defined by new data collection and a new research protocol with the same research question (NWO, 2019). I disagree; the initial research protocol was followed closely and the research question remained unchanged. The added questionnaires can be regarded as an additional investigation that could be
part of a second experiment altogether; remove those questionnaires, and it becomes clear that the original research protocol was respected, even if the nature of the aggressive words was altered.

This replication attempt suffered from many methodological and design-related issues. For example, the sample size was most likely too small, although no proper power analysis was run, and the sample further lacked generalizability. The low number of participants might be explained by the timing of testing, but it remains unclear why so few men participated. Another important impediment to replication relates to the alteration of the aggressive target words since it is doubtful whether the new sexually aggressive words were of a similar salience as the original ones. In other words, the sexually aggressive words might not have been as centrally relevant to the concept of sexual aggression as were the aggressive words employed by Bartholow and Heinz (2006) to the notion of physical aggression. Similarly, the validity of the neutral category of pictures employed remains unclear. Finally, the execution of the hypothesis regarding gender stereotype levels was completely arbitrary and the results uninterpretable.

In conclusion, although the present study failed to replicate the previous finding that alcohol-related cues increase the accessibility of aggressive thoughts (Bartholow and Heinz, 2006; Subra et al., 2010), this does not mean that the effect is non-existent. Rather, it is probable that the methodology employed in this study was significantly flawed and reduced the likelihood of finding significant differences in the variables. This line of research should be further pursued as it bears significant importance on today’s society. If such results were to be replicated and empirically grounded, it would open new avenues for interventions aimed at decreasing the number of violent incidents that result from alcohol consumption. If cognitions related to alcohol truly make sexually aggressive thoughts more readily accessible, social and educational reforms could be undertaken to raise awareness and work on changing these automatic associations. Finally, if the endorsement of gender stereotypes proves to be a moderator in this association, more comprehensive interventions could be targeted at individuals who hold strong gender stereotypes.

References


Jakupcak, M., Lisak, D., & Roemer, L. (2002). The role of masculine ideology and masculine gender role stress in men’s perpetration of relationship vio-
Companion Piece

Abstract. Priming studies, mostly found in the subdiscipline of social psychology, have been the subject of vigorous debates among methodologists, philosophers of science, and priming researchers themselves. This article contributes to the debate about priming studies by carefully examining and dissecting one priming study in particular, namely the 2020 article "Alcohol Cues and their Effects on Sexually Aggressive Thoughts" by Julie Leboeuf, Stine Linden-Andersen, and Jonathan Carriere. By pointing out the flaws of this supposed reproduction study, I reflect on the various levels of complexity that are involved in conducting priming experiments with human subjects. I conclude that the call for more reproductions or replications is worthwhile, but only if the original experiments are solid and theoretically interesting.

License

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
Appendices

Appendix A

Field of study  \( n \)

<table>
<thead>
<tr>
<th>Field of study</th>
<th>( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Psychology</td>
<td>38</td>
</tr>
<tr>
<td>Biology</td>
<td>4</td>
</tr>
<tr>
<td>Languages</td>
<td>3</td>
</tr>
<tr>
<td>Education</td>
<td>2</td>
</tr>
<tr>
<td>Sociology</td>
<td>2</td>
</tr>
<tr>
<td>Physics</td>
<td>1</td>
</tr>
<tr>
<td>Politic Sciences</td>
<td>1</td>
</tr>
<tr>
<td>Biochemistry</td>
<td>1</td>
</tr>
<tr>
<td>History</td>
<td>1</td>
</tr>
<tr>
<td>English Literature</td>
<td>1</td>
</tr>
<tr>
<td>Mathematics</td>
<td>1</td>
</tr>
<tr>
<td>Finance</td>
<td>1</td>
</tr>
<tr>
<td>Arts Administrations</td>
<td>1</td>
</tr>
<tr>
<td>Sports Studies</td>
<td>1</td>
</tr>
</tbody>
</table>

**TABLE 2**  Field of study – Complete list

Appendix B

**Demographic Questionnaire**

1. Age: ____________________________
2. Sex at birth: ___________________ (F/M/other)
3. Gender: ________________________ (F/M/other)
4. University attended: ______________
5. Field(s) of study: ________________

**FIGURE 2**  Demographic Questionnaire

Appendix C

**German Extended Personal Attributes Questionnaire**

In general, I see **MEN** as being:

<table>
<thead>
<tr>
<th>Not independent</th>
<th>1 2 3 4 5</th>
<th>Very independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very passive</td>
<td>1 2 3 4 5</td>
<td>Very active</td>
</tr>
<tr>
<td>Not competitive</td>
<td>1 2 3 4 5</td>
<td>Very competitive</td>
</tr>
<tr>
<td>Decisive</td>
<td>1 2 3 4 5</td>
<td>Not decisive</td>
</tr>
<tr>
<td>Gives up easily</td>
<td>1 2 3 4 5</td>
<td>Never gives up</td>
</tr>
<tr>
<td>Not self-confident</td>
<td>1 2 3 4 5</td>
<td>Self-confident</td>
</tr>
<tr>
<td>Feels inferior</td>
<td>1 2 3 4 5</td>
<td>Feels superior</td>
</tr>
<tr>
<td>Doesn’t stand up under pressure</td>
<td>1 2 3 4 5</td>
<td>Stands up under pressure</td>
</tr>
<tr>
<td>Not emotional</td>
<td>1 2 3 4 5</td>
<td>Very emotional</td>
</tr>
<tr>
<td>Devotes self to others</td>
<td>1 2 3 4 5</td>
<td>Doesn’t devote self to others</td>
</tr>
<tr>
<td>Very rough</td>
<td>1 2 3 4 5</td>
<td>Very gentle</td>
</tr>
<tr>
<td>Not helpful</td>
<td>1 2 3 4 5</td>
<td>Very helpful</td>
</tr>
<tr>
<td>Very kind</td>
<td>1 2 3 4 5</td>
<td>Very kind</td>
</tr>
<tr>
<td>Not aware of feelings</td>
<td>1 2 3 4 5</td>
<td>Aware of feelings</td>
</tr>
<tr>
<td>Not understanding</td>
<td>1 2 3 4 5</td>
<td>Very understanding</td>
</tr>
<tr>
<td>Cold</td>
<td>1 2 3 4 5</td>
<td>Warm</td>
</tr>
</tbody>
</table>

In general, I see **WOMEN** as being:

<table>
<thead>
<tr>
<th>Not independent</th>
<th>1 2 3 4 5</th>
<th>Very independent</th>
</tr>
</thead>
<tbody>
<tr>
<td>Very passive</td>
<td>1 2 3 4 5</td>
<td>Very active</td>
</tr>
<tr>
<td>Not competitive</td>
<td>1 2 3 4 5</td>
<td>Very competitive</td>
</tr>
<tr>
<td>Decisive</td>
<td>1 2 3 4 5</td>
<td>Not decisive</td>
</tr>
<tr>
<td>Gives up easily</td>
<td>1 2 3 4 5</td>
<td>Never gives up</td>
</tr>
<tr>
<td>Not self-confident</td>
<td>1 2 3 4 5</td>
<td>Self-confident</td>
</tr>
<tr>
<td>Feels inferior</td>
<td>1 2 3 4 5</td>
<td>Feels superior</td>
</tr>
<tr>
<td>Doesn’t stand up under pressure</td>
<td>1 2 3 4 5</td>
<td>Stands up under pressure</td>
</tr>
<tr>
<td>Not emotional</td>
<td>1 2 3 4 5</td>
<td>Very emotional</td>
</tr>
<tr>
<td>Devotes self to others</td>
<td>1 2 3 4 5</td>
<td>Doesn’t devote self to others</td>
</tr>
<tr>
<td>Very rough</td>
<td>1 2 3 4 5</td>
<td>Very gentle</td>
</tr>
<tr>
<td>Not helpful</td>
<td>1 2 3 4 5</td>
<td>Very helpful</td>
</tr>
<tr>
<td>Very kind</td>
<td>1 2 3 4 5</td>
<td>Very kind</td>
</tr>
<tr>
<td>Not aware of feelings</td>
<td>1 2 3 4 5</td>
<td>Aware of feelings</td>
</tr>
<tr>
<td>Not understanding</td>
<td>1 2 3 4 5</td>
<td>Very understanding</td>
</tr>
<tr>
<td>Cold</td>
<td>1 2 3 4 5</td>
<td>Warm</td>
</tr>
<tr>
<td>Appendix D</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Neutral Words</th>
<th>Aggressive Words</th>
<th>Nonwords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glide</td>
<td>Touching</td>
<td>Sarf</td>
</tr>
<tr>
<td>Suggest</td>
<td>Incest</td>
<td>Philst</td>
</tr>
<tr>
<td>Observe</td>
<td>Coercion</td>
<td>Wenct</td>
</tr>
<tr>
<td>Vanish</td>
<td>Molest</td>
<td>Jork</td>
</tr>
<tr>
<td>Move</td>
<td>Sodomy</td>
<td>Atand</td>
</tr>
<tr>
<td>Narrate</td>
<td>Harassment</td>
<td>Kesh</td>
</tr>
<tr>
<td>Imagine</td>
<td>Grope</td>
<td>Morpe</td>
</tr>
<tr>
<td>Ignored</td>
<td>Rape</td>
<td>Farn</td>
</tr>
<tr>
<td>Improve</td>
<td>Prey</td>
<td>Lefr</td>
</tr>
<tr>
<td>Joked</td>
<td>Perversion</td>
<td>Chib</td>
</tr>
<tr>
<td>Read</td>
<td>Abuse</td>
<td>Dephps</td>
</tr>
<tr>
<td>Leave</td>
<td>Violated</td>
<td>Anage</td>
</tr>
<tr>
<td>Listen</td>
<td>Choke</td>
<td>Steay</td>
</tr>
<tr>
<td>Transfer</td>
<td>Stalking</td>
<td>Omoga</td>
</tr>
<tr>
<td>Reported</td>
<td>Roofies</td>
<td>Benct</td>
</tr>
</tbody>
</table>

**Table 3** Target Word Type – Complete List
Experiment and Fail: a Comment on “Alcohol Cues and their Effects on Sexually Aggressive Thoughts”

Ruud Abma¹

¹Descartes Centre for the History and Philosophy of the Sciences and the Humanities, Utrecht University, 3584 CH, the Netherlands

Correspondence: Ruud Abma, Descartes Centre for the History and Philosophy of the Sciences and the Humanities, Utrecht University, 3584 CH, the Netherlands

Email: r.abma@uu.nl

Companion Piece: This paper reflects on a previously published empirical article, Leboeuf et al., (2020). Read the abstract at the end of this article.

Abstract

Priming studies, mostly found in the subdiscipline of social psychology, have been the subject of vigorous debates among methodologists, philosophers of science, and priming researchers themselves. This article contributes to the debate about priming studies by carefully examining and dissecting one priming study in particular, namely the 2020 article “Alcohol Cues and their Effects on Sexually Aggressive Thoughts” by Julie Leboeuf, Stine Linden-Andersen, and Jonathan Carriere. By pointing out the flaws of this supposed reproduction study, I reflect on the various levels of complexity that are involved in conducting priming experiments with human subjects. I conclude that the call for more reproductions or replications is worthwhile, but only if the original experiments are solid and theoretically interesting.

KEYWORDS

human subjects, philosophical reflection, priming studies, replication, reproduction

Introduction

The paper “Alcohol Cues and their Effects on Sexually Aggressive Thoughts” (Leboeuf et al., 2020) reports a failed attempt at reproducing two experiments. The massive shortcomings of the reported reproduction are obvious. For a moment I was tempted to think that the authors, in the form of a standard psychological paper, were presenting a philosophical critique of this type of experiment. But they were not. In my comment I will try to formulate such a critique in a more straightforward manner.

I will first give a brief and plain description of what happened in the experiment and why it was done, according to the authors. Then I will say a few things about the complexity of producing and reproducing experiments in general, followed by a section on the problems of the specific type of experiments of which this one is a specimen: priming studies, mostly found in the subdiscipline of social psychology. For almost a decade, priming studies have been the subject of vigorous debate among methodologists, philosophers of science and priming researchers, in scientific journals, but also in newspapers, magazines, blogs and on Twitter. I will end with an assessment of the possibilities and the limits of doing experiments in the human sciences: what can we learn from experiments on alcohol cues if we want to tackle physical, mental and social harm, attributed to the consumption of alcohol?

The Experiment

This is how the experiment conducted by Leboeuf et al. (2020) went. Sixty people, all students, volunteered to come to the Psychological Health and Well-Being lab on Bishop’s University campus (Sherbrooke, Quebec). Upon arrival they are instructed to take part in a word recognition task: how quickly and accurately can they decide whether a string of letters presented to them on a computer screen is a legitimate English word? In total, they are presented with 45 letter strings, of which 15 are neutral, 15 are “nonwords” and 15 are “aggression-related words of a sexual nature” (Leboeuf et al., 2020, p. 10). Prior to the presentation of a target word, a photo is shown of a weapon, an alcoholic drink, or a non-alcoholic drink. Participants have to indicate by pressing on a key whether the letters in their judgment represent a legitimate English word or not.

¹. Interestingly, one of the photos shows bottles of the beer brand Corona, which would nowadays have a totally different association and might possibly ruin the experimental setup. Also, for the association to occur subjects would have to know that Corona is a beer; this is not clear from the photo.
An elaborate and very detailed and precise list of actions worded in common
were detected slower than non-alcohol cues. So, was the alcohol-aggression
(language is needed for this layperson to successfully execute
(1996) by the Dutch philosopher of science and technology, Hans Radder:

(Leboeuf et al., 2020, p. 16): “This replication attempt suffered from many
methodological and design-related issues.”

From their account it becomes clear how complicated reproducing a seem-
ingly straightforward experiment actually is. While an experienced experi-
menter might shake their head at such an imperfect specimen of experimental
research, in my view it is a very instructive case, precisely because of its
faults. It is a specimen of “sloppy science”. Normally when authors want to
publish a paper that is based on weak research, they cover up shortcomings
using methodological decisions and statistical manipulations, but these authors
refrained from such a procedure.

The Intricacies of Producing and Reproducing Experiments
Performing an experiment is quite a complicated task. Apart from a theoretical
description of the required manipulations, the object and the apparatus, there
is the material realization of the actual experiment, which necessitates careful
preparation. The complexity of experimenting in general can be illustrated by
the following example, which is taken from the book In and about the world
(1996) by the Dutch philosopher of science and technology, Hans Radder:

Consider [...] an experiment for determining the boiling point
of a particular liquid. This liquid is our object under study. Our
apparatus consists of a heat source, a vessel, a thermometer,
and possibly some supplementary equipment. On the basis of
our knowledge of the interaction process between thermometer
and liquid, we assume that our readings of the thermometer
inform us about the temperature of the liquid. Part of the
preparation procedure involves making sure that the liquid in
question is pure. This is why it may be necessary first to clean
the vessel that will contain the liquid. (Radder, 1996, p. 11)

Besides guiding the preparation of the object and the necessary equipment,
the theoretical description informs us about the staging of the processes of
interaction between object and equipment and the processes of detection (i.e.
measuring). Finally, the experimental system should be “closed”, which means
that potential disturbances from the outside should be identified and controlled;
this is also part of the theoretical description (Radder, 1996).

This all sounds rather straightforward and self-evident, but experimental pro-
cedures are full of hidden presuppositions, as becomes clear when a researcher
is given the task to instruct a layperson how to perform a certain experiment.
An elaborate and very detailed and precise list of actions worded in common
(non-theoretical) language is needed for this layperson to successfully execute
the tasks involved. It moreover requires that the researcher already knows how
to perform the experiment (Collins, 1985).³

The notions of theoretical description and material realization are both
relevant and helpful to analyze the issues at stake with the reproduction of ex-
Type 1 is the reproducibility of the material realization of an experiment, which
means: (a) it is not dependent on any particular theoretical description, (b) it
can be done by laypersons. In type 2, an experiment may be reproduced under
an identical theoretical interpretation, which allows for slight variations in the
material procedures. Type 3 concerns the reproducibility of the result of an
experiment, which implies that it is possible to obtain the same experimental
result while performing – theoretically and materially – different procedures;
this is, in Radder’s terminology, a reproduction of the original experiment. In
contrast to type 3, type 1 and 2 require a reproduction of the whole of the
experimental process.

In addition to these three varieties of reproduction type, Radder (1996)
distinguishes four possible types of actors in the reproduction process, or four
ranges of reproduction: (1) reproducibility by any scientist or even any human
being in the past, present or future, (2) reproducibility by contemporary sci-
entists; (3) reproducibility by the original experimenter, and (4) reproducibility
by the lay performers of the experiment. Types and ranges combined, there are
thus twelve possible categories in the field of reproduction, which allows a far
more sophisticated assessment and categorization than the usual differentiation
between “direct” (or exact) and “conceptual” reproductions (see below), or the
categorization by the Dutch research funder NWO (Dutch Research Council):
(1) replication with existing data, (2) replication with new data (and the same
research protocol), (3) replication with the same research question, but with a
different research protocol and new data (NWO, 2019).

Is “Alcohol Cues and their Effects on Sexually Aggressive Thoughts” a reproduction?
In what category can we now place the word decision experiment we are dis-
cussing here? Firstly, considering Radder’s aspect of range (Radder, 1996),
it is a reproduction performed by (more or less) contemporary scientists; the
timespan between the original and the reproduction amounts to 14 years. Sec-
ondly, considering type, we learn from the description of the experiment that
the researchers initially aimed at reproducing the original experiment, using
the same protocol: “Dr. Bartholow generously shared the original target word
stimuli and a description of the images he employed, allowing highly similar
material to be used in this study” (Leboeuf et al., 2020, p. 15). In that case, we
would have a reproduction under a fixed theoretical interpretation, i.e. type 2.

However, the reproducers also wanted to study the accessibility of sexually
aggressive thoughts and therefore decided to change the sets of target words and
images, in order to accommodate the addition of a new variable. Nevertheless,
they themselves considered their experiment to be a true reproduction: “the ini-
tial research protocol was followed closely and the research question remained

⁵. This may also mean: disturbances by internal processes, for instance affective and emo-
tional processes in human beings.
³. The intricacies of learning how to do a proper experiment (and to perform reproductions)
are described and analyzed in Collins (1985).
unchanged [...] even if the nature of the aggressive words [and the photos! RA] was altered" (Leboeuf et al., 2020, p. 16). This claim is not irrelevant since, in order to be recognized (and funded) as a proper replication, submitted proposals in this category usually have to conform to strict definitions issued by funding agencies.

In my opinion, the claim made by the authors in this respect is debatable or even false: changing target words and images means bringing about a change in the “apparatus” used, which also implies changes in the “interaction” with the object and probably also in measurement procedures. One way to establish whether the research protocol is really “the same”, as the reproducers claim, would be to explicate in common language the detailed instructions for the material realization of both the original experiment and the reproduction. This is hardly ever done; usually, and also in this case, the method section in journal articles does not give the reader (and the reproducer-to-be) sufficient information about the actual proceedings to create a “lay persons instruction”. It would require getting the protocol from the original experimenters, and even then more detailed information might be necessary (Collins, 1985).

For now, I hold that the reproduction experiment discussed here is at best a replication (cf. Radder, 1996), i.e. an attempt at attaining the same experimental result while performing – theoretically and materially – different procedures. In itself, this could be valuable: the significance of a result is stronger when it can be obtained under different experimental processes or, as Radder (1996, p. 84) puts it: “Abstraction through replication enables us to systematically conceptualize experimental results arising in essentially different situations. As such it constitutes an important step towards theory formation.” Because the reproducers introduced a new variable in their study (“sexually aggressive” instead of “aggressive”), one could argue that this is not even a replication but a new experiment.

### Priming Studies

The reproduction experiment under discussion is a so-called priming experiment, which means that a stimulus is presented that is supposed to subconsciously influence the subjects in the experiment in a systematic way, as measured by their results on a specific task (for an overview of the field of priming studies, see Derksen, 2017). In this case the prime consists of photos of a different nature (weapon, alcohol, non-alcohol) and the specific task is word recognition. According to the “semantic network model of memory”, an individual primed by for instance the image of a beer bottle would be prone to choose a sexually aggressive word faster than a neutral one as a “legitimate English word” (Leboeuf et al., 2020, p. 10).

Since a quarter of a century, this type of experiment has become very popular in social psychology. For the subdiscipline as a whole, it is an attractive type of study because it puts a counterintuitive and (for some) controversial idea center-stage: in human decision making, volition or free will is far less important than is usually thought; instead, people take many – if not most – decisions automatically, influenced subconsciously by environmental factors. For individual researchers, engaging in the priming tradition opens up a variety of topics to study experimentally, and a possibility to share in an almost unlimited market of publication opportunities. Presenting results that are at odds with common-sense thinking is considered an asset (Strack, 2012).

From 2011 onwards, however, a fundamental debate has started about the quality of the ever expanding field of priming research, leading to a “crisis of confidence” in experimental social psychology, or at least in the area of priming research. The main allegation was that within social psychology there was an abundance of “sloppy science”. Researchers were accused of having “photoshopped” their raw data by methodological and statistical manipulations or, as a group of methodologists put it: the field of (social) psychology “currently uses methodological and statistical strategies that are too weak, too malleable, and offer far too many opportunities for researchers to befuddle themselves and their peers” (Derksen, 2017, p. 178). There is in (social) psychology an overly enthusiastic use of “researcher degrees of freedom”, which enables researchers to obtain almost every result they want (Pashler & Wagenmakers, 2012).

Adding fuel to the upheaval were attempts to reproduce “classic experiments” in social psychology, such as Bargh’s study on the effect of subjects being primed by word references to old age, who would afterwards walk slower toward the exit of the building where the experiment was conducted (as elderly people are supposed to do) (Bargh et al., 1996; Derksen, 2017, p. 183-188). The reproducers followed Bargh’s protocol as precisely as possible, but nonetheless were not able to produce the same results. Fearing that this outcome would damage “his life’s work”, Bargh attributed the reproduction failure to the incompetence of the reproducers (Yong, 2012a, 2012b). Several years later, in a more general attempt to do something about the “replicability crisis” and validate psychological research, a massive Reproducibility Project was conducted, leading to a shocking result: no more than one-third of experimental results could be replicated (Open Science Collaboration, 2015).

This ignited a debate on the value of doing what was called “exact (or direct) replications” (following the same protocol) versus so-called “conceptual replications”, in which researchers with the same theoretical background as the original experimenters, use more or less different operationalizations to produce similar effects, in order to strengthen and extend the theory at stake. Whereas methodologists seemed to favor “direct” reproductions, social psychologists considered them of little value: “In psychological research, there are always a multitude of potential causes for the failure to replicate a particular research finding” (Stroebe et al., 2012, as cited in Derksen, 2017, p. 183).

Not surprisingly, these causes mostly had to do with the complexity of experimental work; either the protocol was not precise enough or the reproducer lacked the necessary skills; or there were mediating variables at work, that were thus far unknown. According to Bargh et al. (1996), priming studies in particular are very sensitive to this, which is why priming experiments require precise control and great skill from the experimenters (Derksen, 2017). This line of reasoning is at least half a century old. Already in 1968, two respected experimentalists in social psychology, Elliott Aronson and Merrill Carlsmith, wrote:

> [...] when an attempted replication fails, one must interpret this failure with caution because it is difficult to draw firm inferences. The most we can say is that there was something about the original experiment which was not accurately specified and which seemed to have had an important effect on the

4. See Derksen, 2017, Ch. 9
In the word decision experiment discussed in this commentary, similar issues are at stake. In fact, they all pointed to one and the same methodological problem: the endeavor to ward off disturbing influences from the experimental situation or, in other words, to “keep the system closed”. Although this “closedness” must be qualified – you only need to control those influences that are relevant in view of the problem and the aim of the experiment - it implies careful and systematic action: you “have to produce and maintain them through active intervention” (Radder, 1996, p. 122).

The fact that we’re dealing with interventions by the experimenter might also imply that we are creating an unnatural situation that might diminish the external validity of the results. Collier (2005) recounts a study in which a group of young men is isolated and manipulated to establish how power relations between the participants develop. His question is: what conclusions can you draw about what people tend to do outside the lab? This is his answer:

The men were removed from their families and friends, their jobs and normal leisure activities, even the common light of day. Now of course, all experiments are artificial, but in a good experiment the artifice removes the effects of irrelevant variables on the matter being tested. In this case, it does not remove but introduces such variables, just as putting animals in cages makes it impossible to study their natural behavior. [...] The proper way of studying the effects of power changes and power vacuums on humans would be by studying human behavior in the wild [...] in the open system of history, for instance the history of the French Revolution. (Collier, 2005, p. 332)

In the human sciences the experiment constitutes a miniature social system, in which the subjects are not passive, but respond actively and intelligently to all that goes on in the experimental situation. Using the information that they can get participants will try to guess what the aim of the experiment is and attune their behavior accordingly (demand characteristics). Probably the most important source of information are the experimenters themselves, who might unwillingly influence the responses of the subjects (experimenter bias).

For a good reproduction of an experiment, it is important to know how the experimenter operated, whether they were a man or a woman, what the precise wording was of the instruction to the participants, what knowledge the subjects might have or obtain about the goal and hypotheses of the study, and to what degree deception of the subjects was involved in the experimental setup. As noted, however, this type of detailed information is hardly ever reported in journal articles, which hampers the attempts of reproducers to establish what precautions are taken to minimize experimenter bias or gauge the influence of demand characteristics (Klein et al., 2012).

Experimenting with Humans

Given the complexity of procedures and the insecurity of results, why would human scientists perform experiments? Experimenting with human subjects seems to be far more difficult than with physical objects. This is ironic, because originally it was precisely the association with the rigor of the natural sciences that inspired psychologists in the late nineteenth century to adopt the experimental method as their favorite modus operandi (Danziger, 1990).

Nowadays the usual argument is: “by doing experiments you are able to establish causal connections” . This takes the form of “if p, then q”, which means there is a “constant conjunction” between p and q, where q is “caused” by p (Manicas, 2006). In terms of our experiment: if we present the image of a beer bottle to the subject, they will choose an sexual-aggressive word faster than an alternative.

But what if this does not happen? Should we reject this specific regularity and try to come up with a better one? Not necessarily. There might be one or more methodological flaws in the reproduction attempt, for instance, as in the experiment discussed here. “A sample size [...] was probably not large enough to optimize statistical power”; or “the results of this study might have been negatively impacted by the flawed choices for the aggressive target words”; and “the validity of the neutral category of pictures employed remains unclear” (Leboeuf et al., 2020, respectively p. 13, p. 14, and p. 16). This points to a recurrent problem in experimental social psychology. Already in 1954 Leon Festinger wrote that “ [...] negative results perhaps reveal only the fact that the experiment was not set up carefully and that the experimenter’s attempted manipulation of the variables was ineffective” (Festinger, 1954, p. 143).

In addition, there could be a misguided preconception in the research question, for instance in the idea that alcohol cues are automatically linked to aggressive thoughts, and not to “feeling good” or “having fun with friends”. An indication for this is given in the discussion section, where multiple participants said that “sexual violence was not necessarily associated with alcohol for them” (Leboeuf et al., 2020, p. 15). The authors suggest that cultural differences might be at stake here: multiple participants coming from Europe stated that they were “raised with an open-minded attitude towards alcohol” (Leboeuf et al., 2020, p. 15).

Another shortcoming might be caused by a central feature of priming experiments: they depend on deception. Deceiving subjects about what is the goal of the investigation has been an important instrument for experimental social psychologists since the 1950s to keep the experimental system closed. But do we really know whether we succeed in deceiving our participants? In the case discussed here, one subject was removed from the sample “because it was clear from the debriefing session that this participant had not understood the computer task properly” (Leboeuf et al., 2020, p. 11). This indicates that the task may be interpreted in various ways; how can we be sure that the other participants did not have their own, though maybe less interfering, interpretation of the task? They might even have guessed what the experiment was actually about. This is not far-fetched because the authors themselves admit that, on seeing images of alcoholic and non-alcoholic beverages, some participants suspected that the study was “researching the effects of alcohol as it is contrasted with non-alcoholic beverages” (Leboeuf et al., 2020, p. 14).

5. Manicas argues convincingly that causality requires more than a “constant conjunction”. 
That guess is close enough to open the experimental system to a confounding variable.

These examples of possibly confounding interpretations by experimental subjects point to a fundamental issue: participants will have their own interpretation of the nature and goal of the experiment, and this interpretation may influence their responses in a way that is not intended by the experimenter. This is usually referred to as the problem of the “double hermeneutic”: researchers have to be aware of both their own interpretation of what’s going on in their research and the interpretation that the subjects have of the experiment they are participating in. If researchers fail to assess properly what their subjects think, their interpretation of the results might be seriously flawed. Here we have a fundamental difference between the natural and the human sciences: the objects of study in natural science disciplines do not interpret the experimental interventions they are subjected to (Radder, 2019). In conclusion, experiments may not be the best way to study people.

Does this mean we have to do away with the idea of causality? Yes, if that means sticking to the principle of constant conjunction of isolated variables. No, if causality implies taking into account the characteristics of the object of study, the generative powers that are typical of human beings, for instance their judgmental sophistication, their use of written language, etc. Yes, people do have automatic responses to situations, but they also have the ability to consciously assess what’s going on and choose their course of action. The range of options is not unlimited, but on the other hand it is not possible to predict the outcome with certainty.

Relevance

The importance of this insight can be illustrated if we return to the experiment at hand. Why would we do research into the effects of alcohol use in human beings? Because “alcohol abuse results in three million deaths worldwide every year” (Leboeuf et al., 2020, p. 9). How do these deaths come about? The authors suggest that “intimate partner violence” is among the primary causes, but this is not explicitly stated. Even if partner violence would not be responsible for all of these three million deaths, an investigation into its causes would be most important. The main issue would obviously be: how can we prevent partner violence? And yes, alcohol abuse can be an important causal or facilitating factor, so we would also like to know: how can we prevent alcohol abuse?

Instead, the researchers deflect from these core issues and enter into a different debate: what is it, regarding alcohol, that leads to or increases aggression? That it does is “generally accepted”, but “there is still a debate as to what precisely causes or explains this increase”. Not surprisingly, “it is usually best explained through a combination of multiple theories and viewpoints” (Leboeuf et al., 2020, p. 10). That makes sense, but instead of informing the reader about this combined theory, the authors give a brief overview of rivaling theories (or hypotheses). For instance, drinking alcohol increases aggression, “by anesthetizing the part of our brain that usually keeps our aggressive impulses under control” (Leboeuf et al., 2020, p. 10). The outcome is of course that people are “more likely to express aggressive behaviors” (Leboeuf et al., 2020, p. 10). For some researchers, this is too straightforward, and they propose that alcohol consumption increases aggression “by affecting intellectual functioning and reducing self-awareness” (Leboeuf et al., 2020, p. 10). Finally, it might be “that people tend to associate aggression and alcohol, even if only unconsciously” (Leboeuf et al., 2020, p. 10), which would increase the likelihood of people behaving in an aggressive manner when they have been drinking.

Here we have arrived at the problem that the researchers set out to tackle: is this unconscious association between alcohol and aggression activated by the “belief that alcohol consumption has occurred” or, alternatively, are “alcohol cues alone” sufficient for increasing “the accessibility of aggressive thoughts” (Leboeuf et al., 2020, p. 10)? Within two pages the authors have steered their readers from the urgent issue of preventing three million deaths by alcohol each year to a sophisticated psychological issue, which can be solved by conducting an experiment involving alcohol cues and a “lexical decision task”. The results of the studies Leboeuf et al. (2020, p. 10) aim to replicate “suggest that exposure to alcohol cues without [alcohol] consumption is linked with an increase in aggression-related thoughts”.

According to the reproducers in their closing paragraph, “[t]his line of research should be further pursued as it bears significant importance on [sic] today’s society” (Leboeuf et al., 2020, p. 16). But is there any practical value in this type of research? Even if the results were as expected, they would not help to tackle the alcohol-and-violation-problem. If we would decide to ban all “alcohol cues” from the public domain, alcohol use and abuse would continue. So suspicion arises that the references to the timeliness of a grave social problem are used as a legitimization for conducting an experiment that has to decide on a specific effect of alcohol primes, namely aggression.

The subject of the paper raises many obvious questions, like why wasn’t the research aimed at another, very obvious association: alcohol and pleasure? How and in what circumstances do people learn to associate alcohol with violence and aggression? And when we say people, do we mean both sexes or mainly “the male of the species”? And can we extinguish, for instance by operant conditioning, the automatic association between alcohol and aggression?

Conclusion

The call for more reproductions or replications is worthwhile, but only if the original experiments are solid and theoretically interesting, I would say. Readers may also have doubts about the practical value of this type of study, especially since there is hardly any knowledge of the relation between aggressive thoughts and subsequent behavior. So, despite of the laudable intentions of the reproducers to help solve the reproducibility crisis, it would seem that their efforts could only have been of interest within the milieu of priming specialists. These specialists however would probably not be very pleased with this experiment because of its obvious shortcomings, that are reported with unusual candor by the authors. Why attempt to publish a report like this in the first place? Publication might be instructive to psychology students on how not to perform replications, but is that a sufficient legitimation? Whatever the reason, the paper gave me the opportunity to reflect on the various levels of complexity that are involved in conducting priming experiments with human subjects and maybe help some social psychologists to reconsider their research practice.
Peer review
JOTE aims to make the peer review process accessible to its readers. Therefore, the initial submission with integrated peer review comments is available here.

References


Companion Piece

Abstract. Alcohol and its effects on aggression have been the subject of many discussions and research papers. Despite this fact, there is still a debate surrounding what it is exactly about alcohol that causes aggression. The current study sought to replicate the past finding by Bartholow & Heinz (2006), that alcohol cues without consumption increase the accessibility of aggressive thoughts, which can then influence aggressive behaviors. In the present study, participants had to complete a lexical decision task that was set up to assess whether aggressive words were detected faster in the presence of alcohol-related pictures compared to neutral pictures. The results of this study did not replicate the expected finding as only a main effect of word type was found in which participants detected neutral words faster than aggressive words. Furthermore, the study aimed to assess the role of gender stereotype acceptance levels in this association, but due to faulty design considerations, such analyses were not possible. The results are discussed in terms of the limitations of the study, and propositions for future directions are addressed.


License
Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and...
the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
Trial and Error (-Related Negativity): An Odyssey of Integrating Different Experimental Paradigms

Juliane Traxler¹,²†, Roxane V. Philips¹†, Andreas von Leupoldt¹, Johan W. S. Vlaeyen¹,²

Abstract
Pain can be considered as a signal of “bodily error”: Errors – discrepancies between the actual and optimal/targeted state – can put organisms at danger and activate behavioral defensive systems. If the error relates to the body, pain is the warning signal that motivates protective action such as avoidance behavior to safeguard our body’s integrity. Hence, pain shares the functionality of errors. On the neural level, an important error processing component is the error-related negativity (ERN), a negative deflection in the electroencephalographic (EEG) signal generated primarily in the anterior cingulate cortex within 100 ms after error commission. Despite compelling evidence that the ERN plays an important role in the development of various psychopathologies and is implicated in learning and adjustment of behavior, its relation to pain-related avoidance has not yet been examined. Based on findings from anxiety research, it seems conceivable that individuals with elevated ERN amplitudes are more prone to engage in pain-related avoidance behavior, which may, under certain conditions, be a risk factor for developing chronic pain. Consequently, this new line of research promises to contribute to our understanding of human pain. As in most novel research areas, a first crucial step for integrating the scientific fields of ERN and pain is developing a paradigm suited to address the needs from both fields. The present manuscript presents the development and piloting of an experimental task measuring both ERN and avoidance behavior in response to painful mistakes, as well as the challenges encountered herein. A total of 12 participants underwent one of six different task versions. We describe in detail each of these versions, including their results, shortcomings, our solutions, and subsequent steps. Finally, we provide some advice for researchers aiming at developing novel paradigms.

Keywords
error-related negativity, avoidance behavior, chronic pain, experimental psychopathology, experimental paradigm

Take-home message
Developing a new experimental paradigm is challenging, time-consuming and requires thorough testing. To make this process most efficient, we advise to clearly define the requirements at the beginning, to keep record of all decisions and adjustments made in the process, and to investigate encountered problems and failures which may help to gradually improve the paradigm.

Purpose
Avoidance behavior is part of a natural defense mechanism that is activated when individuals are experiencing chronic pain. In order to shed more light on the neural underpinnings of pain avoidance behavior, it is our long-term aim to investigate its association with the error-related negativity (ERN), an important component of neural error processing. So far, the ERN has been predominantly examined in relation to anxiety disorders and has been suggested as an endophenotype of obsessive-compulsive disorder (Riesel et al., 2011). Initially, ERN research focused on its relationship with the tendency
to learn best from negative feedback (Frank et al., 2005), whereas a few more recent studies also observed greater disorder-specific defensive motivation and avoidance behavior in individuals with elevated ERN amplitudes (Riesel, 2019; Riesel et al., 2019; Weinberg et al., 2016). These findings support the assumption that the ERN may be similarly implicated in pain-related avoidance, but this possibility remains to be tested. Investigating this theoretically plausible association may generally improve our understanding of the neural mechanisms underlying avoidance behavior and help to identify individuals at risk of developing chronic pain. The novelty of this endeavor requires the integration of knowledge of both research strands, as well as the development of an experimental paradigm that is well-suited to their respective requirements. Here, we describe the process of developing and piloting such a paradigm and provide an overview of problems encountered herein. We conclude with a tutorial section on the development of novel experimental paradigms based on our experience with this specific study, which we hope is also broadly applicable.

Introduction

Chronic pain affects as many as 20% of adults in Europe (of Pain, 2017). This poses a considerable health care challenge and causes tremendous individual suffering. Many chronic pain conditions are best understood from a biopsychosocial perspective which acknowledges the complex interplay of organic, cognitive, emotional, behavioral, and social factors. For example, on the neurobiological level, the sensitization model states that following acute pain, hypersensitization of local nerves occurs, causing exaggerated pain perception, or a lowered pain threshold, which may initiating a spiral in which hypersensitization triggers pain, which in turn triggers sensitization (Meeus & Nijs, 2007; van Wilgen & Keizer, 2012). Although the sensitization model is very intuitive, and often used as a metaphor in pain education with patients, it has also been criticized (e.g., van den Broeke & Van den Bergh, 2019). On the cognitive and behavioral level, one of the most prevalent explanations for the transition from a common acute pain episode to chronic disabling pain is the fear-avoidance model (Vlaeyen & Crombez, 2020; Vlaeyen et al., 2016; Vlaeyen & Linton, 2000). It posits that pain-related avoidance behavior, often fueled by catastrophistic (mis)interpretations of pain, contributes to individuals entering a downward spiral of fear, avoidance, inactivity, disability, and negative affect. Pain-avoidance refers to individuals avoiding stimuli that are predictive of pain or pain exacerbations. For instance, persons who have sustained an injury to their back, may avoid lifting objects for fear that it will lead to further pain, or that lifting may create bodily harm that is signaled by pain. Typically, avoidance occurs in anticipation of pain, and hence individuals have little opportunity to test and, if necessary, correct their beliefs about the threat associated with such activities. This failure to correct mistaken beliefs maintains pain-related fear and further increases mood disturbances and pain itself (van Vliet et al., 2018; Vlaeyen & Linton, 2000).

The neural underpinnings of pain-related avoidance behavior are poorly understood. A new and promising avenue to understanding pain-related avoidance is to study its relation to error-related negativity (ERN). The ERN is an event-related potential (ERP), which occurs within ~100ms after error commission (Hajcak, 2012). It can be measured as a negative deflection in the electroencephalogram (EEG) over fronto-central scalp positions, originating in the anterior cingulate cortex (Falkenstein et al., 2000; Gehring et al., 1993; Miltner et al., 2003; Taylor et al., 2007). Differences in the amplitude of the ERN are affected by both contextual and individual factors such as motivational aspects and error salience, and are considered to reflect individual error sensitivity (Hajcak, 2012; Hajcak & Foti, 2008; Taylor et al., 2007). For instance, it has been observed that the ERN is larger when errors are perceived to have worse consequences (Hajcak, 2012). Additionally, the ERN is believed to activate a defensive motivational system, in order to adjust the erroneous behavior and protect the organism (Hajcak & Foti, 2008). In line with these findings, increased ERN amplitudes have consistently been associated with anxiety disorders (Hanna et al., 2020; Weinberg et al., 2012) as persons affected arguably perceive their own mistakes in relation to the respective object of fear as highly threatening. In addition, avoidance behavior (including experiential avoidance) is a typical feature of anxiety disorders (Appperle & Paulus, 2010; Berman et al., 2010; Dikman & Allen, 2000; Dymond & Roche, 2009).

Given the commonalities between anxiety disorders and chronic pain (Asmundson & Katz, 2009), and the high prevalence of negative affect in chronic pain (Geisser et al., 2000), it is theoretically plausible that the ERN may be a biomarker for the development of pain-related avoidance behavior as well. In the case of anxiety, the particular feared object is considered threatening, whereas individuals with chronic pain perceive pain to be highly threatening. Because pain is such a salient and aversive stimulus, which conceptually signals bodily error and is often interpreted as such, avoidance responses constitute a defense mechanism to prevent the feared outcome (Leeuw et al., 2007). As such we may see an association between the ERN amplitude and the level of avoidance exhibited by an individual, with avoidance representing a defensive behavior in response to a motivationally salient and aversive stimulus.

Our long-term goal is to investigate whether individual differences in the ERN amplitude are related to differences in pain-related avoidance behavior, which may help identify individuals more at risk for the development of chronic disabling pain. To this end, we aimed to create an experimental paradigm that would allow us to measure the ERN and avoidance responses simultaneously, the process of which is described in the present manuscript. Previous experiments that studied this relationship in our lab (Traxler et al., in preparation) used the visual Flanker task unrelated to pain and had participants coming in on a separate day to measure pain avoidance behavior using a robotic arm (HapticMaster). A task integrating both the ERN and pain avoidance seems more efficient than measuring them in separate tasks. It also ensures that there is a relation between error commissions and pain-related avoidance, as it is for instance not clear how inhibition errors on a visual task relate to pain-related avoidance. Additionally, it puts less burden on the participant. Such a paradigm would need to meet requirements for both measures (ERN and pain-related avoidance), namely: (1) evoke an ERN; (2) yield a reliable estimate of the ERN, requiring a minimum of six errors (Olvet & Hajcak, 2009; Steele et al., 2016); and (3) be able to detect individual differences in avoidance behavior

1. Readers familiar with the error processing literature may wonder why we do not consider the Pe (error positivity), a later component of neural error processing. The reason is that we have no particular hypotheses about its relationship with pain-related avoidance, as it has not been as extensively studied in anxiety research as the ERN, especially with regard to avoidance behavior. We do aim to explore it in upcoming studies, but our main target of this report is the ERN, therefore we focused on developing a paradigm suitable to measure the ERN.
by providing sufficient opportunities to avoid (at least 20)\textsuperscript{7}. Although there is some debate as to whether one needs to be aware of having committed an error to evoke the ERN (Nieuwenhuis et al., 2001), awareness is essential in this task as participants would otherwise be unable to perform the avoidance response. In addition, the task ought to evoke inhibition errors rather than errors due to lack of skill or knowledge. For that purpose, we decided to integrate two experimental tasks, one of which is frequently used to measure the ERN (the Eriksen Flanker task; Eriksen & Eriksen, 1974) and the other is a common assessment of avoidance behavior (the avoidance task by Vervliet & Indekeu, 2015). We piloted this basic paradigm and continuously adjusted it based on the pilot results, leading to a total of six task variations. The present article focuses on the design and execution of these different tasks, what we can learn from them, and on suggestions for future research.

Methods

<table>
<thead>
<tr>
<th>Participants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twelve participants (8 females) took part in the pilot study between January and March of 2020. They were recruited by word of mouth, had a mean age of M = 29.25 (SD = 10.64), and gave autonomous consent. The exclusion criteria were self-reported: (a) pregnancy, (b) diagnosis of a psychiatric disorder (e.g. depression, anxiety, etc.), (c) a serious medical illness (e.g. diabetes mellitus, cardiovascular disease, neurological problems, etc.), (d) skin disease or condition where electrocutaneous stimuli could cause damage (e.g. an operation scar at the site where the electrodes would be placed), (e) pain: acute (e.g. due to an injury) or chronic, (f) electronic implant (e.g. pacemaker), (g) having been asked by a doctor to avoid stressful situations, and (h) poor vision or hearing that is not corrected. Participants gave written informed consent at the onset of the experimental session. Participants did not receive compensation for participation and mostly consisted of colleagues and acquaintances who were blinded with respect to the main aim of the tasks. The procedures conformed to the Helsinki Declaration and were approved by the institutional ethics committee (Social and Societal Ethics Committee; approval reference number: G-2019 12 1905).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Apparatus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Electroencephalography</td>
</tr>
<tr>
<td>Event-related potentials in response to errors and correct trials were measured by means of electroencephalography (EEG) which was continuously recorded from the scalp using a 129-channel system (HydroCel Geodesic Sensor Net, Philips Electrical Geodesics Inc., Eugene, USA). The sampling rate was set at 250 Hz (Hajcak &amp; Foti, 2008; Hajcak et al., 2005; Trujillo &amp; Allen, 2007), electrode impedances were kept below 50 kΩ and the vertex electrode was used as reference point (Sucec et al., 2019; Tan et al., 2019).</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stimulation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vibrotactile stimuli were administered to the left side of participants’ lower back through two vibrotactors (Dancer Design, St Helens, England) placed 2 cm apart (center to center; Figure 1). The stimulus duration was 112 ms and the vibration intensity, which was set to a clearly perceptible level that remained non-painful throughout the experiment, was kept constant across participants. Tactors were activated in a semi-randomized order.</td>
</tr>
</tbody>
</table>

Electrocutaneous stimuli (e-stim) were generated by a constant-current stimulator (DS7A; Digitimer, Welwyn Garden City, England) and delivered for a duration of 2 ms through two 4mm Ag/AgCl reusable snap electrodes filled with K-Y gel attached 1 cm above the vibrotactors. The intensity of these stimuli was individually calibrated to a level that was “painful and demanding some effort to tolerate”. For that purpose, a series of stimuli of ascending intensity was administered, each of which participants were instructed to verbally rate on a scale ranging from 0 (no sensation) to 10 (worst imaginable pain). The calibrated stimulus intensity was kept constant throughout the task.

The lower back was selected as stimulus location for two reasons: (1) the sensory acuity at the back is relatively low compared to more typical distal stimulation sites, such as the arms or hands (Weissman-Fogel et al., 2012), which was considered advantageous with regard to task difficulty; and (2) a large, even surface area was required to attach the vibrotactors and electrodes.

Software

The experimental task was programmed and presented in Affect 5 (Spruyt et al., 2009), which was run on a Windows 7 Professional (Microsoft Corporation Redmond) 64 bit Dell OptiPlex 780 (Dell Inc, Round Rock, TX) with 4-GB RAM, Duo-CPU at 3.10 GHz.

<table>
<thead>
<tr>
<th>Measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>The purpose of the pilot study was to create a task that allows to measure the ERN and avoidance behavior simultaneously. In each task version vibrotactors emitted vibrations on participants’ left lower back, EEG in response to correct and error trials was recorded, and painful e-stim were applied when participants made an erroneous response which they could omit by pressing the space bar. Avoidance was operationalized as the number of button presses to cancel an e-stim.</td>
</tr>
</tbody>
</table>

After performing the task participants were asked open questions to receive their perspective of the task and to ensure that requirements were met. Specifically, they were asked whether they were able to perceive when they made errors, how difficult they found the task as well as initiating the avoidance response, and any additional comments they would like to provide.

<table>
<thead>
<tr>
<th>General procedure</th>
</tr>
</thead>
</table>
| All task versions were based on a tactile task that was inspired by Eriksen’s Flanker task (Eriksen & Eriksen, 1974) and a fear-avoidance paradigm established by Vervliet and Indekeu (2015). The Eriksen Flanker task typically consists of presentations of arrow stimuli, (e.g. « > »), and participants are asked to indicate the direction in which the center arrow points by pressing one of two buttons as quickly as possible. The flanker arrows can be congruent
or incongruent and thus potentially cause inhibition errors\(^3\). In Vervliet & Indekeu’s avoidance task, participants view an office space on a monitor, in which a desk lamp can take one of three colors. Each color is indicative of whether or not an e-stim will be delivered. Participants may press a button to avoid the e-stim and learn to avoid based on the color of the lamp.

In the tactile task participants were instructed to distinguish between locations of vibrations emitted on their left lower back. We opted for tactile stimuli because these stimuli are more relevant and ecologically valid for pain responses, and more strongly resemble the tactile and proprioceptive input that frequently precedes or co-occurs with pain. Much like the arrow version of the Eriksen Flanker task in which the direction of arrows is the cue for a correct response, participants used left and right mouse buttons to indicate whether they felt a vibration at the left or right location respectively. Participants were instructed to respond as quickly and as accurately as possible. In addition, participants were informed that upon error commission they would receive an electrocutaneous stimulus, which they could cancel by pressing the space bar on the keyboard. This discrimination task was used in all task versions. Tasks employed either two or three vibrotactors (Figure 1). The left and right tactors will be referred to as “target tactors”, as they emit the target stimuli that are to be discriminated by the participant. The third tactor is referred to as the “distractor tactor” as it is meant to make the discrimination task more difficult.

A typical trial proceeded as follows (see Figure 2): a vibration was emitted (112ms), followed by a response window, to indicate the location of the vibration (200-1000ms)\(^4\). After a fixation period (300ms), participants could provide an avoidance response (1000ms). If participants chose not to avoid the e-stim upon error commissioning, an e-stim was delivered following this response window. The trial concluded in a jittered intertrial interval (600-1000ms). To prevent ‘better safe than sorry’ reasoning for performing the avoidance response and to mimic the costs that real-life avoidance typically bears, participants were told that five, and later two, trials would be added to the duration of the experiment every time they cancelled an e-stim. In reality no trials were added. As this proved to be highly aversive to the participants, leading to an unwillingness to avoid, we removed this clause after the eighth participant. The base settings for the tasks are as follows: vibrations are set to the same intensity and the third (distractor) tactor coactivates with every target stimulation. Task versions replace one another rather than being superimposed. Any deviations from the general procedure will be indicated, with Table 1 giving an overview of the task versions. In general, the task versions were composed of four or five blocks of 60 trials each, Table 2 indicates the total number of participants and trials per task version. In total, six different versions of the task were tested: Two Tactor task, Distractor task, 100% coactivation task, Different intensities task, 50% coactivation task, and a Sequence task.

**Note.** Number of tasks refers to whether participants had to perform any additional tasks to the discrimination task. Tasks are listed in the order in which they were tested.

<table>
<thead>
<tr>
<th>Two Tactor task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Procedure</strong></td>
</tr>
<tr>
<td>In this first version of the task, target tactors were attached to the left lower back of the participants (Figure 1, a). Participants indicated via left and right mouse button press whether they felt a vibration at the left or right location, respectively. Upon error commission participants were free to cancel the e-stim by pressing the space bar. One participant completed four blocks of this task version.</td>
</tr>
<tr>
<td><strong>Evaluation</strong></td>
</tr>
<tr>
<td>After running one participant it was quickly established that this task was too easy. This was reflected by participant feedback as well as the number of errors ((M_{\text{error}} = 2)). The participant was aware of having made mistakes but chose not to give any avoidance responses.</td>
</tr>
</tbody>
</table>

3. We seek to elicit inhibition errors, as this is how the ERN is traditionally measured (Ribes-Guardiola et al., 2020) with participants being aware of having committed such an error.

4. Non-responses or late responses did not bear any consequences as other studies conducted in our lab showed that participants do not simply cease to respond, even if this could be an alternative way of avoiding the painful stimulation.
Distractor task

Procedure
In order to increase task difficulty, three tactors were attached to the left lower back, with the two target tactors at their original locations and the third, the distractor tactor, placed centrally below them (Figure 1, b). This distractor tactor emitted vibrations to the beat of a widely known song (e.g., “Seven Nation Army” by The White Stripes). Participants were still expected to perform the same discrimination task (i.e. indicate on each trial whether the left or right tactor became active). Additionally, at the end of each block they had to choose, in a multiple choice manner, which song the distractor tactor had been vibrating to. One participant completed four blocks of 60 trials.

Evaluation
There were several issues with this task version. The task appeared to be difficult, leading the participant to be exclusively focused on the discrimination task ($M_{error} = 10$). Additionally, even though participants wore earplugs to prevent them from distinguishing vibrations based on the sound they made, the distractor tactor acted as a small speaker, and the participant became aware of the song because they could hear it. Due to these issues we did not run any further participants on this task.

Different intensities task

Procedure
This task used the 100% coactivation schedule, with the distractor tactor set to a higher intensity than the target tactors. The goal, once more, was to make the task more difficult, in an attempt to raise the number of error commissions across participants. One participant performed five blocks of this task.

Evaluation
Although this task did raise the number of error commissions ($M_{error} = 33$), according to participant feedback, the participant was less aware of having made a mistake and therefore unable to avoid the e-stim.

50% coactivation task

Procedure
In this task the distractor tactor only coactivated with target tactors 50% of the time. This adaptation was expected to make the task more difficult as participants may habituate less to the sensation of the third tactor. Two participants completed respectively two and one block on this task.

Evaluation
This task proved to be easier than its 100% coactivation counterpart, with two participants making 18.5 mistakes on average. Participants reported having difficulty recognizing error commission, as such further task versions were explored.
| Sequence task |

**Procedure**

In this last task, participants were instructed to recreate sequences of three vibrations using button presses. The target tactors went off in a sequence (e.g., left-right-right), and participants recreated this using the left and right mouse buttons. The distractor tactor followed a 100% coactivation schedule, i.e., coactivated with the target vibrations on each trial. Two participants performed five blocks of the task. We expected that this task would be more difficult as it required a more speeded response from the participants. Additionally, we thought it might lead to more inhibition errors: In the arrow version of the Flanker task, incongruent trials (i.e., those in which the middle arrow points in the opposite direction of the flanking arrows) tend to evoke more errors. Similarly, on the present task, two left activations, for instance, may mislead the participant to expect another left activation, instead of a right.

**Evaluation**

This task provided inconsistent results as the deviations in the number of error commissions varied widely between participants ($\Delta_{\text{ERN},\text{pilot}} = 73$, $\Delta_{\text{ERN},\text{pilot}} = 12$). Though individual differences in performance are to be expected in any task, the task should be reliable in evoking the needed number of error commissions. Participants were aware of their error commissions.

| Statistical Analysis |

The EEG data was processed in Brain Electrical Source Analysis Research 6.0 (BESA GmbH, Gräfelfing, Germany), using a high-cut filter of 30 Hz, a low-cut filter of 0.1 Hz and a notch filter of 50 Hz (Gorka et al., 2019). Ocular artifact removal was performed using the BESA algorithm, which does not require the measurement of an additional electrooculogram (EOG), but uses fronto-facial sensors to detect ocular movements. Bad channels were interpolated or deleted upon visual inspection, with a maximum of 12 bad channels [10% of the total number of electrodes; (Keil et al., 2014)]. Response-locked epochs of 1500 ms (500 ms pre- and 1000 ms post-response) were extracted and averaged per participant, using the 500 - 300 ms pre-response interval as baseline (Gorka et al., 2019; Jackson et al., 2015; Tan et al., 2019). Data were re-referenced to the average reference.

The ERN was operationalized as the mean amplitude in the 0 to 100 ms time window after error commission at the fronto-central site FCz (Geodesic net electrodes 6) (Hajcak et al., 2019). In addition, the correct response negativity (CRN) - a similar yet smaller negative deflection in the EEG signal following correct responses (0 - 100 ms post-response) - was computed in the same way. The difference score between the ERN and the CRN, known as the $\Delta$ERN, will be reported. Previous research suggests that the validity of the $\Delta$ERN is higher than that of the mean amplitude of the ERN alone (Gorka et al., 2019; Riesel et al., 2013). Both the preprocessed EEG and behavioral data were analyzed using R (Team, 2013). Due to the small group sizes across tasks and resulting insufficient statistical power to make reliable inferential conclusions, the analysis consists mostly of descriptive statistics.

| Results |

EEG data were recorded for 10 of the 12 participants, including both Sequence task pilots, the Different intensities pilot, one 50% coactivation pilot, and six 100% coactivation pilots. One participant on the 100% coactivation task committed less than six errors and was therefore excluded from the EEG analyses. No EEG data were recorded for the Distractor task or Two Tactor task. Given the small sample size, the results presented below need to be read with caution and should only be evaluated with regard to the pre-specified task criteria.

**Task performance and avoidance behavior**

On average, participants committed 24.07 errors ($SD = 18.33$, $range = 0-73$, 10.59% of all accepted responses) and gave 203.21 correct responses ($SD = 98.82$, $range = 1-297$) on the discrimination task. A Related Samples Wilcoxon Signed Rank Test revealed that reaction times were significantly faster during correct trials ($M = 527.35$ ms, $SD = 88.39$) compared to error trials ($M = 619.53$ ms, $SD = 114.75$; $Z = 3.059$, $p = .002$).

Table 2 shows the results for the behavioral responses. The Two Tactor task yielded few mistakes ($M = 2$, $SD = 0$) and no avoidance responses. The 100% coactivation task produced more error commissions ($M = 22.5$, $SD = 20.61$), and some avoidance responses ($M = 7.12$, $SD = 16.25$). The Different intensities task version led to a higher number of mistakes ($M = 33$, $SD = 0$), but did not produce any avoidance responses. Moving on to the 50% coactivation task, we observe a comparable average error commission as for the 100% coactivation task ($M = 18.5$, $SD = 19.09$). However, the frequency of avoidance is low ($M = 0.5$, $SD = 0.7$). This might indicate either that participants are unaware of having committed an error and thus are unable to avoid, or that they were unwilling to bear the costs introduced with this task version. The Sequence task produced the highest average error commissions ($M = 42.5$, $SD = 43.13$), but these scores are highly inconsistent across participants. The same applies to the avoidance responses in this task ($M = 27.5$, $SD = 19.09$). The Distractor task shows only $M = 10$ error commissions ($SD = 0$) and no avoidance responses. Note that results for each task version are based on unequal and few data points (Table 2).

We further compared when the most error commissions occurred in each task, in order to determine whether mistakes are more likely to be attributable to the novelty of the task or whether the task manages to consistently evoke error commissions. Across tasks, 56% of errors occurred in the first two blocks of the experiment. Note that for participants for whom the total number of trials was lower than four blocks, only the first block was considered. The task that evoked errors the most consistently throughout the blocks was the 100% coactivation task (45.8% of errors in the first two blocks) and the least was the Different intensities task (81.8% of errors in the first two blocks).

**Error-related negativity**

Figure 3 shows the response-locked ERP topographies for error and correct responses across the four task versions (Sequence, Different intensities, 50% coactivation, 100% coactivation) separately. All four tasks do show the potential for evoking an ERN with negative going deflections being observed at the expected fronto-central locations. Across all tasks, the average difference
FIGURE 3  Response-locked grand average waveforms for ERN and CRN at Cz with topography, averaged for (a) the 100% coactivation task (N = 5); (b) the 50% coactivation task (N = 2); (c) the Sequence task (N = 1); (d) the Different intensities task (N = 1); and (e) averaged across the four task versions (N = 9).

<table>
<thead>
<tr>
<th>Task</th>
<th>N</th>
<th>Total trials</th>
<th>Errors</th>
<th>Mean Avoided</th>
<th>Peak (ms)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Two Tactor</td>
<td>1</td>
<td>240</td>
<td>2</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>Distractor</td>
<td>1</td>
<td>240</td>
<td>10</td>
<td>0</td>
<td>NA</td>
</tr>
<tr>
<td>100% Coactivation</td>
<td>8</td>
<td>2100</td>
<td>22.5</td>
<td>7.12</td>
<td>-3.97</td>
</tr>
<tr>
<td>Different Intensities</td>
<td>1</td>
<td>300</td>
<td>33</td>
<td>0.5</td>
<td>-0.02</td>
</tr>
<tr>
<td>50% Coactivation</td>
<td>2</td>
<td>180</td>
<td>18.5</td>
<td>27.5</td>
<td>-0.05</td>
</tr>
<tr>
<td>Sequence</td>
<td>2</td>
<td>600</td>
<td>42.5</td>
<td>12-73</td>
<td>28</td>
</tr>
</tbody>
</table>

TABLE 2 Results of behavioral and EEG data

The score between ERN and CRN was $\Delta ERN = -2.27\mu V$. The task showing the most pronounced $\Delta ERN$ at FCz is the 100% coactivation task ($\Delta ERN = -3.97\mu V$) and the least pronounced $\Delta ERN$ was observed in the 50% coactivation task ($\Delta ERN = -0.02\mu V$). Though $\Delta ERN$ scores are variable across studies (Barker et al., 2015; Buzzell et al., 2017; Gorka et al., 2019), our results do suggest smaller difference scores than observed using traditional reaction time tasks. Importantly, given the small sample size for each task, no comparative conclusions can be drawn with regards to the ERN and its relation to pain-related avoidance responses across tasks.

Discussion

Here we describe the piloting of a novel paradigm based on the integration of two experimental tasks stemming from different fields of psychology. Our aim was to develop an instrument suited to measure both the neural activity during error commission that is followed by a negative consequence, namely pain, and the avoidance tendencies in direct response to these pain-evoking errors. As described above, this task is required to evoke an ERN and allow for its reliable measurement ($\geq 6$ errors; Olvet & Hajcak, 2009; Steele et al., 2016), evoke conscious errors of inhibition in order to allow participants to avoid the painful stimulation, and create sufficient variability in avoidance behavior ($\geq$ potential avoidance trials, i.e., error trials, and adequate balance between costs and benefits of avoiding vs. not avoiding). Altogether, six different task versions were put to the test, which are discussed in the following.

The first task using only two tactors proved to be too easy, based on the low number of mistakes made and the reported feedback of the participant. Despite the low perceptual acuity at the lower back, the vibrations were reportedly clearly distinguishable and after a first orientation during the practice phase, the participant had no difficulty completing the task. Given the small number of error commissions, this task did not allow for a reliable measurement of the ERN. With the aim of increasing task difficulty and thus the likelihood of error commissions, it was decided to add a third vibrotactor to the paradigm to distract from the target stimulation, which resembles the distracting flankers of the Eriksen Flanker task (Eriksen & Eriksen, 1974).

The Distractor task, the first version using the distractor tactor, was programmed to emit vibrations to the rhythm of famous songs which the participant
was instructed to identify whilst responding to the left and right vibrations. This task turned out highly demanding, leaving the participant focusing entirely on the discrimination task. Although this version successfully elicited a considerable number of errors, according to the participant the task load interfered with performing the avoidance response, which was perceived as disruptive, such that none of the e-stim were avoided.

In order to decrease the demands of the task, we switched to the 100% coactivation task, in which the distractor tactor coactivated with the target tactor on each trial. On average, participants made a sufficient number of errors for ERN analysis, as well as showing a reasonable variability in avoidance responses. Notably, the introduction of different types of avoidance costs during this task version led to a strong decline in avoidance responses and was therefore discontinued. While this version in principle did meet the pre-specified requirements, we decided to further modify this task in an attempt to ensure that the majority of participants would commit sufficient errors.

First, the vibration intensity of the distractor tactor was slightly increased with the aim to better mask the vibrations of the target tactors. This led to the participant being unsure about their performance and thus unable to perform the avoidance behavior. Second, after setting the vibration intensity of the distractor tactor back to its original level, we changed the coactivation to 50% of the trials with the goal of increasing unpredictability. Contrary to our expectations, participants experienced this version as easier than the 100% coactivation version and made fewer mistakes.

Lastly, in the Sequence task, participants were asked to perform a sequence of responses repeating after the vibrotactors. Given the speed of the task, it was hypothesized that participants would be easily led into committing errors of inhibition. This approach was successful for one participant; nevertheless, it appeared that the activation of both target tactors within one trial helped the other participant to orient towards the stimulation locations, causing a large difference in the amount of error commissions.

Overall, the tasks succeeded in eliciting an ERN at fronto-central sites when participants committed errors, which resembled the typical ERN as observed in more classic tasks such as the Eriksen Flanker task (Imburgio et al., 2020). Bearing in mind that the reported results are based on small sample sizes, it is noteworthy that the ERNs produced by the different task versions were slightly less pronounced and some peaked earlier than typically found with the Flanker task (Hajcak & Foti, 2008). Moreover, whereas commonly reaction times for error trials are found to be shorter than those for correct trials (Aarts & Pourtois, 2010; Imburgio et al., 2020), the opposite pattern was found in all of the piloted versions of this paradigm. However, a study that compared the ERN elicited by errors on the visual Flanker task with an interoceptive discrimination task using respiratory stimuli reported that the interoceptive ERN (intERN) did peak significantly earlier than the exteroceptive ERN (Tan et al., 2019). The authors of this article argue that interoceptive stimuli might be more relevant for survival, making the detection of interoceptive errors a priority. It is conceivable that somatosensory errors which indicate proximal threat carry this relevance, too, and thus evoke a similarly early peaking ERN, allowing for fast adaptation of behavior. A possible explanation for the less pronounced waveforms is that participants were less aware of having made a mistake than is commonly the case on the visual Flanker task. In addition, deviations of the present ERN waveforms from those observed in previous studies are most likely also related to the small number of averaged trials and/or participants (Kappenman & Luck, 2011) and therefore need to be interpreted with caution. Future well-powered studies will have to show whether these results are reliable and clarify the origins of potential differences between the processing of interoceptive and exteroceptive errors.

In conclusion, out of the six versions tested, the 100% coactivation task proved the most reliable and useful for investigating whether individuals with larger ERN amplitudes are more prone to engage in pain-related avoidance behavior. Nevertheless, this task does have several limitations: Firstly, error commissions may remain low for a fair number of participants, so that it can be expected that a study employing this paradigm will lose many participants on the exclusion criterion of 20 errors that are needed to obtain a useful assessment of avoidance behavior. Besides, as stated above, this criterion is somewhat arbitrary and future pilots will have to show if this number of error trials is sufficient to detect individual differences in pain avoidance. Hence, it is possible that even more errors would be required. Secondly, the operationalization of avoidance behavior as a single button press may be considered simplistic. Persons with chronic pain tend to engage in a broad variety of frequently subtle alternative behaviors and withdrawal from activities that are thought to evoke or worsen their pain (Asmundson et al., 1999; Volders et al., 2015). Capturing this complexity in an experimental paradigm is highly challenging, and even though better approximations than button presses do exist (e.g., Meulders et al., 2016; Claes et al., 2014), the constraints of the Flanker task, a speeded reaction time task, do not allow for longer, more complex avoidance behaviors in between trials. Even though other tasks can be used to measure the ERN apart from the Flanker task, such as the Stroop task and the Go/No-Go task (Riesel et al., 2013), the constraints of a speeded response task remain and would not allow for the integration of such avoidance responses either. Thirdly, in line with this limitation, engaging in avoidance behavior bears no cost in this paradigm, which causes a discrepancy with real-life avoidance where patients are faced with considerable costs to their social life, physical functioning, and personal well-being (Martínez-Calderon et al., 2019; Prkachin et al., 2007; Vlaeyen & Linton, 2012; Wiech & Tracey, 2013). Yet, when exploring options to add an avoidance cost, in the trade-off between these costs on the one hand and receiving a painful stimulus on the other, participants clearly favored the latter: Participants were instructed that engaging in avoidance behavior would have the consequence of trials being added at the end of the task. As this did not only prolong the duration of the experiment but also further increased the likelihood of error commission, participants appeared to prefer to endure the e-stim. In an attempt to balance the trade-off and increase the cost of not avoiding, the painful stimulation was made more threatening by stating that a “stimulus with a slightly higher intensity than the one previously calibrated” could be delivered occasionally. Surprisingly, this change also did not increase the number of avoidance responses. Adjusting the painful stimulus in length or quality was not feasible given the timing of the trial flow. Based on the low levels of avoidance behavior throughout the pilots, possibly due to the cognitive demand of switching from the discrimination task to avoiding the painful stimulus, a surge in avoidance responses is not expected. Moreover, low-cost clinical avoidance, such as constantly carrying pain medication, is
common and may be problematic as it impedes treatment and maintains fear (van Vliet et al., 2018; Vervliet & Indekeu, 2015; Volders et al., 2012). Hence, studying these low-cost avoidance behaviors is clinically relevant, and it was decided not to integrate any further avoidance costs.

Apart from these task-specific limitations, two limitations of the piloting process need to be mentioned: On the one hand, the small and unequal sample sizes for the different task versions do not allow extensive statistical comparisons. Some decisions were based on just one or two participants who could be outliers, rendering the conclusions invalid. Even though the impressions gained from the pilot participants’ experiences on the Two Tactor and the Distractor tasks were supported by informal pilots conducted on the authors themselves, more pilots per task version would be necessary to provide certainty on the utility of each version. Given that the aim of this study was not to test hypotheses and that a large number of within-subject trials was used, we believe that our observations are nonetheless informative. On the other hand, some changes to the paradigm were made simultaneously, making it impossible to disentangle the individual effects. The decisions to further modify the paradigm were made as soon as clear converging evidence was obtained from behavioral performance and self-report that a given version was either too easy or too complex, often after just one or two pilots, in order to save time and participants. While this allowed us to try out various ideas and establish a functional paradigm within a few weeks’ time, the paradigm clearly requires validation in future studies.

Altogether, developing this experimental task based on the integration of two well-established paradigms was demanding, despite the relative simplicity of the original paradigms. Although the Flanker and the avoidance task seemed compatible on account of their short response intervals and general set-up, many problems arose with regard to timing, task difficulty, and interference of responses, and possible solutions often caused unexpected side-effects to participants’ behavior, such as the introduction of an avoidance cost preventing participants from avoiding altogether. Therefore, a thorough and extensive piloting was (and remains) pivotal in identifying various pitfalls and re-adjusting the task. While the 100% coactivation task seems the most promising, we aim to conduct a second pilot study solely focusing on this task version to ensure (a) that our criteria are consistently met, and (b) that the criterion of minimally 20 error commissions is sufficient to observe inter-individual variability in avoidance, before implementing this paradigm in a larger study. In the following section, we have compiled a list of recommendations based on our experiences for those planning to create a novel experimental paradigm.

Tutorial
This pilot study was an interesting, highly instructive journey. Based on this experience, we propose five key elements to consider when developing a new experimental paradigm or merging existing ones.

1. Clarity: Prior to designing the experimental paradigm, it is advised to clearly define the requirements the paradigm ought to meet. One should consider any known necessities or restrictions inherent to the subject or to the original tasks. Relevant questions in this step are “What is the ultimate goal of this paradigm?”, “How do I operationalize the independent and outcome variables?”, “What does this operationalization require in terms of equipment and measurement?”, and “Which other variables need to be controlled for?”.

2. Thoroughness: Developing a novel experimental paradigm, even if it seems straightforward, requires thorough piloting, perhaps of several different approaches, and is often a process of trial and error. Keeping a detailed record of decisions, adjustments, and their motivation as well as problems encountered, is crucial to understand what works and what does not, and to compare different versions of a paradigm. We recommend considering the use of an open-source repository, for example the OSF platform (https://osf.io/), which may encourage researchers to be more transparent in their decision-making. This would not only be informative for colleagues but also benefit the researchers themselves by stimulating the a-priori formulation of well-founded arguments. Furthermore, it is advised to check all relevant data as seemingly small adjustments of a task may have major effects on participants’ perception and behavior, as illustrated by the apparent drop in avoidance responses upon the introduction of an associated cost in the paradigm described above. Next to documentation, transparency, and data checking, another important aspect of thoroughness is to implement changes subsequently rather than simultaneously, as this allows one to disentangle the respective effects. While we largely adhered to these principles, we introduced the avoidance cost at the first pilot of the 100% coactivation task and terminated both the cost and the pain threat at the same time. Moreover, we did not consistently use the same number of blocks throughout the process, which makes the comparison of task versions more difficult. In hindsight, these changes were hasty, and we eventually reverted to the pure version of the task without cost and increased pain threat.

3. Curiosity: When it becomes clear what does not work, the follow-up question should be “Why does it not work?” Exploring the problem more closely may lead to its solution and to an improvement of the paradigm. Thus, it may prove useful to include additional measures during the pilots: For example, in studies with human subjects the researchers may choose to use a post-experimental questionnaire that addresses participants’ experience of the task as a whole and of any manipulations, as well as their understanding of instructions and materials. An example from the current pilot study is the avoidance cost that was introduced to better reflect real-life avoidance: We initially expected that participants would readily avoid the painful stimulus and hence decided to use penalties to avoid a ceiling effect. However, by adding trials for each avoidance response, we apparently did not meet a good balance between the pain threat and the avoidance cost as it almost entirely prevented participants from avoiding, even after increasing the pain threat. The post-experimental questions revealed that participants experienced the cost as too aversive as it would prolong the duration of the experiment even further. This insight led us to discontinue the use of this avoidance cost. Preparing such questions ahead of the pilots, asking them systematically and adjusting them based on new insights may provide valuable information from a different perspective. Again, it may be valuable to share these experiences on OSF, which may aid others in replicating, extending or modifying the paradigm.

4. Creativity & Collaboration: Sometimes, a problem in the paradigm cannot be easily fixed. Researchers are well-advised to seek inspiration in directly
and indirectly related literature, to reach out to or work together with researchers experienced in the respective domains, and to test the paradigm both on themselves and colleagues to broaden their perspective. Hence, we made sure to first test each task version on ourselves, and shared our ideas and experiences at lab meetings, which inspired some of the task versions. Moreover, joining forces with other labs may facilitate the testing of multiple task versions simultaneously, allowing for larger samples to be tested.

5. Perseverance: Developing new paradigms or integrating existing ones takes time. We strongly advise to allow for both sufficient time and sufficient participants to run several pilots, especially if a version of the paradigm seemingly meets all requirements. A well-tested paradigm will minimize the risks during data collection. In the present case, due to limited resources we opted for small samples for each task version to establish their potential feasibility and utility. Yet, as a next step, we hope to conduct a more extensive pilot study on the most promising task (100% coactivation) in a larger sample (N = 10) before implementing it in a hypothesis-testing study aimed at investigating our initial research question. In fact, whereas a power calculation for a novel paradigm is likely to be difficult, it may be advisable to determine a minimum number of pilot participants a priori.

Conclusion
Developing a novel experimental paradigm is a challenging endeavor that necessitates good preparation and thorough evaluation through extensive pilot- ing. We advise to clearly specify the task requirements in advance, record all approaches and inspirations from various sources. As with conventional research projects, the development of a novel paradigm may greatly benefit from using an open source repository.

Peer review
JOTE aims to make the peer review process accessible to its readers. Therefore, the initial submission with integrated peer review comments is available here.

References


**Companion Piece**

**Abstract.** ‘Trial and Error (-related negativity)’ is a fascinating paper detailing the attempt to develop a new experimental paradigm to study the role of error-related negativity in the development of avoidance behavior. In my comments on this paper I will focus on the interaction between experimenters and participants as the former investigate various ways of designing the experiment, aiming to elicit the right kind of behavior from the participants. As in many psychological experiments, there is a fundamental tension here that experimenters must find a way to deal with: they must guide the subject to the proper performance, without the subject responding to the guidance as such. The performance must be natural, but within tight constraints. Recalcitrance or resistance of the subject must be prevented. Ultimately, the authors of ‘Trial and Error (-related negativity)’ failed in their attempt to do this. Their reflections on their failure are thorough and illuminating, but I will argue that they can be pushed slightly further.


**License**

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
Reflection on “Trial and Error (-Related Negativity)”

Maarten Derksen

Abstract

“Trial and Error (-related negativity)” is a fascinating paper detailing the attempt to develop a new experimental paradigm to study the role of error-related negativity in the development of avoidance behavior. In my comments on this paper I will focus on the interaction between experimenters and participants as the former investigate various ways of designing the experiment, aiming to elicit the right kind of behavior from the participants. As in many psychological experiments, there is a fundamental tension here that experimenters must find a way to deal with: they must guide the subject to the proper performance, without the subject responding to the guidance as such. The performance must be natural, but within tight constraints. Recalcitrance or resistance of the subject must be prevented. Ultimately, the authors of ‘Trial and Error (-related negativity)’ failed in their attempt to do this. Their reflections on their failure are thorough and illuminating, but I will argue that they can be pushed slightly further.

KEYWORDS

tact, performance, recalcitrance, spontaneity, validity

Introduction: three kinds of errors

This paper is about error on three levels. First, it deals with research into the ERN, error-related negativity. The ERN is a negative deflection in the electroencephalography (EEG) signal, which tends to occur within 100 milliseconds of making an error. The authors hypothesize that physical pain can be considered as a bodily signal that a type of error has been committed: there is a “discrepancy between the actual and optimal/targeted state”, as the authors put it (Traxler et al., 2020, p.27). This raises the question whether the ERN is also associated with pain and the avoidance of pain, and if so how. More specifically the authors want to know whether people with an elevated ERN are more prone to avoidance behavior, which in turn can lead to chronic pain.

I am an historian of psychology with philosophical interests and have no expertise in clinical neuropsychology, so I will not comment on this hypothesis. But the paper also deals with error in two other ways, which I do feel able to reflect on. The authors describe their attempts to develop an experimental paradigm for the study of the role of the ERN in pain avoidance. In these attempts they make errors which they then try to correct in a further attempt, six task versions in total. This is the second way this paper deals with errors - those of the experimenters themselves. But there is a third level too: it is crucial for the experimental task to induce the participant to make the right number of errors - not too many, not too few. The second and third aspects are obviously related: the errors of the experimenters concern, among others, the number of errors the participants make.

The experiment as a psychological problem

The authors describe their challenge as an interdisciplinary one: they had to combine elements of neurophysiology (ERN) with clinical psychology (pain avoidance). Specifically: they had to somehow induce an ERN in the participants and elicit and measure some type of avoidance behavior at the same time. Moreover, to determine what each participant’s average ERN is, they needed at least six ERN measures per participant, and thus a minimum of six errors. The errors, finally, had to be “inhibition errors”, not errors due to lack of knowledge or skill. It was not clear to me why this was important, or what an inhibition error is in the first place, but this is no doubt due to my own lack of knowledge in this field.

All in all, the specifications of the task were narrow and demanding: not any type of error would do (only inhibition errors); the errors had to produce a proper ERN; a minimum of six was needed; the participants had to be aware of their error (otherwise they would not show avoidance behavior); and of course there had to be pain associated with the errors, but not so much pain that the ethics committee would reject the pilot study, or the subjects would refuse to participate. The researchers had to set up the experiment in such a way that this narrow target was met, and the participants were the most challenging factor: the experimenters had to elicit the right kind of performance from them, without creating suspicion or resistance. Human participants tend to form their own ideas about the experiment, they do not merely respond to the stimuli.

The first one to discuss this particularity of the psychological experiment
was the American psychologist Saul Rosenzweig. In an article titled ‘The experimental situation as a psychological problem’, Rosenzweig (1933) argued that an experiment, in any scientific discipline, normally combines three factors: experimental materials, one or more experimenters, and instruments. Ideally, Rosenzweig argued, these three factors remain separate: ‘(e)very factor stays within its own domain’ (Rosenzweig, 1933, p.338). In psychology, however, this is often not the case, and this creates problems. The experimenter may unwittingly act as a stimulus in the experiment (thus becoming part of the experimental materials), because his or her presence may have an influence on the behavior of the subject. Conversely, subjects may take on the role of experimenter when they become self-conscious of their role in the experiment and start to modify their responses accordingly. Managing these twin problems requires the experimenter to be tactful in how they guide the behavior of the participant.

In a recent article, Breninkmeijer, Rietzschel, and I reported on a series of interviews we had conducted with researchers in psychology (Breninkmeijer et al., 2019). We had asked them about their informal research practices, that is to say those practices that are not made explicit in the method section of a paper but are nevertheless considered important. We saw two themes in their answers. The first is a strong concern with professionalism, expressing itself among other things in an orderly lab and a smoothly running experiment, and respectful (but not amicable) conduct towards the participants. The second theme is a focus on producing good data by managing the performance of the participant. This second theme is relevant with regard to this paper on ‘Trial and Error’. Much of the work that is done in the lab, including the kind of informal, unwritten work that we explored in our interviews, is geared towards eliciting the right kind of behavior from the participant. What happens in the lab has a theatrical quality: a particular performance is expected from the participants, and they are guided to it by the staging and scripting (or choreography) of the experiment and the conduct of the experimenter. Our interviewees mentioned the importance of the instructions (clear, fool-proof, not too long), of keeping participants focused and motivated (particularly challenging in cognitive psychology, where tasks are often boring and repetitive), and of creating a certain psychological realism, so that the artificiality of the laboratory situation is less salient and the stimuli and tasks more life-like (this was especially a concern for social psychologists).

What makes this work so difficult is that despite the staging and the scripting, the participant’s behavior must be spontaneous and natural. Or rather: the spontaneity is elicited and facilitated by the staging, the instructions, the props, the stimuli, and the conduct of the experimenter. A psychological experiment aims to create natural behavior artificially, aims to produce spontaneity (see also Derksen, 2001) about psychological tests). This paradoxical task demands that the experiment is crafted with subtlety and tact. The experimental situation and the experimenter must be forceful, yet unobtrusive. If the staging and the scripting become too prominent the participant risks become recalitrant, and their behavior is no longer natural and spontaneous (Derksen, 2017; Lezaun, 2007). In designing an experiment, the management of the participants’ awareness is therefore often a major concern: they must be attentive to the stimulus, but not to the stimulus as stimulus, as part of an experiment in an artificial laboratory situation, aimed at probing their responses (Abma, 2020). They must not become reflexive but remain natural and spontaneous.

There is thus a fundamental asymmetry in this endeavor: it is the experimenter who leads the participant according to a particular choreography and in an environment that the experimenter has created. The experiment is like a dance, with the participant following the lead of the experimenter. At the same time, it is essential that the experimental subjects participate in the dance freely and spontaneously, i.e. that they ‘remain themselves’. Everything in the experiment is aimed towards eliciting a response from the participant that is somehow ‘natural’ and ‘real’ rather than contrived and artificial - a response that represents real-life behavior, despite the fact that it occurred in a thoroughly artificial setting. This paradoxical challenge requires a great deal of ingenuity and tact from researchers.

Awareness and distraction

Interestingly, the authors of this paper designed their new experimental paradigm with the help of their twelve participants. Although they are described in the paper in the usual impersonal, technical terms – ‘twelve participants (8 females), ‘mean age of M = 29.25 (SD = 10.64)’ (Traxler et al., 2020, p.29) - the participants nonetheless had an active role in the pilot, almost as collaborators, specifically by supplying the researchers with their experience of the experiment. ‘(T)hey were asked whether they were able to perceive when they made errors, how difficult they found the task as well as initiating the avoidance response, and any additional comments they would like to provide’ (Traxler et al., 2020, p.29).

As noted in this quote, it was vital that the participants were aware they had made an error, because otherwise they would not “initiate an avoidance response”. The experimenters had to hit a very narrow target here. On the one hand, participants had to be induced to make a sufficient number of errors, so the task had to be difficult enough that participants did not always know the right response. On the other hand, once they had responded they had to be aware of having made an error, and thus did have to know the right response. This transition from not knowing to knowing had to occur within 1300 ms, that amount of time being the fixation period plus the time allowed for an avoidance response. This required precision engineering of the awareness of the participants.

It quickly became clear that the basic task - determining where on their lower back they felt a vibration from a “tactor” - did not induce enough errors. To make the task more difficult the researchers decided to distract the participants. The tool that they used to modulate the awareness and attention of the participants was the wonderfully named distractor tactor. The researchers tried various ways to distract the participants, aiming at just the right amount of distraction to induce errors, without distracting the participants so much that they no longer were aware of their errors. Their first attempt was creative, but unsuccessful. I suspect that working out which song is playing merely from the beat is difficult enough in normal circumstances, but if the beat is transmitted by vibrations on one’s lower back it becomes well-nigh impossible. I am not surprised that the one participant in this task version simply gave up trying and focused entirely on the main task of locating the main stimuli. Here we have an example of participant recalitrance, in this case not in response to the mere fact of being manipulated, but in response to the difficulty of the distractor task.

View interactive version here.
The participant refused to be distracted. Moreover, this participant discovered they could hear the beat as well as feel it: the distractor tactor worked as a little speaker, and the ear plugs that the experimenters had given the participant did not stop the sound. In other words, the distractor task was at the same time too difficult and too easy: it was too difficult to perform in the intended way (by attending to the vibrations of the distractor tactor), and at the same time it was too easy to not be distracted by it, and too easy to cheat. The experimenters had failed to tactfully usher the participant into the right state of distraction, and instead met resistance. The participant did not perform as the experimenters had intended. Ultimately, the 100% coactivation task, which the experimenters had tried first, was the only one that met the requirements of a sufficient number of errors combined with awareness of having made an error after responding.

**Stimulating avoidance**

Regarding the other part of the experimental paradigm, the avoidance response, matters were even more complicated. To get a sufficient number of avoidance responses, the experimenters not only had to make sure that participants were aware of having made errors, but also that they were sufficiently motivated to avoid the punitive electrocutaneous stimulus (e-stim) that would follow an error. While in the error-induction part of the task the attention of the participants had to be modulated in such a way that they made errors but where also aware that they made errors, for the avoidance response a different balance had to be struck. The dynamic of control and resistance that is so typical of psychological experiments is quite prominent here.

First, the e-stim had to be calibrated for each individual participant, in such a way that it was ‘painful and demanding some effort to tolerate’ (Traxler et al., 2020, p.29). The researchers, in consultation with the participants, had to find the golden mean between not enough pain (and therefore no motivation to avoid) and too much pain (which would lead participants to reject the experiment altogether, if it would even pass the Ethics Committee). The participants were allowed to choose their preferred level of punishment, so to say, but the researchers were nonetheless afraid that participants would subvert the experiment by always performing the avoidance response (pressing the space bar on the keyboard) as a precaution, whether or not they thought they had made an error. Thus, there had to be a cost to avoidance. At the same time, non-avoidance should be costly, too, otherwise the researchers would not get the kind of responses they were interested in.

To direct their participants to this happy medium between too much and too little avoidance, the experimenters employed threats. Participants were told that they would be punished for avoiding punishment. Every time they pressed the space bar to prevent the e-stim, extra trials would be added to the experiment. This ploy failed: Again participants resisted, this time by simply foregoing avoidance and accepting punishment. The authors imply that participants had made a strategic calculation: If each avoidance response would lead to extra trials, the experiment would take longer to finish. Moreover, participants realized that extra trials would also mean more chances to make errors, which would either be punished with the e-stim, or, if avoided, would lengthen the experiment even more. In other words, the threats of the experimenters unwittingly focused the attention of the participants on the boundaries of the experiment, rather than on the task within the experiment. The experiment became visible to the participants as a social situation that they were in, with a beginning and an end, and they were eager not to stay in it too long.

To ‘balance the trade-off’ (p.31) and make non-avoidance more costly the experimenters again issued a threat: occasionally a punishing stimulus with ‘a slightly higher intensity’ (Traxler et al., 2020, p.31) would be delivered after an error. This failed to strike sufficient fear into the participants and did not result in more avoidance responses. The authors are surprised and do not offer an explanation, but I suspect the reason may be similar to why the first threat failed. Participants perhaps reasoned that a ‘slightly higher’ intensity would still be bearable, bearing in mind that the Ethics Commission sets conservative limits on how much pain a participant may be made to bear. Participants did not attend to the task alone but were aware of the boundaries of the experiment (in this case the limits set by an Ethics Commission) and took these into consideration in the way they performed the task.

Above I wrote that, in psychology, the experimental situation and the experimenter must be forceful yet unobtrusive. It seems to me that, in as much as the pilot failed to elicit the right number of avoidance responses and at the same time increase their cost, it was due to the situation becoming too prominent relative to the task. The experimenters’ threats made the participants attend to the experimental situation as such, as well as to the task within it, and their responses became colored by their thoughts about the boundaries of the situation they were in. It is not clear whether the researchers asked the participants about their experiences with the threats, as they said they did regarding other aspects of the task. It would have been interesting to learn the participants’ reflections and see whether they confirm mine.

**Ecological validity**

Discussing the results of their pilot, the authors argue that the low-cost avoidance responses that they ended up with are nonetheless clinically relevant, since that kind of avoidance behavior, like always carrying pain medication, occurs in real life too. Given that they also emphasize the ‘considerable costs’ (Traxler et al., 2020, p.34) of avoidance behavior in real life this argument seems unconvincing. I am sure they would have preferred high-cost avoidance in their experiment. Perhaps this is an inherent limitation of studying this phenomenon experimentally. First, for ethical reasons it is difficult to make responses in an experiment really costly. Many if not most participants will realize this and understand that the cost they are threatened with is illusory, as indeed it was here. Secondly, if the cost of avoidance in real life concerns people’s ‘social life, physical functioning, and personal well-being’, as the authors write (Traxler et al., 2020, p.34), an experiment may simply be too limited a situation; too small and brief to cause that kind of cost. The authors themselves note that ‘the operationalization of avoidance behavior as a single button press may be considered simplistic.’ (Traxler et al., 2020, p.34) But it is, I think, not just the simplicity of a button press compared to the complexity and variety of real-life avoidance that is the problem. It is also the brief duration and narrow spatial boundaries of an experiment. If I may be permitted one remark on the topic of study itself: It seems to me that the cost of avoidance behavior in real life builds up gradually over time, in a variety of situations. It has a history, one that is longer than the duration of an experiment and that plays out in more locations than a single laboratory. The failure of this
pilot study to elicit high-cost avoidance behavior may well be due to the fact that an experiment is simply too limited an event to successfully emulate the development of pathological avoidance behavior in real life. As became clear in this pilot study, participants know there is an end to the experiment and take it into account in their response strategy. The problem is thus not so much that this particular experiment lacked ecological validity because of the way it was designed, but that any experiment would fail to simulate the development of pathological avoidance behavior.

It might be wise to at least complement experimental studies with research that looks at avoidance behavior in a longitudinal perspective, as a phenomenon that develops over a longer period of time, in various situations. The experience sampling method might be useful here: participants regularly (several times a day, for example) report their current state of mind, behavior, or experiences, typically via an app on their mobile phone. Alternatively, qualitative methods such as interviews could allow a more in-depth exploration of pain and avoidance. Such approaches still require a ‘performance’ from the participant and tactful guidance from the researcher, but their relationship will be different from that in an experiment, if only because it lasts longer. More like a friendship, perhaps, than a single dance.

The authors’ tutorial proposes ‘five key elements to consider when developing a new experimental paradigm or merging existing ones’ (Traxler et al., 2020, p.35). Each of these is solid advice, in my opinion, but it might be worthwhile adding a sixth: an experimental paradigm may not be suitable for your topic of study, so be willing to consider other options.

Peer review
JOTE aims to make the peer review process accessible to its readers. Therefore, the initial submission with integrated peer review comments is available here.

References

License
Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Companion Piece

Abstract. Pain can be considered as a signal of "bodily error": Errors – discrepancies between the actual and optimal/targeted state – can put organisms at danger and activate behavioral defensive systems. If the error relates to the body, pain is the warning signal that motivates protective action such as avoidance behavior to safeguard our body’s integrity. Hence, pain shares the functionality of errors. On the neural level, an important error processing component is the error-related negativity (ERN), a negative deflection in the electroencephalographic (EEG) signal generated primarily in the anterior cingulate cortex within 100 ms after error commission. Despite compelling evidence that the ERN plays an important role in the development of various psychopathologies and is implicated in learning and adjustment of behavior, its relation to pain-related avoidance has not yet been examined. Based on findings from anxiety research, it seems conceivable that individuals with elevated ERN amplitudes are more prone to engage in pain-related avoidance behavior, which may, under certain conditions, be a risk factor for developing chronic pain. Consequently, this new line of research promises to contribute to our understanding of human pain. As in most novel research areas, a first crucial step for integrating the scientific fields of ERN and pain is developing a paradigm suited to address the needs from both fields. The present manuscript presents the development and piloting of an experimental task measuring both ERN and avoidance behavior in response to painful mistakes, as well as the challenges encountered herein. A total of 12 participants underwent one of six different task versions. We describe in detail each of these versions, including their results, shortcomings, our solutions, and subsequent steps. Finally, we provide some advice for researchers aiming at developing novel paradigms.


© The Author(s) 2020
Burst Beliefs – Methodological Problems in the Balloon Analogue Risk Task and Implications for Its Use

Kristel De Groot¹

¹Erasmus University Rotterdam Institute for Behaviour and Biology (EURIBEB) | Institute of Psychology, Erasmus School of Social and Behavioural Sciences | Department of Applied Economics, Erasmus School of Economics; Erasmus University Rotterdam, the Netherlands

Correspondence: Kristel De Groot, Department of Psychology, Education & Child Studies/ Research Methods and Techniques, Erasmus University, Rotterdam, the Netherlands

Email: k.degroot@ese.eur.nl

Abstract
Studies in the field of psychology often employ (computerized) behavioral tasks, aimed at mimicking real-world situations that elicit certain actions in participants. Such tasks are for example used to study risk propensity, a trait-like tendency towards taking or avoiding risk. One of the most popular tasks for gauging risk propensity is the Balloon Analogue Risk Task (BART; Lejuez et al., 2002), which has been shown to relate well to self-reported risk-taking and to real-world risk behaviors. However, despite its popularity and qualities, the BART has several methodological shortcomings, most of which have been reported before, but none of which are widely known. In the present paper, four such problems are explained and elaborated on: a lack of clarity as to whether decisions are characterized by uncertainty or risk; censoring of observations; confounding of risk and expected value; and poor decomposability into adaptive and maladaptive risk behavior.

Furthermore, for every problem, a range of possible solutions is discussed, which overall can be divided into three categories: using a different, more informative outcome index than the standard average pump score; modifying one or more task elements; or using a different task, either an alternative risk-taking task (sequential or otherwise), or a custom-made instrument. It is important to make use of these solutions, as applying the BART without accounting for its shortcomings may lead to interpretational problems, including false-positive and false-negative results. Depending on the research aims of a given study, certain shortcomings are more pressing than others, indicating the (type of) solutions most needed. By combining solutions and openly discussing shortcomings, researchers may be able to modify the BART in such a way that it can operationalize risk propensity without substantial methodological problems.

Keywords
Balloon Analogue Risk Task, risk-taking, uncertainty, expected value, confounding, censoring.

Purpose
The Balloon Analogue Risk Task (BART) is one of the most widely used behavioral tasks in psychology and has an especially strong presence in the fields of decision research, addiction research, and neuropsychology. But despite its popularity, researchers using the BART seem largely unaware of the task’s methodological shortcomings, which sometimes leads to conclusions that are not supported by the data. This is likely a result of these shortcomings not being widely reported, as ‘failure’ is not considered a popular publishing theme. Therefore, the present paper aims to gather and review these shortcomings, as well as potential solutions.

Take-home Message
The Balloon Analogue Risk Task (BART) suffers from various methodological shortcomings. The present paper analyses these shortcomings and offers suggestions to mitigate their effects. Finally, it calls upon researchers to critically evaluate how these shortcomings impact their studies before deciding whether and how to use BART.

Introduction
To a large extent, psychological science rests on the promises of operationalization: defining fuzzy concepts as measurable variables, or in other words,
changing conceptual variables into operational ones (Shuttleworth, 2008). This process is imperative because most concepts researchers hypothesize about are not straightforwardly quantifiable. By defining how a concept is measured, operationalization allows hypotheses to take a falsifiable format and enables us to replicate findings. In a way, operationalizations are arbitrary, as concepts can be defined and thus measured in numerous ways – none of which are surely ‘right’. Nonetheless, some measures may be more suitable than others.

A notable example of a concept that can be operationalized in various ways is risk-taking (Lauriola & Weller, 2018), which has an important place in clinical, cognitive, and developmental psychology, as well as in the fields of criminology, economics, and management. One way risk-taking is operationalized in these fields is through self-report measures, such as the Domain-Specific Risk-Taking (DOSPERT) scale (Blais & Weber, 2006) and the Financial Risk Tolerance assessment (Grable, 1999). Another way is through computerized behavioral tasks, like the Iowa Gambling Task (Bechara et al., 1994), the Cambridge Gambling Task (Rogers et al., 1999), the Game of Dice Task (Brand et al., 2005), the Balloon Analogue Risk Task (Lejuez et al., 2002), and the more recent but already widely used Columbia Card Task (Figner et al., 2009).

Importantly, the quality of a study largely depends on the degree to which its operational measures reflect the underlying concept; in this case, one’s disposition towards risk-taking. If a task is a poor proxy for a concept or is subject to methodological or interpretational problems, any data resulting from it are of limited value to our understanding of the concept. In this regard, several studies have challenged the operationalization ability of the most-cited risk task, the Iowa Gambling Task (see e.g. Brand et al., 2006; Buelow & Suhr, 2009; Figner et al., 2009; Maia & McClelland, 2004). The Balloon Analogue Risk Task, which is the second-most cited, may yet suffer from even more severe issues, hindering its ability to operationalize risk-taking. While some individual issues have been reported in previous publications, no literature so far has discussed these collectively. The present commentary aspires to fill this gap.

The Balloon Analogue Risk Task

In the Balloon Analogue Risk Task, or BART for short, participants are presented with a computer screen showing a small balloon and a pump. They are told that every time they click the pump, the balloon expands, and a fixed amount of money (5 cents) is added to a temporary bank. Every pump also increases the chance of the balloon exploding (marked by a ‘pop’ sound from the computer), resulting in losing all money in the temporary bank for that particular balloon (trial). The point at which a balloon explodes varies across trials, ranging from the first pump to the point where the balloon fills the entire screen. Participants can decide to stop pumping the balloon at any point during a trial by clicking the ‘collect’ button (left in Figure 1), which transfers the money accumulated in their temporary bank to their permanent one, while a slot machine sound is played. Once a balloon explodes or once participants reach a trial by clicking the ‘collect’ button (left in Figure 1), which transfers the money accumulated in their temporary bank to their permanent one, while a slot machine sound is played. Once a balloon explodes or once participants

Figure 1: Set-up of the original Balloon Analogue Risk Task as described by Lejuez et al. (2002). An interactive illustration of the task is provided with the HTML version of this article.

Thus, the probability $p$ of the balloon exploding on the first pump was $1/n$, and the probability of it exploding on pump $i$ (given no prior explosion) was $p_i = \frac{1}{n−1+1}$. For orange balloons, the array ranged from 1 to 8 (hence $p_1 = \frac{1}{8−1+1} = 1/8$), for yellow balloons from 1 to 32 ($p_1 = \frac{1}{32−1+1} = 1/32$), and for blue ones from 1 to 128 ($p_1 = \frac{1}{128−1+1} = 1/128$). Their average explosion points were respectively 4, 16, and 64, with the same (randomly generated) sets of explosion points being used across all participants to limit extraneous variability. Neither the ranges nor the average explosion points were communicated to participants.

The BART’s design is intended to reflect naturalistic decision-making, in which taking more risk generally increases the odds of encountering a loss. This sort of decision-making tends to be emotionally engaging, instigating a sense of increasing tension as the balloon increases in size (Schonberg et al., 2011). In support of the BART’s validity, Lejuez et al. (2002) showed that the average number of times participants pumped the blue balloon significantly correlated with scores on risk-related constructs (sensation seeking, impulsivity) and with real-world risk behaviors, such as polydrug use, gambling, unsafe sex, and stealing. The orange and yellow pumps were originally not examined with respect to risk-related constructs, as their narrow ranges of outcome values (1-8 and 1-32) are less suited for capturing individual differences. Instead, their average pump numbers were analyzed together with those of the blue balloons to show that the number of times participants choose to pump is sensitive to the probability of exploding. Overall, the data showed the BART to have “particular promise as a behavioral index of risk-taking” (Lejuez et al., 2002, p. 82). As would be expected based on this conclusion, the BART (particularly its blue balloon) became a popular instrument for gauging individuals’ propensity for risk-taking, with inconsistent findings being attributed to factors like sampling variability and inadequate statistical power (Lauriola et al., 2014), rather than problems inherent to the BART. However, several authors have argued that such problems exist (De Groot & Thurik, 2018; Gu et al., 2018; Schmidt et al., 2019; Schonberg et al., 2011), and that they limit the BART’s ability to measure one’s propensity for taking risk. The key problems that characterize the BART are 1) a lack of clarity as to whether decisions are characterized by uncertainty or risk, 2) censoring of observations, 3) confounding of risk and
expected value, and 4) poor decomposability into adaptive and maladaptive risk behavior.

Risk or Uncertainty?
In economic theories of decision-making, a key distinction is that between uncertainty and risk, which is often accredited to Knight (1921), and was introduced to psychological thinking in a seminal paper by Edwards (1954) that lies at the origin of behavioral decision theory. When deciding under the condition of risk, the probabilities associated with the possible outcomes are known. When deciding under uncertainty (which some authors call ambiguity), this probability distribution is unknown.

For Knight (1921), this distinction was not only of theoretical but of practical importance as well. According to him, uncertainty – not risk – was the main driver of entrepreneurial success, as only people who recognize hidden opportunities can seize them and profit from them. Since then, the empirical relevance of the uncertainty-risk distinction has been confirmed in various fields of research. In economics, Ellsberg (1961) showed that individuals prefer risk over uncertainty, even if the known probabilities are unfavorable and the uncertain option could be a guaranteed win. In psychology, studies showed that uncertain and risky decisions involve different mental processes, as risk allows for statistical thinking (to optimize) but uncertainty involves heuristics (to satisfice) (Volz & Gigerenzer, 2012). In line with this, decision-making under risk is thought to depend more on executive function (such as categorization and cognitive flexibility) for which the dorsolateral prefrontal cortex is important, whereas decision-making under uncertainty hinges on emotional processes (such as somatic feedback), which are more associated with the ventromedial prefrontal cortex and the amygdala (Brand et al., 2006). This may explain why patients with executive deficits, such as those with Parkinson’s disease, have difficulty deciding under risk but have no trouble deciding under uncertainty (Euteneuer et al., 2009), whereas persons with obsessive-compulsive disorder, for example, show the opposite pattern (Starcke et al., 2010; Starcke et al., 2009).

Given that uncertainty and risk differ both theoretically and empirically, it is imperative for researchers to know the conditions under which participants decide. Unfortunately, despite the word ‘risk’ in its name, these conditions are not straightforward in the BART. Since participants are never given “detailed information about the probability of an explosion” (Lejuez et al., 2002, p. 77), we can assume that at least during early trials, they decide under uncertainty (Bishara et al., 2009; De Groot & Thurik, 2018; Schonberg et al., 2011). As they move further along in the task and ‘sample the distribution’ by pumping balloons and observing their outcomes, they get a better sense of the probabilities, which gradually moves their decisions in the direction of risk. Although not studied in the BART itself, such a shift has been shown for the Iowa Gambling Task, where performance in early trials does not correlate with that in later trials nor with executive function, indicating that people first decide under uncertainty and later under risk (Brand et al., 2007; Brand et al., 2006). While this effect may not be as strong in the BART, studies do show better performance in later compared to early trials, suggesting that participants indeed get a better grasp of the probability distribution over time (De Groot & van Strien, 2019; Lejuez et al., 2002).

The BART’s transition from uncertainty towards risk is problematic for several reasons. First, it is unclear when exactly this shift transpires, making it difficult to determine whether a decision in a given trial is made under uncertainty, risk, or something in between. Second, the point where decisions shift from uncertainty to risk is likely to differ between individuals, and is dependent on task characteristics (Brand et al., 2007; Brand et al., 2006). Third, the shift implies that the BART imposes learning demands, which could inadvertently impact participants’ outcomes on the task, with those capable of updating their knowledge of the probabilities performing better than those who have difficulty doing so. Fourth, once participants manage to derive the task’s probabilities, subsequent decisions are not characterized by what is usually considered risk. Contrary to decisions in which probabilities are explicitly described (‘a priori’ probabilities), probabilities in the BART are derived from experience. Since such probabilities depend on factors like sampling variability and one’s memory of previous events, decision-makers treat experience-based probability differently, which is called the description-experience gap (Hau et al., 2008; Rakow & Newell, 2010). Most notably, when deciding based on experience, people do not act in accordance with prospect theory, but instead, overweight rare events and overweight common encounters. As people have more and more encounters (e.g., trials), their experiences will approach the precision of a priori probabilities, though in practice this is difficult to attain (Knight, 1921).

To address the inability of the BART to differentiate between complete uncertainty, experience-based risk, and description-based risk, several approaches may be used. One option is to apply a model to the BART’s data that allows for participants learning through experience. An early example is a model by Wallsten et al. (2005) in which decision-makers update their probabilities from trial to trial, and continually re-evaluate their options. Alternatively, one could use a different task, in which decisions are either all characterized by uncertainty or risk, or which includes a well-understood shift between the two. Tasks that involve only uncertain decision-making are rather difficult to design, as they require participants to be ignorant of probability-related information and remain ignorant of that as well – automatically disqualifying tasks that have a learning curve. Tasks involving only decisions made under (a priori) risk are much more common and include the Cambridge Gambling Task, the Game of Dice Task, and the Columbia Card Task, the latter of which resembles the BART’s dynamic, affective nature (Schonberg et al., 2011). Finally, a known shift from uncertainty to (experience-based) risk can be found in the Iowa Gambling Task. This task’s shift, while not fully understood, has been studied more thoroughly than that in the BART.

Censored Observations
Statistical censoring refers to a condition in which the value of an observation is unknown because it is beyond a certain limit. This limit can exist by design, which is common in survival analysis. If a study on a surgical intervention follows patients for up to 10 years, the longevity scores of those who live past this term are censored, as their longevity is at least 10 (Young & McCoy, 2019). Censoring can also result from limits on what an instrument can reliably measure. For example, in the study by De Groot and van Strien (2019), on per-block averages is not reported in the published report but will be shared upon request.

1. The relevant data collected by De Groot and van Strien, 2019 on per-block averages is not reported in the published report but will be shared upon request.
measure. For example, the full IQ score of the Wechsler Adult Intelligence Scale ranges from 40 to 160 (Sattler & Ryan, 2009), meaning that IQ scores of people performing either extremely poorly or extremely well are cut off at these boundaries and are thus censored.

In the BART, censoring (by design) occurs if a participant is stopped from taking more risk in a given trial, because the balloon they are pumping explodes, forcing the trial to end. Since such a trial ends prematurely, the number of times the participant pumped the balloon does not necessarily reflect the risk they were willing to take, meaning their risk propensity is censored. This is problematic for various reasons. First, including these censored trials biases the average number of pumps downwards (especially for high-risk takers), underestimating participants’ willingness to take risks (Dijkstra et al., 2020; Pleskac et al., 2008). Likewise, the between-subjects variability across these averages is reduced (Lejuez et al., 2002). Overall, the (unadjusted) average number of pumps is an ill-suited operationalization of risk propensity.

As censoring affects all sequential risk-taking tasks like the BART (involving multiple decisions per trial) and various other research paradigms, like survival analysis, several solutions have been proposed. In the paper introducing the BART, Lejuez et al. (2002) suggest computing an adjusted pump average using only trials in which participants stopped voluntarily, that is, in which the balloon did not burst. However, by omitting explosion trials, censored observations are essentially treated as randomly missing, which is inaccurate (Pleskac et al., 2008). The more risk someone takes, the more likely it is that the balloon bursts, and that the trial forcibly ends. The termination of trials is therefore not independent from participants’ behavior. As a result, Lejuez et al.’s adjusted score tends to discard trials in which participants take a lot of risk. This causes the average number of pumps to be biased downwards, similar to the unadjusted score, but to a lesser extent.

To circumvent the problem of censoring, Pleskac et al. (2008) developed an automatic response version of the BART.† Contrary to the standard BART, in which participants inflate a balloon one pump at a time, the automatic BART lets them indicate their intended number of pumps beforehand. The balloon then inflates to the corresponding size, or until it bursts. This procedure allows for an unbiased statistic of risk propensity, as the intended number of pumps is now observable in all trials (Pleskac et al., 2008). However, it increases the time between decision and outcome, which may make decisions less emotional (impulsive) and more cognitive (planned) (Pleskac et al., 2008), and may reduce the salience of the outcomes. These effects, in turn, can affect participants’ risk-taking (Young & McCoy, 2019). In contrast, however, a study using the Bomb Risk Elicitation Task (BRET; Crosetto & Filippin, 2013), another risk task that uses delayed explosions to circumvent censoring, found that introducing such delays did not impact risk-taking.

Another solution to censoring is using a rigged task (Slovic, 1966). Participants are then told that failure can occur at any moment (in the BART, at any pump), but actually, it is set to occur at the last possible choice. Hence, participants can always stop voluntarily, and no scores are censored. To uphold credibility, ‘mock’ trials are added, in which failure is set to occur early on. Deciding on the number and timing of mock trials, however, is a challenge. Since behavior in a trial is affected by previous outcomes, experiencing (too) few failures could increase risk-taking (De Groot & van Strien, 2019; Dijkstra et al., 2020). Therefore, rigged tasks should be designed such that they produce failure rates similar to non-rigged tasks and should take into account that failure rates differ between participants too. However, research on the Columbia Card Task, another sequential risk-taking task, shows that this is often not the case (De Groot & van Strien, 2019).

A final remedy, which addresses the bias but leaves the BART unchanged, is to apply a statistical model to the resulting data that explicitly incorporates censored behavior. Such models consider all observed data, using the censored trials as lower bounds in determining a participant’s actual risk propensity. Some of them employ Bayesian (generalized) linear mixed-effects regression (Weller et al., 2019; Young & McCoy, 2019); others use maximum likelihood estimation, adding a cumulative distribution function to the likelihood function to account for censoring (Dijkstra et al., 2020; Tobin, 1958). Such models perform significantly better (i.e., have less biased predictions) than those that do not account for censoring. However, as is the case for all statistical models, their soundness hinges on the validity of their underlying assumptions (Schaefer & Graham, 2002), such as that of normality, whose violation not all models are robust against (Powell, 1984).

Confounding and Decomposability

The BART was designed to resemble real-world risk situations, where taking modest risk is generally advantageous, but taking excessive risk is increasingly unfavorable (Lejuez et al., 2002; Wallsten et al., 2005). Within a trial, every successful pump earns participants 5 cents, which are added to their temporary bank. As the amount accumulated in the bank grows, the relative gain of taking additional risk decreases, while the potential loss in case of an explosion increases. Additionally, the probability of the balloon exploding increases with every pump: from 1/128 on the first to 1/127 on the second, and so on. This combination of characteristics makes that the task’s structure entails a serious problem. Since both the balloon value (the amount collected in the temporary bank) and the explosion probability increase with every pump, the expected value of inflating the balloon – the product of the success chance and the reward, minus the product of the explosion chance and the balloon value – changes across a trial (Schmidt et al., 2019). This change is illustrated in Table 1. Early in a trial, the expected value of the pump is positive, so taking additional risk is advantageous. This prospect changes halfway when the expected value turns negative, making additional pumps unfavorable (Lejuez et al., 2002). Due to the expected value changing with each decision, it is confounded with risk (defined as the variability of the possible outcomes), which varies across decisions by design. Although such confounding can happen in real-life decision-making, it is not desirable in a controlled scientific environment: it makes it difficult to measure participants’ risk propensity, as both risk and expected value may influence their decisions. The extent to which individuals are, for example, risk-seeking, can therefore not be determined, because this would require showing a preference for higher variance payoffs, holding expected value constant (Schonberg et al., 2011).

This confounding demonstrates that the BART’s main observable outcome – the number of pumps participants press – cannot be interpreted as a straightforward indicator of risk propensity. Like many behavioral tasks, the BART

† An interactive illustration of this task is provided with the HTML version of this article.
supposedly gauges a single cognitive construct, but it manipulates various other, potentially confounding constructs as well (Schonberg et al., 2011). Expected value is an example of such a construct. As a result, the single score provided by the BART cannot easily be decomposed to identify the cognitive or neural mechanisms involved in the pump decisions. Studying the risk-taking process in isolation using the BART is therefore not possible.

One approach for resolving the confounding and decomposability issues in the BART is to apply a computational model to its data that quantifies the cognitive mechanisms underlying the observed behavior (Bishara et al., 2009). Such models were first proposed by Wallsten et al. (2005), inspired by an expectancy–valence model for decomposing behavior in the Iowa Gambling Task (Busmeyer & Stout, 2002). Wallsten et al. explain decision variability using one parameter for risk-taking, one for response consistency, and two for learning. By applying these models, we can study risk-taking – and other aspects that determine BART behavior – in isolation, by translating “what is observed but relatively uninformative to what is unobserved and relatively informative” (van Ravenzwaaij et al., 2011, p. 95). However, data from the BART may not be rich enough to warrant the use of complicated decomposition models. For instance, a study on Wallsten et al.’s best performing model demonstrated that its learning parameters could not reliably be recovered (van Ravenzwaaij et al., 2011). To allow for more extensive decomposition, one may need to resort to a different task, like the Iowa Gambling Task. Alternatively, one could use a task that by design avoids confounding, such as the Columbia Card Task. Although dynamic and affective like the BART, this task orthogonally varies risk-related constructs, so that they can be decomposed into their underlying mechanisms – like sensitivity to gains, losses, and probabilities – without the use of a computational model (Dijkstra et al., 2020; Figner et al., 2009; Schonberg et al., 2011). Finally, researchers can choose to design a custom task to ensure that the constructs relevant to their hypotheses are not confounded. For example, a risk task presented in Schmidt et al. (2013) varies the level of risk but holds expected value constant. Solutions such as these should be considered carefully so that constructs crucial to a study’s hypotheses can be isolated effectively.

The Normative Solution

The BART is designed in such a way that the balloons’ average explosion point lies at 64, halfway the maximum number of pumps. This is achieved by randomly generating collections of explosion points until one produces an average of 64 over all trials, as well as within each set of 10 trials (Lejuez et al., 2002). Participants can then maximize their earnings by attempting to pump every balloon 64 times, which results in an explosion in about half of the trials, and an optimal overall expected value. Going back to Table 1, we can see exactly why this is the optimal, or normative, solution in the BART. Up to and including the 64th pump, the expected value of pumping the balloon is positive; after 64, the expected value is (increasingly) negative. It is, therefore, optimal to aim for 64 pumps on every balloon, and then stop. Choosing to pump more or fewer than 64 times will decrease expected earnings; and the farther one deviates from the optimum, the lower the expected earnings become (Lejuez et al., 2002; Pleskac et al., 2008; Wallsten et al., 2005). Remarkably, in most trials, participants stop pumping the balloon far before the optimal stopping point (Lejuez et al., 2002). In fact, the average adjusted pump score is typically between 26 and 35 (Pleskac et al., 2008). Real-world risk-avoiders and risk-takers alike rarely pump the balloon enough times to maximize their expected earnings. This is less of a problem in the automatic BART, although participants there still pump fewer than 64 times on average. For example, two recent studies reported averages of 61.9 (Bernoster et al., 2019) and 58.5 pumps (De Groot & van Strien, 2019).

It is yet unknown exactly why participants often stop pumping before they reach the optimal point, but various factors may play a role. First, since the original BART requires participants to inflate balloons one pump at a time, it is plausible that they get tired of pumping after a while. Second, participants may want to limit their effort out of laziness or a desire to finish early (but see Young & McCoy, 2019). Third, they may become satiated: due to diminishing marginal returns, adding 5 cents to a growing temporary bank may stop being an attractive prospect well before reaching pump 64. Fourth, participants may need time to learn which strategy results in maximal earnings (Lejuez et al., 2002). This conjecture is supported by the observation that participants in both the original and the automatic BART on average press closer to the normative solution in the final block of 10 trials than they do in previous blocks (De Groot & van Strien, 2019; Lejuez et al., 2002).1 It also corresponds with the presumed shift from deciding under uncertainty to deciding under risk. In the BART, learning the optimal solution is hard, as the range of possible explosion points is large (1–128), and individual explosions provide limited feedback. This is in line with findings by Lejuez et al. (2002), who show that larger explosion ranges result in larger relative deviations from the optimum.

The fact that participants in the BART often stop pumping before the optimal stopping point has serious implications for how the data can be interpreted. Up to 64 pumps, the risk they take can be characterized as adaptive or functional, as it results in higher earnings. After that point, it can be considered maladaptive or dysfunctional, as it reduces expected earnings. Since people generally pump fewer than 64 times, the BART cannot properly differentiate between adaptive and maladaptive risk behavior, neither within nor between participants. A second, related problem is that experimental manipulations meant to increase risk-taking (such as adding time pressure or administering a certain drug) generally do not lead to lower earnings, as even the resulting higher pump numbers usually do not exceed 64 (Pleskac et al., 2008). For example, if a manipulation causes participants to take more risk and press 50 instead of 30 times, they are actually, on average, better off than before, the opposite of what one would expect in real life. In short, if participants mostly stay under 64 pumps, they simply never reach the point where taking more risk becomes disadvantageous, which limits the conclusions one can draw from the data.

The most straightforward way to mitigate these problems may be the modified BART developed by Pleskac et al. (2008), which differs from the original task in three ways. First, it involves an automatic response mode: participants indicate their intended number of pumps at the start of each trial, after which the balloon automatically inflates to the corresponding size (or until it bursts). Although meant to mitigate censoring, this adjustment may also prevent people from getting tired of pumping and from wanting to finish the task sooner. Second, the adjusted task provides explicit feedback about the explosion point of every balloon, not merely of those that actually explode. This may improve
participants’ learning across trials. Third, participants are (truthfully) informed that the range of pump numbers is 1-128 and that the best overall number of pumps is 64, further increasing the amount of information at their disposal.

These three modifications together successfully moved participants’ behavior closer to the normative solution of 64, with an average pump score of 57.7 for females and 63.7 for males (Pleskac et al., 2008). Part of this effect can be attributed to the automatic response mode, as these averages are higher than those from a manual BART with full feedback and strategy instructions added. Since this manual BART itself resulted in higher averages than the original BART, the feedback and instructions likely also contributed to the effect (Lejuez et al., 2002). Recent research, however, indicates that informing participants about the optimal strategy is not necessary, and even ill-advised. Two studies using an automatic BART with full feedback – but without strategy instructions – found equally high pump averages as did Pleskac and colleagues (Bernoster et al., 2019; De Groot & van Strien, 2019). Additionally, these studies found that a subgroup of participants – often from a STEM background – seem to infer the optimal strategy without any help.\(^2\) Their repeated 64-answers, therefore, reflect cognitive ability rather than risk propensity and reduce task variability. Informing participants about the optimal strategy can increase such problematic responses. Therefore, it seems best to add automatic responses and full feedback to the BART, but not strategy instructions. This will likely elicit sufficiently high pump averages, without compromising the validity of the task.

**Discussion**

Since it was first published in 2002, the BART has become one of the most popular tools in psychology to gauge individuals’ propensity for risk-taking. Halfway 2020, the original article describing the BART (Lejuez et al., 2002) had been cited over 1100 times in Scopus, most often in journals on decision research, addiction, and neuropsychology. This popularity is well-founded. The BART succeeds in recreating the ‘natural’ feeling of exhilaration and tension people experience when taking risk, and thus has excellent ecological validity. Furthermore, it correlates well with self-reported risk-related constructs, such as impulsivity and sensation-seeking, and with real-world risk behaviors, like polydrug use and unsafe sex, supporting its convergent validity. Lastly, it does not correlate with constructs like depression and anxiety, endorsing its discriminant validity (Lejuez et al., 2002). But despite these qualities, the BART

\(2\) The relevant data collected by Bernoster et al., 2019 and De Groot and van Strien, 2019 on individual answering patterns was not published but can be shared upon request.
suffers from methodological problems, most of which have been acknowledged in previous research as negatively impacting its rigor. The present paper is the first to give a comprehensive overview of these problems.

The first problem concerns the lack of clarity as to whether decisions in the BART are made under uncertainty (where outcome probabilities are unknown) or risk (where they are known). Since participants are not given any information about the explosion probabilities, they first decide under uncertainty, which then gradually shifts towards risk as they learn more about the probabilities in the task. As it is unclear exactly when this shift takes place, it is difficult to determine whether a given decision is made under uncertainty, risk, or something in between. The second problem concerns statistical censoring, which occurs in trials where the balloon explodes, as participants are then prevented from taking additional risk. As a result, the average number of times participants pump the balloon underestimates their risk propensity. Third, the BART confounds risk with expected value. Since these constructs change simultaneously throughout a trial, participants’ pump behavior again does not reflect risk propensity, as decisions are influenced by both risk and expected value. This also means that the task is poorly decomposable, as it cannot disentangle the motives underlying a pump decision. A final problem concerns the task’s normative solution. In the majority of trials, participants stop pumping before the point where expected earnings are maximized. Therefore, participants mostly take adaptive risk, which leads to higher earnings. Maladaptive risk-taking hardly occurs, even though one would expect to see such behavior in certain cases.

Despite these problems, much of the research up to now has focused on the empirical findings produced by the BART, rather than on the task itself, with the majority of researchers using the task without critically reviewing whether its problems interfere with their aims. This can have undesirable consequences, such as when it leads to false positives or false negatives. For example, one may fail to show a relationship which only exists for decisions characterized by risk, as some trials in the BART are characterized by uncertainty instead. Conversely, a hypothesis may pertain to people’s response to changing risk and be unjustly supported, as in the BART, risk and expected value simultaneously change and impact individuals’ behavior. Finding true positives and negatives hinges on several factors, an important one being the validity of the measurement instrument. Any data resulting from instruments that suffer from methodological or interpretational problems is of limited value to understanding the concepts they are supposed to operationalize.

For these reasons, it is imperative that researchers critically evaluate the ‘fit’ between their research and the BART before deciding on using it. For many research aims, one will now see that the original BART does not suffice. One may consider rigging it (Figner et al., 2009; Slovic, 1966), providing additional feedback, or automating the responses (Pleskac et al., 2008). Third, one may consider using a different task. This can be an existing (sequential) risk-taking task, like the Columbia Card Task (Figner et al., 2009), which performs better in terms of decomposability than the BART. Alternatively, researchers should consider creating a custom task that exactly suits their research, avoiding methodological flaws that could endanger the soundness of their conclusions. For instance, a task developed by Schmidt et al. (2013) involves decisions under conditions of explicit risk and does not confound risk with expected value. An important goal to keep in mind when designing such bespoke tasks is to combine strong ecological validity with methodological rigor (Schonberg et al., 2011).

Clearly, none of the solutions proposed can be considered a ‘universal’ fix that solves all of the BART’s problems. Depending on the aims of any given study, certain problems will be more pressing than others, indicating the (type of) solutions most needed. By combining solutions, researchers could work towards a task that can operationalize risk propensity without substantial methodological or interpretational problems. For example, an automatic BART with full feedback and explicit information on the probability distribution provides uncensored decisions made under clear risk that are at times risky enough to be maladaptive. If the resulting data from this adapted BART are then analyzed using a model like that by Wallsten et al. (2005) or that by van Ravenzwaaij et al. (2011), all problems reviewed in the current commentary would be addressed. However, this does not necessarily mean that this combination of solutions constitutes a universal fix after all, as the BART may face more problems than the ones discussed here. In all likelihood, the present review is not exhaustive. Researchers using the BART may know of additional problems, although this is unlikely to show in their work, as journals – and by extension researchers – do not consider ‘failure’ a popular publishing theme (Ferguson & Heene, 2012; Song et al., 2009). Therefore, it is important for researchers to not only critically evaluate the instruments they use but to disclose these evaluations as well, so that any and all methodological shortcomings can be openly discussed and addressed, improving the quality of the measures used.

Conclusion

The present paper is the first to review the methodological shortcomings of the Balloon Analogue Risk Task, a highly popular risk-taking task in psychology. The main problems identified are the ambiguity between uncertainty and risk, censoring of observations, confounding of risk and expected value, and poor decomposability into adaptive and maladaptive risk-taking. In addition, the paper reviews solutions that mitigate these problems. By presenting this first-time inventory, the paper highlights earlier mentions of problems in the BART as well as proposed solutions. It calls for a critical attitude towards the BART and experimental tasks in general, as their design deserves at least as much attention as the findings they produce. It also sets the agenda for testing and comparing different tasks and task versions, to explore which designs result in the best usability, reliability, and validity, so that risk propensity can be measured in the most accurate way possible.
Peer review

JOTE aims to make the peer review process accessible to its readers. Therefore, the initial submission with integrated peer review comments is available here.

References


**License**

*Open Access.* This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
Towards an Expanded Understanding of Publication Bias

Nicole Nelson

1Department of Medical History and Bioethics at the University of Wisconsin–Madison. Other affiliations are with the Holtz Center for Science and Technology Studies and the History department

Correspondence:
Email: nicole.nelson@wisc.edu

Abstract

This short essay argues for an expanded conception of publication bias. In addition to considering the selective publication of results, I argue that we need to also consider the selective publication of epistemic by-products—observations and knowledge that scientists accumulate incidentally in the process of carrying out their work. There are three reasons why we should be concerned about the exclusion of epistemic by-products from the published literature: first, because they play an important role in robust replication attempts; second, because their absence can result in misplaced scientific certainty; and third, because they contribute to a holistic understanding of natural phenomena. However, identifying and addressing publication bias against epistemic by-products and other undervalued forms of knowledge is more difficult than identifying bias against quantitative findings. I argue that scientific pluralism and making data publicly accessible are two potential remedies for addressing this form of publication bias.

KEYWORDS
epistemic by-products, publication bias, replication geneticization, scientific pluralism

Introduction

Publication bias (also known as selective publication, or the “file drawer” problem) has long been recognized as a problem in the sciences (Dickersin, 1990; Dickersin, Kay et al., 1987; Rosenthal, 1979). The term describes problems arising from how scientists choose which studies (or which elements of studies) to publish from the many studies that they have conducted. If researchers decide whether to publish based on a property of the result—say, whether the result reaches commonly-used thresholds of statistical significance or supports the researcher’s hypothesis—then the subset of study results that appears in the published literature will be biased.

In this short essay, I argue for an expanded conception of publication bias. In addition to considering the selective publication of results, I argue that we need to also consider the selective publication of epistemic by-products—observations and knowledge that scientists accumulate in the process of carrying out their work (Nelson, 2018). Researchers typically choose not to publish these findings not because they fail to reach statistical significance, but because they were never intended to be knowledge in the first place. This implicit distinction between the entities that scientists consider to be scientific findings and those that they consider to be anecdata, tacit knowledge, or lab lore acts as an additional filter that prevents some types of knowledge from circulating widely. Expanding the notion of publication bias to include these processes allows for a deeper understanding of how users of the scientific literature might arrive at a false sense of certainty, and also offers insight into how gaps form between individuals’ understanding of natural phenomena.

Existing research on publication bias

Researchers studying publication bias have tended to focus on studies that have clearly specified questions, and outcomes that are assessed with quantitative measurements. Studies that ask questions such as “Is psychotherapy effective?” or “Does smoking increase risk of heart disease?” can be grouped in a way that allows researchers to define a universe of studies and then ask questions about the subset of those studies that appear in the published literature. The quantitative measurements typically employed in these studies—p-values, effect sizes, hazard or odds ratios, correlation coefficients, and so on—allow researchers to compare across studies to look for evidence of bias. For example, “funnel plots” of outcome measures allow researchers to visually detect bias in the group of studies that they have collected (Begg & Mazumdar, 1994; Egger et al., 1997).

In these plots, the effect sizes of the individual studies should theoretically be evenly distributed across the plot, and a plot where studies are clustered to one side allows researchers to see a “missing mass” of unpublished studies (Figure 1). Multiple proposals for new techniques to visualize and quantify publication bias have emerged in the past few years (Schimmack, 2020; Simonsohn et al., 2013; van Assen et al., 2015).

The main harms of publication bias, according to this research, are that evidence will appear more certain or interventions will appear more effective than they actually are. If the studies that find weak effects or no effects for an intervention are not published but the studies that find strong effects are, then the users of that published literature will be accessing a biased sample of results. For example, if a doctor is reviewing the published literature on a particular
drug, selective publication will make it more difficult to assess the true efficacy of that drug or its potential harms. Worse still, if the doctor assumes that studies available in the published literature are an unbiased sample, then they are likely to arrive at the conclusion that the evidence for using the drug is more conclusive than it actually is. As one group memorably put it, without the publication of negative results, false claims are “likely to be canonized as fact” (Nissen et al., 2016). This false sense of certainty can result in quite direct harms, as in the case of selective serotonin reuptake inhibitor (SSRI) drugs, where researchers have argued that doctors were prescribing for years on the basis of evidence that was far more equivocal than it appeared when seen through the published literature (Melander et al., 2003; Turner & Tell, 2008).

Not all research programs, however, have clearly defined starting/stopping points or outcomes, which may make it difficult to delineate a universe of studies in which to study publication bias, or to compare across studies. This may explain why systematic review approaches to analyzing publication bias are more common in fields such as clinical research and psychology than they are in basic or preclinical fields such as laboratory animal research (Korevaar et al., 2011). Even within animal research, those studies that have addressed publication bias have tended to focus on studies of drug efficacy that resemble clinical studies (MacLeod et al., 2004; Sena et al., 2010) rather than more exploratory research. Some researchers have circumvented the question of how to define a universe of studies or compare studies that don’t employ null hypothesis significance testing by asking simply what proportion of publications report any kind of result that support the author’s hypothesis (Fanelli, 2010, 2012). Others have used surveys to gather researchers’ estimates on what proportion of results go unpublished in fields such as laboratory animal research, where existing techniques for assessing publication bias are more difficult to employ (Riet et al., 2012).

**Epistemic by-products and publication bias**

Even these more expansive approaches, however, rely on a fairly narrow understanding of what counts as a result in scientific research, and therefore what counts as publication bias. In my ethnographic studies of animal behavior genetics researchers, I found that what these researchers learned in the course of conducting their experiments far exceeded the outcomes they envisioned in their research proposals, if what counts as a “outcome,” “result,” or “knowledge” is conceptualized broadly (Nelson, 2018). The researchers that I studied were aiming to understand genetic contributors to behaviors such as anxiety and drug addiction. A typical research project might involve breeding mice who voluntarily drank large quantities of alcohol, and then examining their genomes to see what regions were being selected for through this breeding process. But in order to identify genetic contributions to behavior, researchers needed to understand and control for the many other factors that might also contribute to behavior. Through the process of implementing these controls, they ended up learning as much—if not more—about the impacts of the environment on behavior as they did about genetics. This learning, however, was qualitative rather than quantitative, and took place within small collectives of scientists rather than within the broader scientific community.

Learning about environmental factors was almost unavoidable because of the impossibility of controlling the many factors that might change mouse behavior. Construction and fire alarms introduced stressors into the lives of their animals, shifts in facility policies meant shifts in the animals’ diets and bedding, the distinctive smell and demeanor of each researcher introduced uncontrolled variation, and changes in researchers’ lives over time sometimes produced noticeable changes in their results. On occasion, these experiences came close to resembling controlled experiments, where all of the seemingly important elements were held constant but for one source of environmental variation that researchers could not control. For example, one graduate student told me about how her results shifted around the time that she adopted a dog. She had done many iterations of this same experiment and had previously struggled to get a strong behavioral response from her mice, but after the arrival of the dog her behavioral responses went up. She hypothesized that the mice were reacting to the smell of the dog on her clothes, since mice are known to change their behavior in response to odors from predator species (Apfelbach et al., 2005).

I have argued that this knowledge, gained through the management of everyday troubles in the laboratory, is best thought of as a by-product of experimental work (Nelson, 2018). In the community that I studied, only rarely did researchers explicitly set out to study the impact of loud noises or diet or predator odor on behavior. Most of their research programs were designed with the aim of producing genetic results, and the numerous small...
insights about environmental contributors to behavior that they gained were incidental to the process of carrying out this research. Knowledge about environmental factors, in this setting, was like the sawdust produced in the process of building a piece of furniture—a by-product of constructing the desired object. The researchers I studied typically did not work up findings about environmental factors for publication. When environmental knowledge circulated, it was in methods papers or through informal channels such as training sessions or conversations between researchers. Researchers might talk about the importance of controlling light levels when helping a colleague set up a new experiment. Or, they might gossip about how other researchers would yell or play loud music in front of the mice, potentially imperiling their experiments by introducing uncontrolled stressors.

Scholars studying scientific practices have long recognized that much is left out of scientific publications, and that non-textual channels for circulating knowledge are a critical component of scientific ecosystems. Collins (1974, 1992, 2001, 2010) has used the concept of tacit knowledge to enumerate a variety of reasons why knowledge may not appear in the scientific literature: it may be difficult to communicate in writing or photographs, it may be unrecognized by the researchers themselves and unconsciously transmitted through observation and apprenticeship, scientists may wish to conceal it to gain a competitive advantage, and so on. A large literature now exists on the role of tacit and informal knowledge in a wide variety of knowledge communities, from experimental psychology (Brenninkmeijer et al., 2019) to sound engineering (Horning, 2004) to technology transfer (Grimpe & Hussinger, 2013).

The argument I am making is that epistemic by-products are not unrecognized or unable to be conveyed in words; they are undervalued. There is nothing intrinsic about knowledge of environmental contributors to behavior that prevents researchers from studying it systematically and publishing scientific papers about it. When knowledge of the laboratory environment connects to valued goals such as the improvement of human health, then researchers can and do expend the effort needed to transform unsystematic observations into findings that can be published, as in the case of the decades-long project funded by the National Institute of Alcohol Abuse and Alcoholism on alcohol abuse and stress (Grant, n.d.). Much like the sawdust produced in the process of building furniture, the status of knowledge is contextual—in some markets sawdust might be a waste product that businesses pay to dispose of, while in others it might be a valuable product in its own right. The value of epistemic by-products is thus influenced by individual researchers’ judgements about what is worthwhile to know, and by broader factors such as societal beliefs about the capacity for genetic research to improve human health and the generous funding that accompanied those beliefs, particularly in the early days of the Human Genome Project (Lippman, 1992).

**Reasons to attend to the selective publication of epistemic by-products**

It is tempting to think that the lack of epistemic by-products in the published literature is no great loss. By definition this knowledge tends to be anecdotal rather than systematically collected, and so it is typically low-quality evidence that would not provide a strong foundation for supporting future action. The fact that researchers (or funders) did not see enough value in the information to design or support studies to collect it systematically also implies that there is little to be gained by including them in the published literature. Unlike knowledge about drug treatments, which has a clear connection to the socially-valued goal of improving human health, knowledge about how the smell of a technician’s pet dog can change mouse behavior seems far less consequential.

Both analysts of scientific practices and scientists themselves acknowledge that this knowledge does have one important role: it is often critical for getting experiments to work. Collins (1974) first drew attention to the necessity of face-to-face connections for getting experimental setups running in his ethnographic study of TEA lasers. He showed that not one of the research groups that he studied was able to build a functional laser based on the published literature alone; the groups that had success all had access to sources of informal/tacit knowledge that were critical for getting the laser to work. The researchers that I studied likewise saw knowledge about environmental influences on behavior as something that they needed to know about in order to produce the genetic findings they valued. In conversations about replication in experimental psychology, scholars have argued further that omitting informal/tacit knowledge from the methods sections of published papers is problematic because it hampers efforts to replicate results (Brenninkmeijer et al., 2019). There is growing evidence to support the argument that replication efforts are more likely to be successful when published methods are supplemented by interactions with the original authors (Chatard et al., 2020; Klein et al., 2019). All of these analyses assign value to informal/tacit knowledge or epistemic by-products, but portray their value as merely instrumental: they are useful only insofar as they aid in the production in other kinds of knowledge.

There are two additional reasons why we should consider selective publication of epistemic by-products to be an important problem. The first is that uneven access to this knowledge further exacerbates the problems of misplaced certainty that scholars studying publication bias have already identified. Just as the omission of studies with null or negative results from the published literature can make interventions seem more effective than they actually are, too can a lack of access to epistemic by-products result in an inflated sense of certainty. In the laboratories that I studied, researchers drew heavily on their personal experience with conducting animal behavioral experiments when evaluating the quality of similar experiments in the published literature. They used their knowledge of how environmental factors altered mouse behavior to fine-tune their sense of confidence in findings, mining the methods sections of papers for clues about how well-controlled the experiments were. Their environmental knowledge also allowed them to make sense of incongruent findings. By identifying subtle variations in protocols that could explain the differences in outcomes, a literature that appeared at first glance to be highly contradictory could be made more coherent.

Readers who have no personal experience of conducting behavioral experiments or who did not have access to the knowledge circulated through informal channels would likely arrive at different conclusions when examining the published literature. This phenomenon has been explored by MacKenzie (1998) in his study of intercontinental ballistic missile technology. He compared the levels of confidence in the technology expressed by those who had intimate familiarity with the missiles (e.g. their designers), those who were committed to the technology but less directly involved (e.g. senior management in the
firm producing the technology, users of the technology), and those who were opposed to the technology (e.g. activists calling for disarmament, proponents of another technology). MacKenzie found that those who expressed the highest degree of confidence in the technology were not the researchers who had intimate familiarity with it; it was the managers and users of the technology, who were supportive but distant from the site of knowledge production. Those who lacked the “insider” knowledge about potential sources of inaccuracy in the missile data were more likely to take the data at face value and to have greater trust in the missiles than those who designed them. This suggests that biases against publishing epistemic by-products have an effect similar to that of biases against publishing null findings—they both increase the risk that users of a technology or a technology will have an inflated sense of confidence in those results.

The second reason to care about the selective publication of epistemic by-products is that these by-products play an important role in shaping individuals’ understandings of natural phenomena. A lack of access to epistemic by-products can result in a distorted or overly simplistic view of the factors contributing to a particular phenomenon. In observing animal behavior genetics researchers as they went about their work, I came to realize how much epistemic by-products contributed to their understanding of behavior. Day in and day out, they worked with dozens of genetically identical mice who differed in how much alcohol they chose to drink, how they performed on behavioral tests, or even what color of fur they developed. The experience of reading off widely varying numbers from alcohol bottles on rows of cages containing genetically identical animals was a powerful reminder that genetics alone could not explain behavior. This is not to say that the published literature was unimportant—when describing their views on behavior, researchers often quoted data from human behavior genetics studies showing that genetic and environmental factors both mattered in developing alcohol use disorders (Prescott & Kendler, 1999), or mouse studies showing that the genetic makeup of the mice influenced how much they would drink (Rhodes et al., 2007). But the experience of working with mice in the laboratory also contributed substantially to how researchers viewed behavior, and in particular to their beliefs that subtle changes in the environment could have a strong effect on how much an individual drank.

These same researchers complained that the non-scientists they interacted with had inaccurate, all-or-nothing views about genetics and behavior. Not infrequently, they made similar complaints about other scientists within their field, and even other scientists within their own department. Some of their friends and family members believed that alcoholism was a choice and that inherited predispositions had nothing to do with it. Others believed that a family history of alcoholism meant that their children were destined to become alcoholics and needed to abstain from drinking at all costs. And their colleagues tended to be too quick to assume that any differences in drinking they saw in their animals could be attributed to genetics, rather than considering the role of environmental factors. From the point of view of the scientists that I worked with, both members of the lay public and scientists from outside their subfield frequently misunderstood the important contributions that both genes and environments made in the development of behavioral disorders.

Examining the selective publication of epistemic by-products could help explain these gaps. The firsthand knowledge of mouse behavior that was important for my researchers’ understandings of behavioral disorders was accessible only to a few people—those doing the experiments themselves, and those who were part of the informal channels through which these by-products circulated. Members of the lay public would have little access to this kind of experiential knowledge about how genetically identical mice differed in their behavior. Even some of their colleagues in the same department lacked access to this knowledge, if they were performing different kinds of experiments and not part of the informal channels for sharing this lab lore. To the extent that other scientists’ and lay public understandings about behavior were informed by this research, they were based on a biased sample of what the researchers that I worked with knew—outsiders only had access to the (largely genetic) findings that my researchers considered valuable enough to be published in journals, touted in press releases and presented in public talks. Of course, non-scientists also draw on their own life experiences in forming their views about behavioral disorders, and scientists’ and non-scientists’ views are not necessarily as divergent as the scientists I studied believed they were (Richards, 2006). However, looking at differences in access to epistemic by-products may help to explain divergences in understandings of natural phenomena where they do arise between scientific communities, between researchers at prestigious institutions and disadvantaged ones, between early career researchers and experienced ones, or between scientists and non-scientists.

Conclusions and potential solutions

As noted previously, the bias against funding, systematically collecting, and publishing knowledge about environmental contributors to behavior and human health is driven by individual and collective assumptions about what is valuable to know. Other analysts have written at length about the potential harms arising from the values underlying the processes of selection I have described here. Lippman (1992)’s well-known argument about geneticization, written in the early years of the Human Genome Project, expressed concern that researchers and funders were devoting so much time and money to studying genetics as a means of addressing illness, rather than focusing on social determinants of health. She argued that this single-minded focus on genetics was also colonizing and co-opting public thinking and debate about health. Evelyn Keller (2000) has similarly critiqued “gene talk,” and has argued that the privileged place of genetics in both scientific research and public discourse limits possibilities for thinking and action.

The view I have presented is somewhat more optimistic: my research shows that even in settings where scientists are focused on genetics, they can learn quite a bit about the importance of environmental factors and develop an appreciation for the many, intertwined contributors to health and illness. It is not necessarily as divergent as the scientists I studied believed they were (Richards, 2006). However, looking at differences in access to epistemic by-products may help to explain divergences in understandings of natural phenomena where they do arise between scientific communities, between researchers at prestigious institutions and disadvantaged ones, between early career researchers and experienced ones, or between scientists and non-scientists.
Identifying publication bias against epistemic by-products and other undervalued forms of knowledge is more difficult than identifying bias against quantitative findings, as is identifying solutions for this problem. Unlike bias against publishing negative findings, the bias associated with other forms of selective publication is far more difficult to quantify or visualize. A funnel plot will be of no help in identifying the knowledge of important caveats and limitations to results that their producers picked up along the way. Funnel plots and other quantitative analyses of publication bias direct attention only towards the main outcome measures of a study. The omission of incidental findings gained during the process of conducting the research will be entirely invisible in these plots. Common policy solutions for combating publication bias, such as pre-registration of research plans, would be similarly ineffective.

Pre-registered protocols describe the findings that researchers intend to produce at the outset of their research and the means by which they will produce them. But epistemic by-products are, by definition, things that researchers do not intend to find, and so registering a research plan in advance would not help identify or communicate this missing mass of knowledge.

One potential solution to combat the selective publication of epistemic by-products is to promote scientific pluralism. Researchers who self-describe as geneticists may see environmental knowledge only as a means to an end, but for other researchers, the impact of environmental stressors on the development of behavioral disorders may be interesting in their own right. Longino (2013) argues that the behavioral sciences already exhibit a pluralistic structure, with different groups of researchers “parsing the causal space” of behavior differently: while molecular behavior approaches aim to create genetic knowledge by holding environmental factors constant, social-environmental approaches focus explicitly on the environment as a causal factor, and integrative approaches take the interaction of genes and environments as their topic of interest. Pluralism embraces the idea that one persons’ by-product might be another person’s valued product, and that by diversifying the scientific ecosystem we will capture a broader range of knowledge in the published literature. This approach is not without its limitations, however. One limitation is that the balance of researchers and research approaches in a given field will be impacted by the same values that drive processes of selective publication. Collective assumptions about what is valuable to know pose barriers for individual researchers seeking funding or publication opportunities for research on undervalued topics. Another limitation is that scientific pluralism does not address the information loss that happens through selective publication. Pluralism assumes that the environmental knowledge not communicated by one group of researchers will be re-discovered, formalized, and communicated by another group. While this process may eventually result in a balanced portfolio of published knowledge, it is inefficient because it requires re-discovering knowledge already held by others.

Making data publicly accessible is another potential means of addressing the selective publication of epistemic by-products. The Mouse Phenome Database, for example, invites researchers to submit data on phenotypic measurements on commonly used mouse strains, along with protocol information about the conditions under which those data were collected (Bogue & Grubb, 2004). Researchers often collect this kind of data to use as a comparison for their intervention groups, and the Mouse Phenome Database acts as a kind of epistemic recycling bin for what is typically single-use data, making it available for new kinds of analyses. Researchers might collect baseline data on the body weight of their control mice to compare with the body weights of genetically manipulated mice, and aggregating this data and making it public allows researchers to ask questions about how the baseline body weights of a mouse strain vary across laboratory environments. Many open science/open data efforts of this kind, designed to promote data reuse, are emerging across the life sciences (Leonelli, 2016). However, these initiatives work best for structured, quantitative data that researchers already record because they recognize it as having some value, even if that value is only as a foil for making other phenomena visible.

The approach taken by the Journal of Trial and Error offers a potentially more expansive solution to these challenges, one that integrates the strengths of multiple other approaches. Creating a venue specifically targeted towards insights that arise through the process of trial and error encourages researchers to communicate their by-products. The flexible online format allows for the dissemination of both structured data and qualitative findings. And by pairing original reports by scientists with commentaries from humanists, JOTE avoids the problem of information loss associated with other forms of pluralism: it encourages the formation of multiple accounts of the same data, hopefully allowing scientists to see new forms of value in what they have produced. I look forward to seeing what kinds of unwieldy, undervalued findings will be collected in its pages over the coming years, and to the insights that will come from them.

References
Chatard, A., Hirschberger, G., & Pyszczynski, T. (2020). A word of caution about many labs 4: If you fail to follow your preregistered plan, you may fail to find a real effect. https://osf.io/ejubn

License

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
CrowdDetective: Wisdom of the Crowds for Detecting Abnormalities in Medical Scans

Veronika Cheplygina

Abstract

Machine learning (ML) has great potential for early diagnosis of disease from medical scans, and at times, has even been shown to outperform experts. However, ML algorithms need large amounts of annotated data – scans with outlined abnormalities – for good performance. The time-consuming annotation process limits the progress of ML in this field.

To address the annotation problem, multiple instance learning (MIL) algorithms were proposed, which learn from scans that have been diagnosed, but not annotated in detail. Unfortunately, these algorithms are not good enough at predicting where the abnormalities are located, which is important for diagnosis and prognosis of disease. This limits the application of these algorithms in research and in clinical practice.

I propose to use the “wisdom of the crowds” – internet users without specific expertise – to improve the predictions of the algorithms. While the crowd does not have experience with medical imaging, recent studies and pilot data I collected show they can still provide useful information about the images, for example by saying whether images are visually similar or not. Such information has not been leveraged before in medical imaging applications. I will validate these methods on three challenging detection tasks in chest computed tomography, histopathology images, and endoscopy video.

Understanding how the crowd can contribute to applications that typically require expert knowledge will allow harnessing the potential of large unannotated sets of data, training more reliable algorithms, and ultimately paving the way towards using ML algorithms in clinical practice.

KEYWORDS
machine learning, artificial intelligence, medical imaging, crowdsourcing, computer-aided diagnosis

Research proposal

<table>
<thead>
<tr>
<th>General information</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Institution of employment at time of application</strong></td>
</tr>
<tr>
<td>Eindhoven University of Technology</td>
</tr>
</tbody>
</table>

| **Prospective host institution** |
| Eindhoven University of Technology |

| **Main field** |
| Artificial Intelligence, Expert Systems |

Other fields

| Medicine, other |

Number of words

| Description of the proposed research: 1999 (max. 2,000 words) |
| Knowledge utilisation: 719 (max. 750 words) |

Description of the proposed research

Overall aim and key objectives

Machine learning (ML) has seen tremendous successes in recent years, for example in classifying everyday objects such as cats in images. This progress
is driven by the availability of powerful computational resources, sophisticated algorithms, but most of all, large amounts of labelled data. Data is annotated at different levels: outlining the cat in the image (a strong annotation) or tagging the whole image (a weak annotation). Strong labels are the most effective, but weak labels can also be used by weakly supervised or multiple instance learning (MIL) algorithms (Fig.1).

In medicine, ML offers invaluable opportunities for diagnosis of disease, reaching expert performance [1] or even outperforming the experts [2, 3]. However, while the amount of medical data is growing rapidly, the data is often only weakly annotated (a scan with a diagnosis, but without outlined abnormalities). Although MIL is gaining popularity in medical imaging [4–9], there is an overlooked, fundamental problem. MIL algorithms are optimized to predict weak annotations [10], but the classifier best at predicting weak annotations, is often not the best at predicting strong annotations [6, 8, 11, 12]. In practice this means that without strong annotations, MIL algorithms are poor at localizing abnormalities [13].

I propose a novel way of using crowdsourcing [14] to improve the abnormality localization. Crowdsourcing is the process of gathering input by enlisting the services of a large number of people, either paid or unpaid, typically via the internet, allowing large-scale data annotation [15]. A widely used example is tagging people on social media, used to improve face recognition. Crowdsourcing has therefore been successful in computer vision [15, 16] with non-expert tasks, such as recognizing everyday objects. However, internet users are not medical experts – how could they annotate abnormalities? A key insight is that the crowd does not need to mimic the experts to improve MIL for detecting abnormalities. For example, for detecting abnormalities in chest CT images, I propose to instead leverage the human visual system by asking:

- To outline airways, which are easily recognizable structures (Fig.2)
- Whether patches are similar to each other

These annotations can be intuitively provided by the crowd, as demonstrated by my pilot results [17, 18]. Although such annotations cannot be used to directly train algorithms, they are still informative.

The visual information in these annotations can help the MIL algorithm to find better representations for the data via multi-task learning [19] with labels for related tasks, such as outlining airways, and with similarity-based learning [20, 21] with patch similarities.

The goal of this project is to improve the prediction of strong annotations by MIL, which is important for medical imaging, but also applications in other fields where annotations are scarce. Furthermore, the project will provide insight into the value of crowdsourcing in expert-level tasks. This is essential to leverage the value and scientific potential of big datasets, which are growing at an exponential rate.

**Originality and innovative aspects.** The problem of predicting strong labels with MIL is underexplored [10, 12]. Using crowdsourcing to improve MIL is novel in general - in a recent survey of crowdsourcing in computer vision [16] none of the 195 papers address MIL, while a recent book on MIL [22] does not mention crowdsourcing. While techniques for multi-task and similarity-based learning exist, these often have assumptions not compatible with medical images annotated by crowds. I will develop methods which better address such data.

Crowdsourcing is new to medical imaging, with most studies published in the last five years [17, 23–27]. The crowd typically mimics the experts in labelling images with a diagnosis or outlining structures in image patches, and the results are compared to expert labels. Several studies show promising results, while in other cases the crowd does not achieve the level of the experts [25, 27]. Note that in most studies the crowd only labels the images – the labelled images are then not used in ML approaches. Only [27] uses the labels to train algorithms, but collects labels and does not address the MIL scenario. There is no prior work on collecting labels for related tasks and similarities, except in my recent pilot studies [17, 18].

Next to my expertise in these topics I started building a community around the topic of crowdsourcing for medical imaging (MICCAI LABELS workshop, recently funded eScience-Lorentz workshop), have many international collaborators and interest from industry. The project also benefits from my outreach and open science efforts (Section on Knowledge Utilisation).

**Approach.** I will investigate three strategies in which expert weak labels and crowd annotations can be combined to train more robust MIL algorithms:

- Experts + crowd labels for the expert task
- Experts + crowd labels for related tasks
- Experts + crowd similarities

My hypothesis is the best strategy will depend on the application, as ex-

---

Labels from the crowd can be combined with expert labels to create a more robust classifier by decision fusion (DF) [28, 29]. In traditional DF approaches each classifier is trained on a (random) subset of the data [30–32] – classifiers of similar quality, and their decisions are averaged. An alternative is to weight classifiers by their estimated accuracy, which requires additional data or assumptions, such as that per image, at least half of the classifiers are correct. This assumption may not always hold in medical images. In more difficult tasks where expert knowledge is needed, I expect that this strategy will not outperform the baseline, where the MIL algorithm is trained only on the expert weak labels.

Labels for related tasks, such as outlining airways, when the target is to detect abnormalities, can be leveraged via multi-task learning (MTL) [19]. The intuition is that there are underlying features helpful for both tasks, which helps to find a more robust classifier. This is achieved by optimizing a joint loss function for the classifier. Such approaches are used in computer vision to simultaneously detect different everyday objects in images [33]. In medical imaging MTL was used to simultaneously predict an Alzheimer’s diagnosis and a cognitive test score [34], and to simultaneously predict different abnormalities related to lung cancer [35].

However, the computer vision approaches have assumptions that might be incompatible with medical imaging, such as well-defined boundaries for each object, and the medical imaging methods do not address MIL and expect expert labels for both tasks. I will investigate the applicability of these and similar approaches, while addressing the specifics of crowd-annotated medical images, for example, by weighting the loss function such that the expert labels are given more emphasis. I expect that the MTL strategy will be most successful when the related task is simple enough for the crowd to do, but has some relevance to the target task.

Similarities, such as “image patch A is more similar to B than to C” (Fig. 3) can also add information during training. The idea is that such judgements are more intuitive for crowds to provide, even if they are not able to label the images. The similarities provide relative information about the expected classifier outputs, favouring classifiers for which these relationships hold. Examples include metric learning [20] and similarity embedding [21, 36], which I group under the name similarity-based learning (SBL).

Again, the assumptions of current methods that focus on MIL may not be applicable. For example [37] assumes that an image label can apply to at most one patch in that image, which is not true for many abnormalities. Based on my experience with different (medical and non-medical) MIL problems [38, 39], I will investigate how to create approaches with assumptions more suitable for medical images. I expect SBL to be successful when the crowd can focus on a single aspect of similarity, such as image texture. If there are multiple aspects to focus on, the similarities are likely to be inconsistent across annotators, and might hurt performance of MIL instead.

I will develop MIL algorithms which learn both from expert weak labels and crowd annotations, via DF, MTL or SBL. The strategies I will develop are general, therefore I will use them both with traditional [40, 41] and more recent [39, 42] MIL classifiers. Similarly, the strategies will apply both to traditional features and features extracted by deep learning.

To investigate which strategy is better in which scenario, I will apply these algorithms to three applications, with different image characteristics and task difficulty:

- Localizing patches with emphysema, in chest CT images [43, 44]. This is a task of medium difficulty, where I expect best results from MTL because the underlying disease affects the appearance of airways, or SBL based on recent pilot results [18].
- Localizing different surgical instruments in endoscopy videos [45]. This is a relatively easy task where DF is already likely to be successful.
- Localizing cancerous regions in histopathology images [46]. This is a difficult task, where I would expect best results for SBL, or MTL if a simple but powerful related task can be defined in collaboration with experts.

The datasets have expert weak labels, but also some expert strong labels available, which I will use for validation purposes. I will collect annotations via two platforms, with different trade-off between collection time and quality [47]:

- Amazon Mechanical Turk with paid crowds, which reduces the time to 1–Rajpurkar2017 days, but could reduce quality.
- Zooniverse with unpaid volunteers interested in science, which increases time, but also quality.

The next step is to decide which patches to annotate and how many annotators to assign to each patch. Combining annotations increases the quality but, given the same budget, decreases the total amount of annotated data. Because ML (and thus also MIL) algorithms benefit from seeing a larger set of data and my goal is not to produce expert-quality annotations, I expect it will be more valuable to maximize the amount of annotated patches. To keep the number of annotations within budget, I will investigate active selection of patches, for example based on their uncertainty or diversity according to the MIL algorithm.

Lastly I will compare the strategies to baselines, such as training MIL only...
with expert weak labels, and to each other. In this process I will vary the number of expert labels and crowd annotations used, in order to create more scenarios than the three applications investigated.

By analysing these results from three different applications, I will aim to extract general rules on when each strategy is best to use.

Research plan
The project has four work packages:
WP1: Decision fusion with labels
WP2: Multi-task learning with related labels
WP3: Representation learning with similarities
WP4: Generalization

In WP1-WP3, I will crowdsource the corresponding type of annotations and develop MIL algorithms for combining these annotations with expert weak labels. The deliverables are extended versions of my publicly available MIL toolbox [48] and papers published in high impact conferences and journals. In WP4 I will compare the strategies and draw up recommendations for their use. The deliverable is a paper with a related blog post.

I will conduct the research at the Medical Image Analysis group, TU/e. The main collaborators (I will visit both for 3 months are

- Prof. Lena Maier-Hein (DKFZ, Germany). I will collaborate with her on crowdsourcing aspects, and apply my algorithms to endoscopy video data from her group.
- Prof. Eric Granger (ETS, Canada) and Dr. Luke McCaffrey (McGill, Canada). I will collaborate with Prof. Granger on MIL and apply my algorithms to histopathology data provided by Dr. McCaffrey.

I also have contacts in medical imaging and crowdsourcing communities whom I could approach for advice, including Dr. Marleen de Bruijne (Erasmus MC), Dr. Javed Khan (TU/e), Dr. Alessandro Bozzon (TU Delft).

<table>
<thead>
<tr>
<th>Knowledge utilisation</th>
</tr>
</thead>
</table>

Potential
Research. In medical imaging the results of my research will be beneficial to many research groups, as the annotation problem is present in many different applications. In machine learning (ML) the research will demonstrate the potential of improving algorithms with crowd annotations which are not typically leveraged. This is an underexplored area, but is likely to spark interest in the community. I expect this to assist in development of (general purpose) algorithms focusing on such annotations.

Furthermore I expect interest from other applications where annotated data is scarce, such as remote sensing [49] and ecology [50]. Lastly, the behaviour of the crowd (who contributed what and why) could be of interest to research groups in human-computer interaction [47] and social sciences.

Industry. There is already great interest from the industry in ML, and medical imaging is catching up, for example in 2017 the leading conference MICCAI had IBM, Nvidia, Siemens and many medical imaging startups as sponsors.

The interest in my project is demonstrated by the companies in Table 2, including IBM Research. These companies could integrate the research outcomes in their products, which are designed to be user-friendly and secure. This will enable translation of the proposed research to products and (in case of medical imaging) the clinic.

Society. The research area is relevant to several questions on the National Science Agenda, including health (questions 105, 89, 94) and society and technology (112, 108). In health, there are long term benefits of improved medical imaging algorithms, such as better prognosis and diagnosis of disease, and better use of the experts’ time.

For society, there are benefits in involving the public as annotators, which can raise awareness about health. In the long term, contributing to such projects could create jobs or volunteering opportunities, accessible to people who are unable to work due to health or care responsibilities [51], increasing their well-being.

Implementation
Research. I will publish in relevant venues (MICCAI, Pattern Recognition) and make the papers/data/code available online. I will continue organizing events I’m already involved in (MICCAI LABELS, eScience-Lorentz workshop, NVPHBV) and giving talks about my work. In years 1 and 2 my research visits to Germany and Canada will help in reaching other groups I am not yet in contact with. I expect other groups will use my methods for their data from the first year of the project.

In year 2 I will organize a workshop on the intersection of ML and crowdsourcing in label-scarce applications beyond medical imaging. I will invite
researchers via my network and online communities. I’m a member of (ml-news, crowd-hcomp). The impact is more difficult to predict, but I expect that at least some groups will be interested in either the data and/or code generated by my project, and translating these to their own applications.

**Industry**. I have established a user group (Table 2) of industry representatives interested in the research outcomes. I will organize meetings with this group **every year**. Translation to industry could start within 2-3 years of starting the project. This collaboration could lead to other joint projects, ensuring impact even after completion of the current research.

<table>
<thead>
<tr>
<th>Group</th>
<th>Contact</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM Research</td>
<td>(anonymized)</td>
<td>IBM Research is working on a crowdsourcing solution for “internal crowds” (e.g., colleagues) and is interested in a medical imaging application.</td>
</tr>
<tr>
<td>Thirona, ClinicalGraphics</td>
<td>(anonymized)</td>
<td>Thirona and ClinicalGraphics develop software for medical imaging. They train employees for annotating images, and want to scale up the annotations without an equal increase in cost.</td>
</tr>
<tr>
<td>Cosmonio</td>
<td>(anonymized)</td>
<td>Cosmonio develops an app that allows interaction between the medical expert and the algorithm. My research will help to optimize the type of interactions needed.</td>
</tr>
</tbody>
</table>

**Table 2** User Group

**Society**. I will reach a broader group of people (people with an interest in science and/or annotators from AmazonMT and ZooUniverse platforms) via outreach through my blog (3K visits per month) and Twitter (1.8K followers). I will blog about my project **every quarter**, explaining my project to an audience interested in science, but without a technical background.

In **year 3** I will record lectures for an online course on DataCamp, where I am already setting up an image analysis course. The course will combine machine learning, medical imaging and crowdsourcing, but will not require technical prerequisites. This is a more long-term strategy than the others.

**Cost estimates**

| Total budget requested | £250,000 |

**Intended starting date**

September 1, 2018

**Application for additional grants**

No

**Data management plan**

**Will data be collected or generated that are suitable for reuse?**

Yes. I will collect annotations (labels and similarities) for patches extracted from already available medical images. The annotations will be stored in the widely reusable JSON format.

I will also generate features – numerical data which describes the images, but from which the images cannot be reconstructed. The features will be stored as CSV files to facilitate reuse across different platforms.

The annotations and features (hereafter referred to as relevant data) can be reused in development of other machine learning algorithms.

**Where will the data be stored during the research?**

All data is stored electronically. The annotations will be collected via the cloud (Amazon Mechanical Turk or Zooniverse) and thus a secure back-up will always be available. The annotations will then be copied to secure network drives available at the host institute. This is also where the generated features will be stored. Daily back-ups are made on this storage facility.

Upon publication of a paper, I will also upload the relevant data to the Figshare repository (see below).

**After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties? To whom will the data be accessible?**

I will store the relevant data on Figshare which will ensure its long-term preservation. Figshare has agreements with publishers such as Nature, PLOS, and others to ensure the data persists for a minimum of 10 years. I will share the relevant data under the Creative Commons Attribution-NonCommercial-ShareAlike (CC-BY-NC-SA) license.

**Which facilities (ICT, (secure) archive, refrigerators or legal expertise) do you expect will be needed for the storage of data during the research and after the research? Are these available?**

*ICT facilities for data storage are considered to be resources such as data storage capacity, bandwidth for data transport and calculating power for data processing.

I will use in-house computing facilities, which are already available at the host institute.

2. The author now distances herself from this company. She would not use their software after an investigation has shown an incident of sexual misconduct from the former CEO and the ensuing mismanagement of this incident. For transparency, the editors are maintaining the company name.

3. Conferences, outreach, publication, and travel
Ethics

Use of extension clause
No

Ethical aspects

| Approval from a recognised (medical) ethics review committee | Not applicable |
| Approval from an animal experiments committee | Not applicable |
| Permission for research with the population screening Act | Not applicable |

**TABLE 4** Ethical aspects

**Declarations.** By submitting this form I endorse the code of conduct for laboratory animals and the code of conduct for biosecurity/possibility for dual use of the expected results and will act accordingly if applicable.

☐ I have completed this form truthfully.
☐ By submitting this document I declare that I satisfy the nationally and internationally accepted standards for scientific conduct as stated in the *Netherlands Code of Conduct for Scientific Practice 2014* (Association of Universities in the Netherlands)
☐ I have submitted non-referees.

**Name:** Veronika Cheplygina  
**Place:** Eindhoven  
**Date:** 8 January 2018

Society

**Public summary**

Crowds as medical detectives (ENG). *Dr.ir. V. (Veronika) Cheplygina (v), TU/e – Computer Science*

Detecting abnormalities in medical images is essential for diagnosis and treatment of illness. Computer algorithms can learn to do this using manually annotated scans, but the annotation process is costly for experts. This project studies how annotations made by untrained internet users can improve the detection accuracy of computer algorithms.

**Internetgebruikers als medische experts? (NL).** *Dr.ir. V. (Veronika) Cheplygina (v), TU/e – Informatica*

Het detecteren van afwijkingen in medische scans is belangrijk voor diagnose en behandeling van de ziekte. Computeralgoritmos kunnen dit leren van handmatig geannoteerde beelden, maar dit kost experts veel tijd. Dit project onderzoekt hoe annotaties van gewone internetgebruikers de automatische detectie van afwijkingen kunnen verbeteren.

Reviews

Grant

Vernieuwingsimpuls Veni ENW 2018

**Title**

CrowdDetective: wisdom of the crowds for detecting abnormalities in medical scans

**Applicant**

Dr. ir. V.V. Cheplygina

**File number**

016.Veni.192.066

Your rebuttal on the referee reports is in progress

Referee report of referee 1

**Assessment of the quality of the researcher.**

**Explanation.** Criteria - Quality of the researcher:

- in terms of profile fit in the target group;
- from an international perspective belongs to the 10 to 20% of his/her peer group;
- academic excellence as demonstrated by the PhD thesis, publications and/or other relevant achievements in the field
- inspiring enthusiasm for research and/or technology;
- persuasiveness;
- clear indications of an outstanding talent for academic research.

The Veni scheme aims at outstanding researchers only: the top 10-20% of his/her international peer group.

**Question a:** What is your opinion on the past performance of the researcher (as demonstrated by his/her doctoral thesis, publications, and other relevant scientific achievements)?

**Comments:** The applicant shows a clear focus in her research on multiple instance learning (MIL), one of the key components of the present application. She has several years of postdoctoral experience dealing with the application of MIL related algorithms to tasks in medical image analysis. I see some publications in journals, such as Pattern recognition and Pattern recognition letters, one Miccai paper. She has co-organized a workshop at the same conference, dealing with data annotation and crowd sourcing, the other key component of the present application. She also organized a workshop at ICML, and lists several international research visits. After only two years of national postdoc experience, she became assistant professor.

Overall, the applicant presents herself with a profile dedicated to an academic career, a driving research interest, all aligned well with the present research proposal. As a minor grain of salt I would have hoped for (more) high impact publications, i.e., either highly cited papers (and 8 years after the MSc there may be one or two highly cited papers, whatever the journal or conference), or contributions to high visibility conferences, such as ICML, NIPS, CVPR, or a second or third MICCAI paper.

**Question b:** Does the applicant belong to the top 10-20% of his/her international peer group? Which scientific achievements or talents of the applicant show he/she belongs to this top?

**Comments:** I see a number of activities, such as international collaborations and research visits, co-organization of events and workshops that actively shape the discussion of the research community. She actively disseminates research
ideas via new media with followers being interested in her views and opinions. I think this does set her apart from many of her peers. There are some noticeable publications, although with a 5 year PhD and 3 years of postdoc (and assistant professor) level research, a number of other researchers in the field might have a stronger publication record. Depending on how to weigh both aspects she may be among the top 20% of her peers.

Assessment of the quality, innovative character and academic impact of the proposal

**Explanation:** Criteria - Quality, innovative character and academic impact of the proposed research

- challenging in terms of content;
- originality of the research topic;
- innovative scientific elements;
- potential to make an important contribution to science;
- effective in terms of proposed methodology.

**Question a:** Please comment on the relevance of the problem and on the originality and challenging content of the proposal.

**Comments:** The problem of how to include crowd-sourced expert and non expert annotations is a relevant problem in machine learning and, hence, in medical image processing research. The problem is not solved yet and any solution would have the potential to impact significantly on the design and dissemination of machine learning in diagnostic clinical image analysis. Still, the proposed project could be stronger if it would not only focus on the comparison of different (more or less) existing techniques to a few selected (and more or less well defined) problems, but would promise to contribute to the advancement of related machine learning algorithms itself. I would see that generating these algorithms may be a natural second step, but - as it is - the application only promises to “investigate three strategies” that are likely to be data set dependent (as the applicant suggests). Overall, I feel the application promises to deliver solid and systematic research that, however, is far from offering new innovative concepts and contributions to the field.

**Question b:** What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?

**Comments:** The main contribution will be a systematic comparison of different analytic strategies on different data sets. As such, it promises some ‘best practice’ guidance in a field that would, indeed, would benefit from such systematic research.

**Question c:** What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?

**Comments:** see 2.b)

**Question d:** To what extent is the proposed method effective? Please comment.

**Comments:** The research strategy is well described, and the aims the applicant is presenting are likely to be reached.

Assessment of the knowledge utilisation

**Explanation:** Criteria - Knowledge utilisation (= KU)

NWO uses a broad definition of KU; not only innovative end of pipe product creation is considered, but also purposeful interaction with (potential) knowledge users, such as industry or public organisations and knowledge transfer to other scientific disciplines. NWO asks applicants to fill out the paragraph on KU. An applicant may however explain that KU cannot be expected given the nature of the research project. In that case, we still kindly ask you to assess whether the applicant has provided reasonable argument or evidence sufficiently.

**Potential:**
- contribution to society and/or other academic areas;
- disciplines and organisations that might benefit from the results.

**Implementation:**
- action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- if and how the potential knowledge users will be involved;
- (concrete) outcomes for society;
- the period over which KU is expected to occur.

**Question a:** What is your opinion on the described relevance of the results of the research?

**Comments:** In the best case the project will help paving the road for simplifying research in medical image computing and the translation of medical image computing technology into clinical practice. The outcome of the project, i.e., a description of the optimal strategy for structing diagnostic information linked to a given medical image set, can be used both in the design, or evaluation, or continuous quality control of these technologies. As such, I would consider the overall research direction to be interesting and relevant also from a wider perspective.

**Question b:** Please comment on the effectiveness and feasibility of the proposed approach for knowledge utilisation.

**Comments:** Data sets and best practice recommendations, together with related algorithms, will be the promised output: “By analysing these results from three different applications, I will aim to extract general rules on when each strategy is best to use”. Whether these general rules exist, will only be known upon completion of the projection. I am somewhat missing a ‘basic methodological research’ component in the research objectives, for example, exploring one particular machine learning algorithm for MIL on top of the promised systematic comparison (whether this is the ubiquitous deep learning or any other). Similarly, a ‘driving clinical problem’ that would be solved at the end of the project, would have been nice as well (e.g., new solutions to one interesting problem that would be relevant - whether these rules generalize to other tasks, or not).

**Question c:** Only answer this question in case the applicant argued that knowledge utilisation is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why knowledge utilisation is not applicable for his/her research project (see also the information under criterion 3 listed above)?

**Comments:** Knowledge Utilization is expected.

**Final assessment**

**Question a:** How do you assess the entire application? Please give your final scoring (A+/A/B/UF/U).
Comments: 03, B

Question b: Could you please summarize (point by point) the strengths and weaknesses of the grant application focusing on the candidate, proposal and knowledge utilisation?

Comments: The applicant presents herself with a dedicated and good career path in medical image computing. There is a focus on MIL, that is relevant for the present project, although I feel that publications are cited for systematic comparisons, benchmarks, and implementations than for innovative methodological contributions. The proposed project is very timely and, in case "general rules" can be found, has the potential to have significant impact. It promises these contributions from a systematic study, rather than from new innovative concepts and ideas.

Feedback datamanagement

Question. Feedback datamanagement

Comments: no

Referee report of referee 2

Assessment of the quality of the researcher.

Explanation. Criteria - Quality of the researcher:

▷ in terms of profile fit in the target group;
▷ from an international perspective belongs to the 10 to 20% of his/her peer group;
▷ academic excellence as demonstrated by the PhD thesis, publications and/or other relevant achievements in the field
▷ inspiring enthusiasm for research and/or technology;
▷ persuasiveness;
▷ clear indications of an outstanding talent for academic research.

Question a: What is your opinion on the past performance of the researcher (as demonstrated by his/her doctoral thesis, publications, and other relevant scientific achievements)?

Comments: The researcher has a very good profile, with organising of workshops and interesting publications in good venues. It is a junior researcher so major impact in terms of citations can not be expected. The research direction is in line with her previous work and she has experience with research visits as well.

Question b: Does the applicant belong to the top 10-20% of his/her international peer group? Which scientific achievements or talents of the applicant show he/she belongs to this top?

Comments: She has received recognition by being part of the workshop organisers at LABELs and has had several interesting publications.

Assessment of the quality, innovative character and academic impact of the proposal

Explanation. Criteria - Quality, innovative character and academic impact of the proposed research

▷ challenging in terms of content;
▷ originality of the research topic;
▷ innovative scientific elements;
▷ potential to make an important contribution to science;
▷ effective in terms of proposed methodology.

Question a: Please comment on the relevance of the problem and on the originality and challenging content of the proposal.

Comments: The problem of limited annotation is an important problem in medical image analysis and a major limiting factor of machine learning in image analysis. Another important factor is the amount of data available, which increasingly gets easier with large repositories such as TCIA and TCGA that were not even mentioned in the text.

Question b: What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?

Comments: Using Crowdsourcing has been done many times in medical image analysis and a few examples are mentioned. The link of weak annotations is somewhat novel but it seems that there are other approaches that would need to be combined. Just getting more data is often not enough and getting the right images annotated, so those that are on the decision boundaries would seem most important. I did not see any strategy of quality control of the crowdsourced annotations and this seems like the major factor that is important.

Question c: What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?

Comments: There is an opportunity to advance the area medical image annotation but to a limited degree with the approaches give if no quality control is done ad if only weak labels are given. MIL is important and finding a link between the annotation and the best approaches could be very interesting.

Pretty much all medical images have reports associated to them, so ignoring the available weak labels would be a pity.

These can be radiology and pathology reports and may be more effective than getting labels of limited quality.

Question d: To what extent is the proposed method effective? Please comment.

Comments: It is very hard to judge if the method will work. Some approaches have been using crowdsourcing in the past and they show that with strong quality control this works well. It is not clear how this will be leveraged by the proposed approached.

Assessment of the knowledge utilisation

Explanation. Criteria - Knowledge utilisation (= KU)

NWO uses a broad definition of KU: not only innovative end of pipe product creation is considered, but also purposeful interaction with (potential) knowledge users, such as industry or public organisations and knowledge transfer to other scientific disciplines. NWO asks applicants to fill out the paragraph on KU. An applicant may however explain that KU cannot be expected given the nature of the research project. In that case, we still kindly ask you to assess whether the applicant has provided reasonable argument or evidence sufficiently.

Potential:

▷ contribution to society and/or other academic areas;
▷ disciplines and organisations that might benefit from the results.

Implementation:
Action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- if and how the potential knowledge users will be involved;
- (concrete) outcomes for society;
- the period over which KU is expected to occur.

**Final assessment**

**Question a:** How do you assess the entire application? Please give your final scoring (A+/A/B/UF/U).

**Comments:** The research has potential to increase clinical decision making if it is working well. However, quality control is not clear how exactly the interaction will be done and how intellectual property rights can be shared. AI in medicine is a hot topic in industry as well, so there definitely is potential if things work well.

**Question b:** Please comment on the effectiveness and feasibility of the proposed approach for knowledge utilisation.

**Comments:** The research proposes contact with industry and has an industrial panel. It is not clear how exactly the interaction will be done and how intellectual property rights can be shared. AI in medicine is a hot topic in industry as well, so there definitely is potential if things work well.

**Question c:** Only answer this question in case the applicant argued that knowledge utilisation is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why knowledge utilisation is not applicable for his/her research project (see also the information under criterion 3 listed above)?

**Comments:** Knowledge utilisation is expected.

**Feedback data management**

**Question.** Feedback data management

**Comments:** The data management only concentrates on the annotations to be collected and not on the raw data. Where will the images originate from? How many are available? Who will test the system and who will generate the ground truth? Is ethics approval for the CT and histopathology images available? By whom? How will it be insured that these data are treated properly? These are human data!

Sharing only annotations without the raw data would have a very limited usefulness.

**Rebuttal**

**File number** 016.Veni.192.066

**Name of candidate** Veronika Cheplygina

**Title** CrowdDetective: wisdom of the crowds for detecting abnormalities in medical scans

I would like to start by thanking the committee and the reviewers for taking the time to provide feedback on my application. Below I first address the overall opinion of the reviewers and then discuss a few specific points. Direct quotes from the reviewers are in red, and direct quotes from my proposal are in blue. The page numbers refer to pages in my submitted PDF, where my research proposal is on pages 4 through 9.

The reviewers are positive about my profile as researcher, mentioning important publications and my leadership role in the community around my research topic. R1 comments that my publication record could have been stronger. I would disagree with this, since several of my publications have been cited at 5 or 10 times the impact factor of the venue. Furthermore, since submitting the Veni, my citations increased from 230 to 254, and my h-index from 9 to 10. Two papers I published in 2017 already have 6 and 4 citations, therefore I would expect the h-index to further increase 12 in 2018. This is exceptional for somebody at my career stage, after 6.5 years in total (not 8, as the reviewer calculated) spent on research.

The reviewers agree that the problem is important, the method has potential and describe the project as well-defined. R1 is concerned with a more innovative contribution of the project, and suggests it would have been better to focus on one specific MIL algorithm, AND on one clinical problem. First, a key innovation of the proposal is to focus on different types of annotations that have been collected from the crowd, which has not been addressed before in medical imaging. Furthermore, I have specifically chosen to focus on a range of methods and applications, providing general guidelines for the field. I would
argue that this is more innovative than developing specific methods for specific applications, which is what is regularly being done at most conferences on the topic.

In contrast to R1, R2 seems to find the proposal too innovative, and suggests it would be better to follow the existing approach – collecting labels from the crowd, and comparing them to expert labels. As I discuss in the proposal, this is likely not to be an optimal strategy. My proposed methods, which focus on alternative (not yet investigated) types of annotations, are more promising in this regard. Since they rely on more intuitive characteristics of the images, the quality control is also less of an issue than suggested by the reviewer. Of course, I will still perform validation, as described on page 7 of the proposal – “The datasets have expert weak labels, but also some expert strong labels available, which I will use for validation purposes”. R2 suggests a number of other improvements, most of which are either addressed in my proposal, or could not be addressed within the scope of the project. I briefly respond to these below:

Active learning. This is indeed an important point to investigate. I mention this in my proposal on page 7 “will investigate active selection of patches, for example based on their uncertainty or diversity according to the MIL algorithm”. Although I did not mention this explicitly, both of my collaborators have recent work on active learning [Carbonneau2017] and the related concept of uncertainty estimation [Moccia2018]. These methodologies can be incorporated in the algorithms I develop.

Existing large repositories such as TCIA and TCGA. TCIA is a repository of genomic data, which is not relevant to my proposal. TCGA could be an interesting resource, but does not provide local annotations, which is precisely what is necessary for validation/quality control. As I describe in the proposal, I choose to focus on three applications for which local annotations are available for validation.

Patient reports. Patient reports provide weak labels for images, and are indeed often the basis of the expert weak labels I have available for my datasets. It is incorrect that I ignore these labels – these are in fact the expert labels my methods will use, in combination with the crowd labels. Processing the patient reports with natural language processing is outside the scope of my research.

Redundancy of expert and crowd labels. The reviewer writes “if expert labels exist than crowd labels do not seem necessary.” This is incorrect. The use of expert weak labels alone leads to unstable MIL algorithms, as I have detailed on page 4, “In practice this means that without strong annotations, MIL algorithms are poor at localizing abnormalities [13].”

Questions on data management. These questions, together with the other comments, suggest that the reviewer has overlooked an entire page of my proposal (page 7), where I discuss the public datasets I will use, and which already have expert labels available for validation. Overall, given the many positive comments of the reviewers, and the fact that several weak points are not justified, I hope that the committee will consider my proposal for the interview stage.

References

Decision
Project number
016.Veni.192.066
Applicant
V. V. Cheplygina

Title
CrowdDetective: wisdom of the crowds for detecting abnormalities in medical scans

Scores
Quality of the applicant: 4.2
Quality of the research proposal: 5.8
Knowledge utilization: 3.9
Final score: 4.8

Qualification: Your research proposal received the qualification “good”, based on the application, the reviewer reports and the rebuttal.

Quality of the Candidate
The committee and reviewers agree that the candidate has a clear research focus and an average to good publication record, although high-impact publications are still missing. One reviewer is therefore hesitant to place the candidate in the top 20% of her international cohort, which is agreed upon by the committee. The candidate was however found by the committee to be an ambitious researcher who has spent a significant amount of time on academic services (workshop and conference organisation, reviewing duties, board member) and outreach.

Quality of the Proposal
The reviewers agree that the proposal tackles a very timely and relevant research topic within medical image processing. They also notice that the methodology is logical, though not overly compelling. One reviewer questions the novelty of the envisioned contributions while another reviewer raises the issue of crowdsourcing quality control that should have been included in the proposal. The committee shares the doubt of the reviewers on these aspects.

Knowledge Utilization
The committee and the reviewers find the knowledge utilization plan convincing. The plan aligns well with the candidate’s prior experiences and targets different audiences with diverse activities. The inclusion of an industry panel is valued by the committee. One reviewer misses more details on issues of intellectual property rights as well as further details on the setup of the industry panel, to which the committee agrees that more details should be provided.
References


**License**

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020
Global Giving

René H. F. P. Bekkers

Abstract
Why do citizens in some countries take more responsibility for the well-being of others than in other countries? This project seeks to understand the genesis of prosociality, investigating its biological foundations, the influence of cultural traditions, and effects of political, economic and legal structure. The dominant theory in economics views philanthropy as a solution to social illnesses that the market and the state are not solving, a view complementary to political science theory on preferences for government provision. Sociologists focus on social norms emerging from religious traditions. Cultural evolutionary theory highlights the instrumental value of trust. Still other theories have suggested a role for natural selection of genes.

However, these theories have not been tested stringently nor simultaneously. Also the project includes a very important factor largely ignored thus far: political, legal and economic institutions also affect the level of giving as well as who gives to which causes. Therefore, the objectives of Global Giving are (1) to map country differences in the size and nature of philanthropy across the world; (2) to develop and test multidisciplinary theories explaining these differences; (3) to facilitate international collaboration across disciplinary boundaries in research on philanthropy. The research draws upon 200 surveys recently harmonized by the PI and on new data on philanthropy to be collected among large samples in 145 countries across all continents. Collaboration with international networks of academics safeguards the validity of the questionnaires and experiments. Appropriate multilevel regression models will be used, the lack of which caused biases in previous research.

An integrated understanding of philanthropy is useful not only for theory development, but also for government policy makers and practitioners in nonprofit organizations seeking to mobilize philanthropic contributions and make them more effective. The application in practice is ensured through collaboration with a large network of practitioners.

KEYWORDS
philanthropy, charitable giving, cooperation, international comparative research

Main field
Sociology

Other fields
Social and Organizational Psychology, Political Science, Macroeconomics, Public Administration

Explanation of the multidisciplinary character of the proposal
Societies do not neatly observe the boundaries between academic disciplines, and neither do people. Giving can be a political act; it can be religiously motivated; it can be a form of conspicuous consumption; and it may spring from humanitarian concerns. The same gift can result from all of these. Yet different
disciplines have focused on explanations at different levels, and with particular clusters of determinants. Political science has understood philanthropy as political behavior based on liberty and democratic institutions. Sociology has emphasized cultural traditions, trends and social influences at the national and social group level; social and organizational psychology has studied group processes and the behavior of individual citizens. Biologists have explained philanthropy as an extreme form of altruism resulting from evolutionary processes selecting for certain genetic polymorphisms. In reality, these factors are all interwoven in a complex web of determinants. To explain cross-national differences in philanthropy, monodisciplinary explanations are incomplete. We need to understand how individual motivations and societal circumstances interact. For instance, personality and self-identity theories explain giving as the result of individual traits with genetic underpinnings, while institutional theories suggest that such traits matter only for individuals in groups that are actively mobilized.

**Number of words**

Description of the proposed research: 7,957 (max. 8,000 words)
Knowledge utilisation: 849 (max. 1,000 words)

<table>
<thead>
<tr>
<th>Description of the proposed research</th>
</tr>
</thead>
<tbody>
<tr>
<td>This proposal and background materials are publicly available at <a href="https://renebekkers.wordpress.com/global-giving/">https://renebekkers.wordpress.com/global-giving/</a></td>
</tr>
</tbody>
</table>

Why do citizens in some nations give more money, time and other resources to the benefit of society than citizens in other nations? How can these differences between nations be explained? *Global Giving* moves the frontier of research beyond the focus on western countries and narrow mono-disciplinary accounts that dominate the literature today. Going beyond micro-level theories that explain who gives what and why, *Global Giving* will provide an integrative multidisciplinary explanation of cross-national differences in generosity, adding institutionalization and organization of the non-profit sector as macro-level context factors. Also the research will improve the methods and statistics currently used in behavioral social science research.

To do so, *Global Giving* will (1) map country differences in the size and nature of philanthropy among large (n > 1,000) representative samples across a large number (145) of countries; (2) develop, operationalize, and test theories explaining differences in philanthropy between countries; (3) facilitate international collaboration across disciplinary boundaries in research on philanthropy and provide methodological innovations that will benefit researchers in the social sciences as a whole.

Philanthropy – the voluntary contribution of private resources to collective welfare [1] – is an increasingly important social behavior contributing to the well-being of citizens in nations across the world. Governments are calling upon citizens and corporations to take responsibility for society as a whole. Engagement in philanthropy has been linked to better health and well-being of the giver [2–5] but more importantly, society benefits from generosity. Recent history provides dramatic examples of the contribution of philanthropy to a better world. Long before the international community came to a coordinated response, nonprofit organizations such as Unicef and Doctors Without Borders provided emergency shelter, food, and health care to victims of the civil war in Syria. The first Ebola vaccine was produced in a lab funded by the Wellcome Trust. Through coordinated efforts supported by a coalition of the WHO, Unicef, and the Rotary International Foundation, polio was eradicated from Africa. The Ocean Cleanup system was recently launched to remove plastic waste from the oceans. In countries like Greece, still suffering from the economic downturn, volunteers in churches and welfare organizations provide for basic needs through soup kitchens and food banks.

Across the globe, countries differ markedly in the prevalence of philanthropy, as Figure 1 shows. Such behavioral differences in prosociality are at the heart of *Global Giving*. Countries with high levels of engagement in philanthropy are the Netherlands, Ireland and the UK (75-80% giving to charity), Australia and New Zealand (around 70%), and the US and Canada (60-65%). Much lower levels of engagement in philanthropy are found in Russia and China (5-10%), India and Japan (~25%). Also within Europe we find considerable differences, where citizens in the Netherlands, the UK and Ireland, are 8 times more likely to engage in philanthropy than citizens in Lithuania, Russia or Greece. Where do these differences come from? How can they be explained? Are the differences indicative of generosity, or do they merely reflect economic prosperity? Due to a lack of data these questions cannot be answered. The data in Figure 1 say nothing about the amounts donated. In the Netherlands, the only country beyond the US where such a figure is available, philanthropy amounts to about 0.8% of GDP [6]. So while the Netherlands may seem to be a generous country in Figure 1, the amounts donated in acts of giving tend to be small for most households [6]. In the US, philanthropy amounts to about 2% of GDP [7].

An attempt to harmonize existing data on philanthropy from various countries in the world in the International Philanthropy Database (IPD) [9] concluded that a comparative analysis of amounts donated is difficult. As a result, previous comparative research on philanthropy is limited to analyses of the likelihood that citizens donate money to charity [10].

Theoretical models of philanthropy discussed below have focused at the individual level and have been developed in western countries. They suggest that different motivations underlie the decision to give money and the amount donated. However, it is unclear how prevalent these motivations are in different countries. This is important to know because academic research on philanthropy has been conducted primarily in western countries. While philanthropy is an increasingly global phenomenon, the use of ‘WEIRD’ samples of Western, Industrialized, Rich and Democratic participants limits the generalizability of existing research [11]. The generalizability of research on philanthropy is also limited because of the overwhelming influence of seemingly trivial context factors [12, 13]. Experiments with charitable giving as an outcome variable have produced widely different results as a function of minimal differences in the design of these experiments.

Motivations for giving are likely to differ between countries. But before we can even begin to answer the question how characteristics of countries affect motivations for giving, it is important to know the basic numbers: where do which citizens give to what to which causes? The global increase in charitable giving in the wake of the economic recovery after the financial crisis suggests that improvements in the economy can increase giving across societies [6]. However, even long periods of economic growth and strong increases of wealth
have not increased philanthropy in China, India and Russia by much, and the ranking of countries has remained about the same. Why are people living in countries such as the Netherlands, the UK, the USA, Canada, Australia and New Zealand much more likely to give to charity? What proportion of income and wealth are donated in these countries? Are relative levels of generosity there also higher than in other countries?

Overview. The research I propose on country differences in philanthropy is based on an integrated model of giving behavior involving determinants at three levels, displayed in Figure 2 [14]. The model departs from macro-level differences between countries in values and institutions in the top left corner, which affect meso-level differences in the activities of nonprofit organizations about mobilization strategies (arrow A). The strategies of nonprofit organizations, in turn, are attuned to and affect the likelihood that individual citizens will be asked to contribute, and the characteristics of the situations in which they are asked (arrow B). For individual citizens, the decision to give is influenced by the characteristics of the giving situation and their resources and prosocial values (arrow C). Both the organizations that solicit contributions (arrow B) as well as the national context in which individual giving decisions are made (arrow D) affect the nature and level of giving. They also affect the influence of situational and individual donor characteristics (arrows E and F).

The simultaneous consideration of factors at three levels is a novel contribution to the field of research on philanthropy. While the three-level model is commonly used in research on education, stratification and health, most of the academic research on giving behavior has been limited to characteristics at the micro-level [15, 16]. For social scientists, however, the enormous differences between countries in the proportion of the population engaging in philanthropy that the map in Figure 1 shows pose intriguing questions. Why do some countries have much lower proportions engaging in philanthropy than others? How much do people give as a proportion of their income and wealth in countries around the world? How can country differences in generosity be explained? Could situational characteristics that increase generosity in western countries be leveraged in non-western countries? The research I propose is the first to answer these questions. It is also relevant for practitioners in nonprofit organizations. Knowledge on situational characteristics and organizational
strategies that make people give are obviously of interest to organizations that collect contributions. *Global Giving* will show how factors that drive giving differ between countries.

**Theory.** Philanthropy is a unique form of human cooperation. It poses a puzzle: in contrast to informal helping, it cannot be explained from principles of reciprocity [17]. People help their neighbors, friends and family, both in their current country of residence and in their country of origin through remittances. People even help strangers by doing things for them, or indirectly by giving them access to their resources. These are examples of informal giving because there is no formal organization involved. Informal giving can be explained by kinship altruism, [18, 19] direct reciprocity and exchange [20, 21]. However, people also engage in cooperation by giving time to nonprofit organizations by volunteering, donating money to charitable organizations and giving blood to blood banks. These examples of giving all involve an intermediary organization that channels the help offered by individuals to the beneficiaries. Even though friends and family often raise funds for charitable causes in their direct social environment, acts of philanthropy cannot be driven entirely by direct reciprocity, and require different explanations [22–24].

*Global Giving* focuses on such acts of formal giving. In these acts people usually do not have direct contact with the beneficiaries, and they cannot be sure that their efforts will be compensated in the future. In these cases, philanthropy does not hold the promise to the donor that once she will be the beneficiary of an act of reciprocated giving. Because direct reciprocity is unlikely to explain formal giving it poses an intriguing puzzle for social science.

That said, there is no dearth of theories explaining philanthropic behavior, especially at the micro-level. Philanthropy is so multi-faceted that it cannot be explained by a single theory, but requires a multidisciplinary perspective [25, 26]. Therefore, an important aim of the current research is to synthesize the theories that purport to explain philanthropy in different disciplines. In addition, we draw special attention to theories that identify factors at the meso- and macro-levels of analysis, because they have been ignored in previous research and hold the key to explain cross-national differences.

**Micro-level theories: Who Gives When?** Researchers have tried to solve the puzzle of philanthropy in two different ways, each with its own type of explanation and methodology. The first type of explanation uses as evidence answers to the question ‘who gives?’ In this strand of research, random sample surveys are often used to measure the tendency of a large number of people to engage in philanthropy through a series of questions on various forms of giving, including donations to charity and volunteering. The second type draws upon answers to the question ‘when do people give?’ In this strand of research, specific examples of giving are observed among small samples of participants in experiments that manipulate situational characteristics. An important contribution of *Global Giving* is that these two types of explanations will be combined, and investigated in conjunction with explanations for country differences.

1. **Who gives?** Broadly speaking, two groups of factors dominate the literature on characteristics of charitable donors and countries: resources and values. At the individual level, citizens with more resources at their disposal in the form of human and social capital and with stronger prosocial values give more.

   **Resources.** As charitable giving is costly, those with more resources at their disposal will be more likely to engage in it [27]. There is a broad consensus in research relying on surveys that education, status and income are positively correlated to charitable giving [28–30]. The level of education is one of the ubiquitous correlates of charitable giving. In all countries where this relation has been investigated, higher educated citizens are more likely to contribute and give more [28]. Analyses of donations using data from income tax returns and longitudinal panel surveys in the Netherlands and the USA show that as their income and wealth increase citizens spend higher amounts on giving, but a lower proportion of income [31, 32]. Results from several other countries tend to be similar [30]. In my previous research I have used life course data and genetically informed designs to examine how the relationship between the level of education and philanthropy can be explained within countries [33, 34]. To some extent, the correlation between the level of education and the amount donated is due to higher levels of income and wealth accumulated [35] but controlling for them does not diminish the relationship to zero [28]. Charitable giving is also dependent on resources of others accessed through social networks [35, 36]. To some extent, social capital explains why human capital is associated with charitable giving [35]. Also the relation between religious affiliation and philanthropy is reduced when social capital is taken into account [37].

   **Prosocial values.** While resources increase the budgets that constrain the giving behavior of individual citizens, they do not give a positive explanation of why they give, and how generously they give relative to their budgets. Charitable giving will only occur when people prioritize helping others and collective welfare. The moral principle of care, the internalized value that citizens attach to helping others [38, 39] is thus an important correlate of charitable giving and a variety of prosocial behaviors in daily life [40]. The principle of care may also be able to explain why some people are more generous than others and why members of religious groups give more. As all major religions stress the moral value of helping others, it is likely that the principle of care explains a part of the remaining relationship between religious affiliation and charitable giving. The *Global Giving Survey* will be the first survey to study the giving behavior of members of non-western religious groups in conjunction with the moral principle of care. In addition, generalized social trust is an important trait of citizens who engage in prosocial behavior [41, 42].

2. **When do people give and why?** Research on philanthropy commonly contrasts purely altruistic motivations to give with ‘impure’ or ‘egoistic’ motives, sometimes labeled ‘warm glow’ [43]. The hypothesis of pure altruism states that the desire to help others, if aware and given the opportunity to have an impact, leads individuals to adapt their level of giving to the societal need for contributions. Experiments finding that most participants do not do so proportionally is taken as evidence for non-altruistic motivations for giving, as if people derive utility from the act of giving and not so much from the increase in well-being among recipients of donations [44–46]. The reasons why people enjoy the act of giving more so than the increase in the well-being of recipients include social image concerns (the anticipated reputational benefits of giving), the anticipation of guilt, and the desire to express support for certain values by giving time or money to a specific cause [46, 47]. While these motivations were found in experimental research on charitable giving, they can also be
identified in research on volunteering [48] and blood donation [49].

In addition to these motives, there are also external factors that are largely beyond the control of individual citizens, but have an influence on giving. One particularly strong finding is that a vast majority of acts of giving occur in response to a specific solicitation to contribute [50, 51]. However, even though solicitations are strongly related to the likelihood that people contribute, they do not explain much variance in the amount contributed. These decisions seem to be governed by other circumstances and motivations. Strategies employed by nonprofit organizations and the level of professionalism in mobilization are important meso-level factors that influence private contributions [52–55]. The defaults that nonprofit organizations use and the suggestions they make about potential contributions strongly affect individual giving [56–64]. Another external factor is the price of giving [65]. When donations to nonprofit organizations are tax-deductible, the effective price of giving is reduced, and giving increases [66].

3. Integration of research on who gives and when and why people give

Theories on when and why people give can be used to explain who gives because some citizens are more likely to encounter situations that include mechanisms and motivations that drive giving [28, 30]. Thus far, few studies have explicitly tested arguments using measures or manipulations of the mechanisms that drive giving, simply because there are virtually no survey datasets on charitable giving that also include such measures. The Giving in the Netherlands Panel Survey (GINPS) is an exception to this rule [67]. An important contribution of Global Giving is that mechanisms theoretically predicted to drive giving behavior will be measured explicitly. I will build on the GINPS to include measures and manipulations of the mechanisms driving philanthropy.

Macro-level theories: where do people give? The preceding discussion has focused on giving at the individual level of citizens. Global Giving extends the research within countries discussed above to a comparative analysis across countries. The basic tenet of comparative research is that it is the location of citizens in a certain nation that influences their giving, regardless of their individual characteristics. Giving is shaped directly by the resources and values that countries influence (arrow D in Figure 2 above) and indirectly through the presence of nonprofit organizations and their activities (arrows A and B). Arrows E and F visualize the exciting possibility that characteristics of countries and nonprofit organizations also affect who gives what and when.

Previous comparative research has often used aggregate measures of characteristics of individual citizens. It is striking that some of the relations between characteristics of citizens and philanthropy consistently found at the individual level do not emerge at the country level. While religion and education are positively associated with engagement in philanthropy within countries, this is not the case across countries. More educated countries are not more charitable on average [68] and the average level of religiosity of a country is not consistently related to engagement in philanthropy [69]. For education we see a pattern of equalization. In countries with a high average level of education, the difference between lower and higher educated citizens is smaller as the higher educated become less likely to contribute than in countries with a low average level of education. For religiosity, it has been found that religious citizens are more likely to give when their religious group forms a minority in that country [69]. The finding that individual levels of education and religiosity are positively related to giving while aggregate levels are not, can be explained by a single hypothesis: that the norm of giving is observed less by the groups in which it originated, as it is adopted widely by other groups. As it spreads from the religious to the non-religious, and from the higher to the lower educated. The Global Giving Survey allows for a test of this hypothesis, using both self-reported data on giving in the past year as well as prosocial choices observed in abstract social dilemmas.

Global Giving moves beyond an investigation of the characteristics of citizens and investigates characteristics of countries that affect giving by citizens in these countries. Obviously, it is important to take the influence of the composition of the population in the countries investigated into account. Without the presence of citizens who are able and willing to give the establishment of institutions that facilitate giving does not make much sense. But because the composition of the population cannot be influenced by policy makers and nonprofit organizations it may not be very useful to know which characteristics of citizens are correlated with generosity (arrow C in Figure 2) [70]. For practical purposes it is much more useful to know which groups of citizens are willing to give but are not yet asked to do so in practice. The Global Giving Survey will produce such knowledge. The research will also reveal which characteristics that are amenable to change influence charitable giving.

The major contribution of Global Giving is that the research answers the question which factors explain why some countries have higher levels of giving, taking population composition effects into account. I build this explanation on a comprehensive review of the literature, identifying six core theories:

1. Three failures theory [71, 72]. This is the dominant theory in economics, in which philanthropy is viewed as a solution to social needs that the market and the state are not meeting successfully.
2. Social origins theory, in which preferences for government provision are key [73]. Nations have different worlds of welfare, depending on the level of de-commodification and universalism.
3. Social integration theory, advocated primarily in sociology, identifying group cohesion and social norms emerging from religious traditions [37].
4. Cultural evolutionism, highlighting the instrumental value of fairness and trust in areas with high levels of economic integration [74].
5. Biological and physical theories, perhaps the most controversial group of theories that have suggested that natural selection of genes and of the climate determine the level of prosociality in countries [70, 75–79]. In this view, humans are more charitable by nature in some countries than in others.
6. Institutionalism, emerging from recent work specifically in the field of philanthropic studies. From this perspective, national institutions affect the level of giving as well as who gives to which causes [54, 55, 80].

The puzzle that Global Giving seeks to solve is how these theories are related to each other and to the facts. I conjecture that geographic and biological conditions determine the prosocial potential of countries, and that cultural traditions and political preferences also provide a fairly stable foundation for prosociality. Economic and legal institutions, however, are more flexible or malleable, open to choice. Societies have developed institutions
and different traditions of giving as a result of the accumulation of human and social resources, and specific values and attitudes. Global Giving will therefore focus primarily on the institutions that countries have developed to facilitate giving. These explanations involve both national institutions as well as organizational responses. They are largely independent of individual characteristics of citizens.

The core idea of Global Giving is that institutions are key not only to how much is given in a country, but also to who gives to which causes, and why. It is the presence of nonprofit organizations, the sophistication and professionalism of strategies of nonprofit organizations that determines whether the willingness and the capacity of citizens to contribute to society is mobilized and transformed into action. A study of blood donation in the European Union in the 1990s showed that the organization of the collection of blood strongly affects the size and composition of the pool of blood donors [54]. Recently, such differential mobilization effects of institutions on the composition of the population of donors have been documented in the UK [81].

The institutional characteristics of countries are important potential influences explaining philanthropy. Based on the empirical analyses of philanthropy in 26 countries in the International Philanthropy Database, Wiepking & Handy [9] listed a set of facilitating factors for philanthropy. Specifically, Global Giving investigates how philanthropy is related to (1) government support to charitable causes, (2) the regulation of nonprofit organizations, (3) fiscal incentives for nonprofit organizations, and (4) the professionalism of fundraising.

Government support. In comparisons between the US and Europe it is often argued that public goods paid by taxes and philanthropy are substitutes and that private donations are crowded out by government subsidies. Are the US more charitable than Europe because taxes and welfare state spending are higher in Europe? Such a ‘crowding-out’ hypothesis would explain why hospitals, universities and cultural nonprofits in the US receive a larger share of their income from donations than in Europe. However, a fair test of the crowding-out hypothesis requires more than just two countries. A recent meta-analysis of research on the crowding-out hypothesis demonstrates that the evidence in favor of the crowding-out hypothesis is rather thin [45]. In fact, the US may be the exception rather than the rule. In Europe, citizens in countries with higher taxes, like the Netherlands and Sweden, are actually more likely to give to charity rather than less [82]. This result speaks against the crowding-out hypothesis, and suggest that other characteristics of the US explain its high level of philanthropy. Without evidence on the amounts donated and appropriate statistical models, however, these results are not very convincing as a test of the crowding-out hypothesis. Thus it is important that Global Giving will measure amounts donated to charitable causes. Also welfare state provisions may not affect the proportion of the population donating to charitable causes or the amounts donated, but rather the causes supported. Only when government subsidies cover basic needs, education, and health, citizens with prosocial values will direct their giving to overseas causes, including international relief and development, and human rights [83, 84].

The ‘crowding-out’ hypothesis is part of a broader political debate on the role of government in the provision of public welfare. A seminal contribution to this debate was made by Salamon and Anheier in their social origins theory [85], based on the work of Barrington Moore [86] and Esping-Andersen on welfare states [73]. The theory distinguishes four types of nations, and predicts that giving will be more widespread in liberal democracies, followed by corporatist and social-democratic states. Preferences for provision of public goods by charities rather than the state would be expected to be strongest in liberal welfare states. Citizens in the UK indeed prefer charities in many areas, and do so more strongly than 10 years ago [81]. However, the hypothesis has not been thoroughly tested across countries. The data collected in the IPD and the Johns Hopkins International Comparative Philanthropy project, however, do not give much support for social origins theory. Einolf recently concluded that social origins theory “fails to predict present-day cross-national variation in charitable giving and the size of the nonprofit sector.” [87]

Regulation and fiscal incentives. Countries vary strongly in the legal framework and regulation of nonprofit organizations [86–90]. The regulation of fundraising and the level of transparency of charitable organizations are likely to affect the level of charitable confidence among the general public [41]. In conjunction with the practice of fundraising the aggregate level of charitable confidence may affect charitable giving [52]. Using data on the legal framework for giving across a large number of countries in the world Global Giving will investigate how giving is related to tax laws and regulation [90]. Changes in such tax laws and regulations can be viewed as natural experiments [91, 92] that create exogenous shocks in the price of giving for individual citizens. The deductibility of donations in the income tax lowers the price of giving. In more progressive income tax systems, donations are a more attractive way for citizens to reduce taxes.

Resources. At the country level, the proportion of the population engaging in charitable giving increases with GDP [8, 16]. Due to a lack of data, it is unknown whether this relationship holds for the amounts contributed and for relative levels of generosity. As a country becomes more resourceful, it is likely to become more charitable because citizens feel more financially secure. Cultural evolutionism predicts that citizens in more open economies are more likely to give. Henrich et al. found that stronger market integration is positively related to offers in social dilemma games [74]. The number of nonprofit organizations is also likely to be higher in countries with a higher level of trust and with more economic and political freedom, because it is easier for citizens to establish them. As the number of nonprofit organizations increase, citizens are more likely to be asked for donations. In a study of cadaveric organ procurement in the US, Healy found that donation rates increase with the resources devoted to procurement [55]. In addition to the level of resources, also the distribution of income may indirectly affect charitable giving, because income inequality reduces trust [42] and prosocial values [93]. Trust and civic norms are also stronger in more educated countries [94].

Prosocial values. As a country becomes more trusting and convinced of the value of helping others, it is likely that its citizens become more likely to engage in charitable giving. It is striking that the countries with high levels of engagement in philanthropy in Figure 1 are all countries with a protestant Christian religious tradition or past. Lower levels of engagement in philanthropy are found in countries with a Catholic majority where trust is lower, though Ireland is an exception to this pattern. Also at the individual level, charitable giving is higher among Protestants than among Catholics [47]. Some studies suggest that aggregate levels of trust are indeed positively
Thus far, no studies have examined aggregate levels of prosocial values in relation to philanthropy. Another unique contribution of Global Giving is that a comparative test of the mechanisms will be conducted across countries. Global Giving allows for tests of hypotheses on where these influences are likely to differ. Macro-level factors influence not only the prevalence and nature of collection strategies, but also their effects. The influence of reputation depends on the norm of giving [96]. Who gives what depends not only on the characteristics of donors and decision making situations, but also on how societies have institutionalized the provision of welfare, and on how organizations solicit contributions [54, 55].

Mobilization strategies used by nonprofit organizations affect the likelihood that citizens are asked to give as well as the composition of the donor pool [97]. In the case of blood donation, countries that collect blood through academic research hospitals have a much higher proportion of students as blood donors than countries that collect blood through donor associations [54]. For donations of money, collection systems and mobilization strategies have been described recently [52]. In the Netherlands, I have found that institutions created to raise funds have a strong influence on levels of charitable giving and the composition of the pool of donors and volunteers [98, 99]. Global Giving will investigate how institutions and mobilization strategies affect the level of giving and the composition of donor pools in countries across the globe.

**Ground-breaking nature of the project and its innovative and multidisciplinary aspects**

Global Giving is ground-breaking because of its multi-disciplinary theoretical perspective, its multi-national scope, its integration of experimental and survey research methods, and its connection with practice. The theoretical contribution of Global Giving is the systematic development of a set of hypotheses based on the institutional perspective, and the comparison with the other theories. The empirical contribution of Global Giving is that it will provide the best possible test of the six theories. None of these theories have been tested stringently or simultaneously with other theories. The practical contribution of Global Giving is that it will provide knowledge that policy makers and executives in the charity and nonprofit sector can apply directly in practice.

**Approach**

Global Giving uses both existing data, as well as new data from a high quality survey, modeled after the Giving in the Netherlands Panel Survey (GINPS) [100, 101], including experiments with giving behavior, administered to large, random samples. The survey also includes social dilemma games, in which participants make choices about valuable points, which will be converted in local currency after the survey is completed. Appropriate hierarchical multi-level models will be used to analyze differences between countries. To ensure the validity of the module in non-western countries, Global Giving works with an international review board from the Center for Global Generosity (CGG), including experts from the PI’s extensive global network of researchers on philanthropy.

**Measuring giving** Cross-national comparative research on philanthropy has lagged behind due to a lack of data. The first objective of Global Giving, to map differences in giving behavior across countries, requires data that are comparable across a large number of countries with very different traditions of philanthropy. The multi-country datasets that are available to researchers are limited to aggregate figures and often used without critical attention to the quality of the data [10]. Funding for Global Giving would solve both these problems, as it includes data collection for the Global Giving Survey using one single methodology among large samples (n > 1,000). The data collected in the Global Philanthropy Survey will enable researchers across the world to test theories on determinants of giving at both the individual and national level, using one single validated methodology.

This objective is difficult to achieve, because different words are used in different countries to talk about giving, and the same words can mean different things to respondents in different countries or even within the same country. As a group of US researchers in a study of different survey modules concluded: ‘methodology is destiny’[102]. The wording of the questionnaire matters not only for the resulting estimates of the size and composition of giving [102], but also for the profile of who is a generous donor [10, 103]. Therefore we develop the questionnaire for the Global Giving Survey in close cooperation with country experts from the Center of Global Generosity. The development of a master questionnaire will start with a critical review of the Giving in the Netherlands Panel Survey (GINPS), a high quality survey measuring donations to charitable causes in the Netherlands in an extensive questionnaire module [28, 100]. The module produces estimates which are valid (r = .85 between self-reports and archival records of donations received) [28] and reliable [104], also for non-native Dutch citizens with culturally different norms and traditions for giving [105].

The GINPS was successfully adapted for comparative research in France [106]. The results for France show that the questionnaire for the Netherlands can also be adapted to other countries, retaining measurement equivalence. The lessons learned in the creation of the Giving in France Survey include that it is important to define giving in behavioral terms, to adapt the examples of organizations for areas in which people can give to the country context, and to tailor the terminology used to national vocabulary and country specific giving traditions. Problems of adaptation are particularly severe for countries beyond western Europe. The questionnaire for the survey will be translated into the primary EU languages, following the TRAPD methodology (Translation, Review, Adjudication, Pretesting and Documentation) [107], which is also used in the European Social Survey. In the Netherlands, the experience of the adaptation of the GINPS to surveys for members of ethnic minorities has yielded valuable lessons on the formulation of questions about giving in non-western cultures [105]. For instance, respondents with a Muslim affiliation typically do not report donations to their mosque. In Islam, zakat is an important religious duty that is not viewed as a ‘donation’. Not counting such contributions to religious institutions, however, would lead to a serious underestimation of giving for Muslim respondents.

As predictors of giving, in addition to standard questions on demographic and economic indicators, the GINPS includes questions on prosocial values such as trust and the principle of care and a variety of prosocial behaviors, such as informal helping, providing care, blood donation, and remittances. Inclusion of these questions enables a theoretically informed comparison of models of
Measuring prosocial values and behavior

Developmental and personality psychologists have long believed that the internalized moral value attached to helping others is an important motivation for helping behavior [39, 111]. Yet no scale was available to measure this value. Starting from three items in the General Social Survey 2002 in the US [40], Mark Ottoni-Wilhelm and I have developed such a survey measure which we call the principle of care and validated it in two national sample surveys in the Netherlands, and one in the US [112].

An alternative approach to measuring prosocial values and preferences is found in social psychology and behavioral economics, in which choice behaviors in abstract situations are observed to measure revealed preferences and perceptions of others [113]. In the online survey, participants will play a variety of games from behavioral economics with other participants. While experimental games have been around for decades, the recent advent of behavioral economics has spurred research on prosocial behavior [114]. In this line of research experimental games are used to study abstract forms of prosocial behavior, typically among convenience samples of students. These experiments include prisoner’s dilemma games, trust games, ultimatum games, dictator games, and common pool resource dilemmas [113]. The dictator game is most often used to model giving behavior [115]. The participant decides about the division of an amount of money between him/herself and ‘another person’, who is not involved in the game and has no power to refuse the amount allocated by the dictator. The participant does not know the other person and will not meet this person after the experiment. The design of the dictator game resembles the situation in which people decide about donations to charitable causes. Choices typically made in these games stand in contrast to the prediction based on a completely self-interested model of man that dictators keep all money for themselves, especially if the recipient is deserving [116]. The amount allocated to the ‘other’ decreases as the decisions of the dictator become more anonymous for third parties such as the experimenters [117]. But even in the absence of strategic concerns participants still give [118].

Critiques to this line of research are the lack of external validity (generalizability) of abstract game experiments [119–121], the use of non-representative samples of participants [11], and the use of windfall gains (‘house money’ or ‘manna from heaven’) [91]. Participants in experiments are more generous with windfall gains than they have received from the experimenters than with earned income [92, 122, 123]. The Global Giving Survey lets participants play with windfall gains as well as with earned income. In previous research, I have successfully included a set of hypothetical games to measure social value orientation in the GINPS. Prosocial choices in abstract game situations are positively correlated to amounts donated to charitable causes in real life [124]. In the games token points are at stake. The points earned in these games will be converted into an amount in local currency that will be added to a fixed reward for participation. At the end of the survey, participants enter a dictator game, in which they are paired with either a charity or another participant. The GINPS has incorporated such donation experiments, with high external validity. Correlates of donations observed in these experiments are similar to correlates of self-reported donations [125]. Similar findings have recently been reported in experiments in Germany and other countries [126, 127].

Context data

Measures of the level of institutionalization and professionalization will be used from existing sources [9] and collected for additional countries. Aggregate data at the country level from the 200 surveys in the Harmonized Trust Database will be used as a measure of trust [128]. Other data on country characteristics such as welfare state spending, legal characteristics, trust, corruption, democracy will be added from established databases and trusted sources including the OECD, Eurostat, the European Social Survey, the World Values Survey, the World Inequality database, the Center for Global Prosperity, the Charities Aid Foundation, Transparency International, the European Foundation Center.

Statistical models

A problem that has plagued comparative research is the use of inadequate statistical models and low statistical power [10]. While many characteristics of countries could explain country differences which are all interrelated, the number of countries for which data are available is typically low. To obtain robust evidence on country-level correlates from multi-level regression analyses, the number of countries included should be as high as possible, but certainly higher than 25 [129]. The availability of data from more than 25 countries allows for robust estimates of country differences, provided that country characteristics are not too strongly correlated. The models will not include variables that vary over time, so a standard nesting structure is adequate [130]. Appropriate statistical models typically reduce strong correlations at the country level to very weak relations at the individual level of citizens, explaining only a small proportion of the variance [8, 10, 68].

Research team and planning of the project

Global Giving consists of four work packages (WPs). With the assistance of a co-PI (postdoc) and a research assistant, I will coordinate Global Giving. We will consult with a carefully selected international advisory board, in teleconferences twice a year, and face-to-face at two project conferences.

In WP1, we will prepare a master questionnaire in English for WP3, based on the mega-analysis. The co-PI will act as daily supervisor of the PhD candidates, who write their dissertation in WP2 and WP3. PhD#1 and PhD#2 will both assist with the development of the survey, and write at least three empirical journal articles on country differences in giving, one on macro-level institutions, one on organizational strategies, and one analyzing all three levels. PhD#3 will assist with the development of the experiments, and write at least three empirical journal articles on cross-national experiments, one on the behavioral games, one on the donation experiments, and one on the interrelation of these with survey reports on giving in the past calendar year. A research assistant will provide assistance with data management, website design and maintenance, conference organization, and correspondence with a network of experts throughout the project.

Schedule

WP1: Preparations for a new survey through mega-analysis and questionnaire development. In WP1 the PI and co-PI test the validity of previous surveys including the World Giving Index (WP2) and prepare the methods and materials for the Global Giving Survey (WP3). The general research question answered in WP1 is: How can philanthropic behavior be measured accurately? Specific research questions answered are in WP1.1: How do characteristics of donors
Coordination and dissemination

Advisory board meetings

Mid-term and final conference

Workshops

Select 3 PhD candidates

<table>
<thead>
<tr>
<th>YEAR 1</th>
<th>YEAR 2</th>
<th>YEAR 3</th>
<th>YEAR 4</th>
<th>YEAR 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

WP1: Preparation

1.1 Theory development (PI, Co-PI)

1.2 Mega-analysis (PI, Co-PI, PhD1)

1.3 Develop questions (PI, Co-PI, PhD1)

WP2: World Giving Index

2.1 Prior waves (PI, Co-PI, PhD1, PhD2)

2.2 All waves (PI, Co-PI, PhD1, PhD2)

WP3: Global Giving Survey

3.1 Pilot (PI, Co-PI, PhD2)

3.2 All countries (PI, Co-PI, PhD2)

WP4: Experiments

4.1 Pilot (PI, Co-PI, PhD3)

4.2 In survey (PI, Co-PI, PhD3)

**TABLE 1** Schedule. The red lines mark the mid-term and final conference.

Vary between surveys as a result of survey design features? and in WP1.2: How can measures of giving practices and giving contexts be optimized to capture country specific giving practices?

In WP1.1, we use a method called mega-analysis to estimate the effect of survey methodology on results of survey data analyses on philanthropy and prosocial values. Mega-analysis takes advantage of the fact that in several countries, multiple surveys of varying quality have been conducted at the same time [131]. By using the primary data reported in previous analyses rather than results reported about the data, mega-analysis does not suffer from the key disadvantages of meta-analysis. It can be used for any variable. In an ongoing mega-analysis responses to questions about trust in ‘most people’ are compiled in the Harmonized Trust Database, a big data file including 3.7 million observations from 200 different surveys. WP1.1 applies mega-analysis to the first comparative research data file of survey data on philanthropy created by Wiepking & Handy, the International Philanthropy Database (IPD) [9, 80]. This file compiles data from various surveys in countries across the world that have included questions on charitable giving. Because the methodology used in the different countries varies considerably, it is difficult to interpret country differences without considering methodological differences. In the IPD only one survey for each country is included: the one with the highest quality (e.g., the GINPS for the Netherlands and the Philanthropy Panel Study from the US) [128]. In WP1.1, all available data from surveys that have been conducted at the same points in time as those already included in the IPD, also those of lesser quality, will be added to the database to enable such a mega-analysis. The analyses will demonstrate how conclusions published earlier based on analyses of data of lower quality from existing surveys differ systematically from conclusions based on higher quality data. I have demonstrated the feasibility of this approach in two analyses pooling data from various surveys [10, 100].

In WP1.2, the questionnaires developed for the Giving in the Netherlands Panel Survey and the Giving France Survey will be translated in the major world languages and adapted for cross-national use in the Global Giving countries. Also measures of the social, political, economic and legal context [88, 90] in which individual citizens make decisions on giving will be collected and harmonized.

**WP2: World Giving Index.** The general research question answered in WP2 is: How can differences between countries in generosity be explained? In WP2.1, we analyze data collected through the Gallup World Poll on three forms
of prosocial behavior: giving to charity, volunteering, and helping a stranger. Previous analyses of the GWP in the annual World Giving Index reports (CAF, 2010 – 2015) have shown aggregate differences between countries as shown in figure 1. In WP2.1 we apply hierarchical (‘multilevel’) regression models to (1) estimate variance components at the individual and country level, and (2) explain context effects by including data on country characteristics collected in WP1.2. In WP2.2 we collect other context data to measure relevant country characteristics and match these to the individual level data from the GWP, to exhaustively test theories on generosity, both at the individual and national level.

WP3: The Global Giving Survey. Research questions: How much is donated to which causes by citizens in different countries (WP3.1), and how can these differences be explained? (WP3.2) In close collaboration with Kantar Public and a network of experts in Europe (ERNOP) and beyond (Center for Global Generosity), the Global Giving Survey will be conducted in 31 countries with sample sizes of at least 1,000 respondents. The 20 largest EU countries will be included and a selection of 11 countries in three other continents (Canada, the US, Mexico, Chile, Brazil; China, Japan, India; Australia, Russia). This selection maximizes the proportion of the world population represented (57%) as well as the heterogeneity in country characteristics that are theoretically relevant for giving [132]. Africa is excluded because of the high costs for face-to-face surveys. An extension of Global Giving to more countries, including those in Africa, is feasible as soon as the infrastructure is in place. Also it will be possible to collect subsequent waves of data among the same respondents who have participated in the Global Giving Survey, following the example of the GINPS.

WP4: Experiments on philanthropy. The general research question answered in WP3 is: How does philanthropic behavior change in response to changing conditions in the decision situation? To answer this question, the Global Giving Survey includes two types of experiments: behavioral games during the survey [124], and donation experiments after the survey. The donation experiment is a modification of the ‘All-or-nothing Dictator Game’, which has been validated in the GINPS [124, 125, 133]. For the behavioral games, the lab will be set up online to enable interactive experiments. Participants in different countries can communicate with each other in real time. In countries where online research is not yet possible among representative samples parts of the fieldwork need to be conducted face-to-face. In these subsamples the strategy method will be used [134], which rules out communication between participants but produces valid results in social dilemma game experiments [135].

An important innovation in the experiments in Global Giving is that they will be conducted by international groups of researchers working together across disciplinary boundaries. I will offer small grants to teams of researchers selected through a request for proposals (RFP). A multidisciplinary international board of advisers will review the proposals following the example of the Science of Philanthropy program at the University of Chicago (http://spihub.org). Grants awarded entail presentation of the research plans at mid-project and final conferences. All experiments in Global Giving will be piloted through online platforms such as M-Turk, Crowdflower and Prolific, which quickly provides high-quality data from diverse samples [136–138]. Positive findings dominate the social sciences,[139] also in experiments on philanthropy [15]. Publication bias is likely to be an important reason for this [140]. Therefore the experiments in Global Giving are conducted as registered reports [141]. The replicability of research on philanthropy has not been assessed previously. Typically, registered reports yield smaller effect sizes and fewer effects below commonly used significance levels [142–144]. To eliminate publication bias, all experiments in Global Giving are conducted as registered reports [141]. The design and a power analysis will be made publicly available.

| Knowledge Utilisation |

Global Giving will be the first comparative study of the size and nature of philanthropy ever conducted. Global Giving will have a considerable impact on both practitioners in nonprofit organizations and on academic researchers from a variety of disciplines. A mapping of countries based on the data collected will make philanthropy visible on a global scale, demonstrating which countries are most generous and which are the least generous. Practitioners and policy makers will benefit from insights on what fundraising strategies and legal conditions make people give to charity.

Global Giving will have a significant impact in numerous social science disciplines. The question why people help others at a cost to themselves is a classic in economics, psychology and sociology [145, 146]. The science of philanthropy has grown exponentially [147, 148] in an increasing number of disciplines [149]. Explanations of cross-national differences in philanthropy originate in a variety of disciplines, including sociology, political science, law, economics, marketing and communication science. The applications of theories from these disciplines are relevant for scholars in these disciplines. The technique of mega-analysis[128] is applicable in many areas, such as health, happiness, and political interest. The survey will provide the first global data on philanthropy, using one single, validated methodology. Global Giving facilitates international interdisciplinary collaboration by funding proposals for survey experiments, which will encourage researchers in other fields such as behavioral economics and social psychology to implement survey experiments. Finally, the impact of Global Giving will also extend to research on other forms of prosocial behavior than monetary giving, such as volunteering, giving blood, organs and informal helping, because the survey will also include questions on these behaviors.

Global Giving follows the principles of Open Science. The research design, data collected and analyses in Global Giving will be made publicly available through its website. Experiments will be conducted by other researchers, who submit applications in a competitive grant scheme and will be attending the conferences along with practitioners from the global philanthropy sector. In this way, knowledge will be shared on circumstances that influence charitable giving and volunteering.

An international review board consisting of both academics from multiple disciplines, policy makers at the national, European and international level and practitioners from non-profit organizations will be formed to ensure that the research appeals to relevant stakeholders. The research will demonstrate how charitable giving relates to (1) government support to charitable causes, (2) legal regulation of nonprofit organizations, (3) fiscal incentives for nonprofit organizations, and (4) the professionalism of fundraising. The findings in
these areas will be discussed with the advisory board, grantees of the RFP and a selective group of researchers at a mid-period and final conference. In conjunction with the European Foundation Centre (EFC, www.efc.be) we organize workshops to discuss the implications of the findings with leaders and executives from the charity and non-profit sector, and with government representatives and public policy makers.

The PI is well-connected to networks of practitioners and philanthropy advisors in the Netherlands as well as in Europe and beyond. The European Research Network on Philanthropy (www.ernop.eu), co-founded by the PI, includes >200 members from almost all countries in Europe. All research institutes on philanthropy across Europe are members of ERNOP and will be involved in the research as co-producers and members of the Advisory Board. Also leaders from other networks and associations of researchers (ASGE, ARNOVA, ISTR) will be invited, also from outside the western world (AROCSA in Africa, the ISTR regional network in Asia). The European Foundation Centre will provide accommodation for a workshop at the Philanthropy House in Brussels. Further dissemination activities include a series of working papers throughout the project, an interactive website with open data, displayed on adaptable world maps. At the conclusion of the project we will publish a book about the results of the Global Giving Survey.

Not only will the research produce substantive research on origins of generosity, but also valuable knowledge on social science research methodology. The research will generate a survey instrument to measure charitable giving and volunteering in a valid and reliable manner across the globe. Knowledge on the effects of survey design is relevant to academics in all disciplines that rely on surveys. The research will demonstrate how behavioural experiments can be included within surveys. Through the RFP (WP4), academic researchers are actively encouraged to conduct field experiments, working with practitioners from the non-profit sector. The experiments will demonstrate the applicability of the findings in real world settings.

Similar to the World Values Survey, which has had a great influence on our thinking about values systems across the globe, Global Giving has the potential to have a similar impact by providing open access to our thinking about prosocial behavior.

In sum, the knowledge produced by Global Giving will be useful for researchers and policy makers beyond the horizon of the project itself. The data and publications will be available in open access mode through its website and public depositories. The networks formed are likely to generate new research well beyond the project period. The data collected will be available for eternity, providing generations of researchers with a treasure trove of data on prosociality.

<table>
<thead>
<tr>
<th>Cost estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total budget requested</strong></td>
</tr>
<tr>
<td>£1495.400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Intended starting date</th>
</tr>
</thead>
<tbody>
<tr>
<td>September 1, 2020</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data management plan</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Will data be collected or generated that are suitable for reuse?</strong></td>
</tr>
<tr>
<td>Yes</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Where will the data be stored during the research?</th>
</tr>
</thead>
<tbody>
<tr>
<td>During the data collection phase, data will be stored at contracted firms, at VU servers and in a shared folder on Surfdrive, protected by passwords and only accessible by project members.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>After the project has been completed, how will the data be stored for the long-term and made available for the use by third parties?</th>
</tr>
</thead>
<tbody>
<tr>
<td>All data collected and harmonized in WP1, WP3, and WP4 will be stored on the Open Science Framework (<a href="https://osf.io">https://osf.io</a>) in public projects. These data will also be provided through Dataverse.nl, using guidelines of DANS/EASY. Personal data will be removed from the datasets that will be deposited. The</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Cost Estimates</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Personnel</strong></td>
</tr>
<tr>
<td>Year 1</td>
</tr>
<tr>
<td>Year 2</td>
</tr>
<tr>
<td>Year 3</td>
</tr>
<tr>
<td>Year 4</td>
</tr>
<tr>
<td>Year 5</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

1. Conferences, outreach, publication, and travel
micro-data collected in WP2 are proprietary, owned by Gallup, and cannot be shared publicly. A public use file will be made available through the OSF.

**Which facilities do you expect will be needed for the storage of data during the research and after the research? Are these available?**

No particular research facilities are needed for this project other than standard computing facilities provided by the VU.

---

**Ethics**

**Use of extension clause**

No

**Ethical aspects**

| Approval from a recognised (medical) ethics review committee | Not yet applied for |
| Approval from an animal experiments committee | Not applicable |
| Permission for research with the population screening Act | Not applicable |

**TABLE 3 Ethical aspects**

**Declarations**

*By submitting this form I endorse the code of conduct for laboratory animals and the code of conduct for biosecurity/possibility for dual use of the expected results and will act accordingly if applicable.*

- [ ] I have completed this form truthfully
- [ ] By submitting this document I declare that I satisfy the nationally and internationally accepted standards for scientific conduct as stated in the Netherlands-code-of-conduct-for-research-integrity (Association of Universities in the Netherlands)
- [ ] If applicable: I have submitted a list of non-referees with my pre-proposal.
- [ ] If applicable: I have included one or more authorised form co-funding from the host institution (or a third party) guaranteeing to meet part of the costs of this research project

**Name:** René Bekkers  
**Place:** Amsterdam  
**Date:** August 27, 2019

---

**Society**

**Public summary**

**Global Giving (NL)** Waarom geven mensen in sommige landen zoals Nederland meer tijd en geld aan goede doelen dan in andere landen? Dit onderzoek in 145 landen gaat na welke invloed biologische factoren, culturele tradities, economische omstandigheden, de overheid, wetgeving, belastingvrijstelling en goederenorganisaties zelf hebben op de vrijgevigheid van burgers.

**Global Giving (ENG)** Why do citizens in some countries like the Netherlands give more time and money to charitable causes than in other countries? This research in 145 countries examines influences of biological factors, cultural traditions, economic conditions, government support, legal regulation, fiscal compensation and nonprofit organizations on the generosity of citizens.

**Reviews**

**Grant**

Vernieuwingsimpuls Vici SGW 2019

**Title**

Global Giving

**Applicant**

Prof. dr. R.H.F.P. Bekkers

**File number**

VLC.191.063

Your rebuttal on the referee reports is in progress

---

**Referee report of referee 1**

**Assessment of the quality of the researcher**

**Explanation.** Explanation. Criteria - the quality of the researcher

- ✔ in terms of profile fit in the target group;
- ✔ in the top 10% of his/her international peer group;
- ✔ academic excellence as demonstrated by numerous publications of international standing and/or other academic achievements;
- ✔ inspiring enthusiasm for research and/or technology;
- ✔ persuasiveness;
- ✔ demonstrably capable of successfully developing own new innovative line of research;
- ✔ has both a national and international prominent position;
- ✔ demonstrable leadership and coaching skills.

The Vici scheme aims at outstanding researchers only: the top 10% of his/her international peer group.

**Question a:** What is your opinion on the past performance of the researcher (as demonstrated by publications and other relevant scientific achievements)?

**Comments:** This researcher is internationally known as a leader in this field. He is innovative, highly cited, a frequently published in the most important journals.

**Question b:** Does the applicant belong to the top 10% of his/her international peer group (Please indicate which Group/Research Group or Peer Group you are referring to in this comparison.)? Which scientific achievements or talents of the applicant show that he/she belongs to this top?

**Comments:** This researcher is internationally known as a leader in the field of philanthropy research, well within the top 10% of his peer group. I would place him in the top 1%. He uses a broader range of methodologies with excellence than perhaps any other leader in this field.

**Question c:** To what extent is there sufficient evidence that the applicant has the ability to lead and supervise a research group and support staff and to
We have many individual country results from a very few number of nations, but almost nothing that seeks to harmonize results from across the globe. This has greater potential to make a major contribution than any project I have seen in this field in the last 10 years. No previous research has comprehensively explored intra-national differences in this field of activity in this way. Having one of the top researchers in the field undertake such an ambitious project could lead to not only practical results but deep insights into the origins of cultural differences. This researcher has completed similar substantial investigations in the past.

**Assessment of the quality, innovative character and academic impact of the proposed research**

**Explanation.** Criteria - the quality, innovative character and academic impact of the proposed research

- challenging content;
- originality of the research question;
- innovative scientific elements;
- aimed at building up a new line of research;
- potential to make important contributions to science;
- effectiveness of proposed methodology;
- international importance of the proposed research area.

**Question a:** Please comment on the relevance of the problem and on the originality and challenging content of the proposal.

**Comments:** This project and its results could serve as the next important step for the entire field of philanthropy research. It addresses fundamental questions that cannot be advanced outside of such an ambitious undertaking. We have many individual country results from a very few number of nations, but almost nothing that seeks to harmonize results from across the globe.

**Question b:** What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?

**Comments:** No previous research has comprehensively explored intra-national differences in this field of activity in this way. Having one of the top researchers in the field undertake such an ambitious project could lead to not only practical results but deep insights into the origins of cultural differences.

**Question c:** What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?

**Comments:** This has greater potential to make a major contribution than any project I have seen in this field in the last 10 years.

**Question d:** To what extent is the proposed method effective? Please comment.

**Comments:** As always, this researcher uses the most appropriate methodologies to precisely measure the key factors.

**Assessment of the knowledge utilisation**

**Explanation.** Criteria - knowledge utilisation (KU)

Potential

- contribution to society and/or other academic areas;
- disciplines and organisations that might benefit from the results.

Implementation

- action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- if and how the potential knowledge users will be involved;
- (concrete) outcomes for society and/or other academic disciplines;
- the period over which knowledge utilisation is expected to occur.

The selection committee assesses:

- whether the applicant has given a realistic description of the potential for knowledge utilisation;
- and to what extent the applicant has presented a concrete and convincing plan for the implementation of the available potential.

If a researcher is of the opinion that the proposed research is not appropriate for KU then he/she should explain why he/she thinks that KU is not applicable. The selection committee will assess the arguments given for this

**Question a:** What is your opinion on the described potential for knowledge utilisation?

**Comments:** Understanding motivations for civic cooperation in a practical, measurable, context such as philanthropy holds great promise for numerous social science disciplines and well as government and non-government organizations. It is a fundamental question with widespread potential impact.

**Question b:** Please comment on the effectiveness and feasibility of the proposed approach for realizing KU (implementation).

**Comments:** This researcher is ideally placed to effect knowledge utilisation. No one has broader international contacts or is more respected among those groups both with practitioners and academics.

**Question c:** Only answer this question in case the applicant argued that KU is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why KU is not applicable for his/her research project (see also the information under criterion 3 listed above)?

**Comments:** Not applicable

**Final assessment**

**Question a:** How do you assess the entire application? Please give your final scoring (A+/A/B/UF/U).

**Comments referee 01, A+** Highest quality, significance and recommendation for funding

**Question b:** Could you please summarize (point by point) the strengths and weaknesses of the grant application focusing on the candidate, proposal and knowledge utilisation? Strengths:

- This researcher is ideally placed to effect knowledge utilisation. No one has broader international contacts or is more respected among those groups both with practitioners and academics.

- This is a potentially revolutionary project proposed by the person most able to complete it which holds the potential for impact in academia, practice, governmental, and non-governmental organizations across the globe.

We do not have this information in the field and it is desperately needed.

**Data management**

**Explanation.** Responsible data management is part of good research. For the collection/generation of data and the analysis of this data, timely measures need to be taken to ensure the storage and later reuse of the data. This means that prior to the start of the research project researchers must ascertain a) which data could be relevant and b) how these data could be stored. After a proposal has been awarded funding, the researcher will draw up a detailed data management plan in which the researcher explains how all relevant data will be made findable, accessible, interoperable and reusable (FAIR). The data management section( 2e) is not an assessment criterion for the research proposal. However for all the data management sections of these proposals, you

- action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- if and how the potential knowledge users will be involved;
- (concrete) outcomes for society and/or other academic disciplines;
- the period over which knowledge utilisation is expected to occur.

The selection committee assesses:

- whether the applicant has given a realistic description of the potential for knowledge utilisation;
- and to what extent the applicant has presented a concrete and convincing plan for the implementation of the available potential.

If a researcher is of the opinion that the proposed research is not appropriate for KU then he/she should explain why he/she thinks that KU is not applicable. The selection committee will assess the arguments given for this

**Question a:** What is your opinion on the described potential for knowledge utilisation?

**Comments:** Understanding motivations for civic cooperation in a practical, measurable, context such as philanthropy holds great promise for numerous social science disciplines and well as government and non-government organizations. It is a fundamental question with widespread potential impact.

**Question b:** Please comment on the effectiveness and feasibility of the proposed approach for realizing KU (implementation).

**Comments:** This researcher is ideally placed to effect knowledge utilisation. No one has broader international contacts or is more respected among those groups both with practitioners and academics.

**Question c:** Only answer this question in case the applicant argued that KU is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why KU is not applicable for his/her research project (see also the information under criterion 3 listed above)?

**Comments:** Not applicable

**Final assessment**

**Question a:** How do you assess the entire application? Please give your final scoring (A+/A/B/UF/U).

**Comments referee 01, A+** Highest quality, significance and recommendation for funding

**Question b:** Could you please summarize (point by point) the strengths and weaknesses of the grant application focusing on the candidate, proposal and knowledge utilisation? Strengths:

- This researcher is ideally placed to effect knowledge utilisation. No one has broader international contacts or is more respected among those groups both with practitioners and academics.

- This is a potentially revolutionary project proposed by the person most able to complete it which holds the potential for impact in academia, practice, governmental, and non-governmental organizations across the globe.

We do not have this information in the field and it is desperately needed.

**Data management**

**Explanation.** Responsible data management is part of good research. For the collection/generation of data and the analysis of this data, timely measures need to be taken to ensure the storage and later reuse of the data. This means that prior to the start of the research project researchers must ascertain a) which data could be relevant and b) how these data could be stored. After a proposal has been awarded funding, the researcher will draw up a detailed data management plan in which the researcher explains how all relevant data will be made findable, accessible, interoperable and reusable (FAIR). The data management section( 2e) is not an assessment criterion for the research proposal. However for all the data management sections of these proposals, you
can make suggestions and give advice that could be helpful for the researcher in drawing up the data management plan to be submitted after funding was awarded. See also www.mwo.nl/datamanagement

**Question** Datamanagement: advice and/or suggestions

**Comments:** The data management approach appears appropriate.

---

**Referee report of referee 2**

**Assessment of the quality of the researcher**

**Explanation.** Explanation. Criteria - the quality of the researcher

- in terms of profile fit in the target group;
- in the top 10% of his/her international peer group;
- academic excellence as demonstrated by numerous publications of international standing and/or other academic achievements;
- inspiring enthusiasm for research and/or technology;
- persuasiveness;
- demonstrably capable of successfully developing own new innovative line of research;
- has both a national and international prominent position;
- demonstrable leadership and coaching skills.

The Vici scheme aims at outstanding researchers only: the top 10% of his/her international peer group.

**Question a:** What is your opinion on the past performance of the researcher (as demonstrated by publications and other relevant scientific achievements)?

**Comments:** Dr. Bekkers has proven to be a leader in the field of philanthropic studies. His 2011 literature review on the matter is a "go-to" in the literature and the networks he has established for the widespread study of this topic have been extremely influential. His productivity is strong (CV lists 50 peer reviewed papers, 60 chapters, and 6 books), but not exceptional for someone of his status/level/career stage. However, Dr. Bekker is well-informed, appears to have strong leadership/mentoring skills, is very well connected with others who can assist with his vision, and an exceptional enthusiasm for the topic and project.

**Question b:** Does the applicant belong to the top 10% of his/her international peer group (Please indicate which Group/Research Group or Peer Group you are referring to in this comparison.)? Which scientific achievements or talents of the applicant show that he/she belongs to this top?

**Comments:** I would say that Dr. Bekker is in the top 10% of researchers studying philanthropic science. Perhaps the strongest piece of support for this claim is his exceptional network in ERNOP.

**Question c:** To what extent is there sufficient evidence that the applicant has the ability to lead and supervise a research group and support staff and to coach young researchers?

**Comments:** My read of this application indicates that Dr. Bekker has mentored (or is currently mentoring) 3 doctoral students and 3 post-docs. For his career stage, this seems a little low, especially since the current proposal suggests employing 1 post-doc and 3 new PhD students.

However, Dr. Bekker notes that he has mentored 25 other researchers at various stages, so that may indicate a longer history of supervision than I recognize.

---

**Assessment of the quality, innovative character and academic impact of the proposed research**

**Explanation.** Criteria - the quality, innovative character and academic impact of the proposed research

- challenging content;
- originality of the research question;
- innovative scientific elements;
- aimed at building up a new line of research;
- potential to make important contributions to science;
- effectiveness of proposed methodology;
- international importance of the proposed research area.

**Question a:** Please comment on the relevance of the problem and on the originality and challenging content of the proposal.

**Comments:** The current proposal seeks to develop a large, international survey to probe several unanswered questions about the frequency, nature, and motivation for philanthropy. The proposal has numerous strengths, including the use of consistent methods/survey questions to allow comparison, large samples for robust estimates, as well as a mix of self-report and behavioural measures. Moreover, the proposal seems to be poised to explore multi-level processes (at the national, meso, and micro-level), as well as test several complementary and competing models.

To my knowledge, the scope and focus of this project is extremely novel. As the applicant notes, several existing and very large surveys do capture philanthropy around the globe using representative samples (e.g., the Gallup World Poll), but the items probing philanthropy are shallow (e.g., did you donate to charity in the past month? yes or no). As such, a larger and more thorough survey probing the nature, frequency, and motivation for formal giving is of value, especially if objective behaviour and other level variables are captured as well.

**Question b:** What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?

**Comments:** I think this project has the potential to reveal new, multi-level understanding of the variables that shape human philanthropy and generosity.

**Question c:** What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?

**Comments:** I think this work has the potential to make a large impact on our understanding of philanthropy.

**Question d:** To what extent is the proposed method effective? Please comment.

**Comments:** I appreciate the space limitations of the current proposal limit the applicant’s ability to discuss ideas in great detail.

However, there were several questions I had about the proposed methods. For instance, the applicant mentions a clear focus on the measurement and exploration of “formal philanthropy” and seeks to provide a thorough mapping of this behaviour in a large swath of representative samples around the world. While I certainly admire and see value in this goal, I wonder how limiting one’s focus to formal philanthropy restricts the samples and likely findings. However, in many places around the world, people may not have the opportunity to
engage in formal philanthropy due to their location, government, etc. (e.g., people living in rural areas around the world may not donate to organized charities because they aren’t asked, or lack a means of doing so). In fact, formal philanthropy may be an odd feature of WEIRD /developed countries where infrastructure allows for more far-reaching assistance. As such, I wonder whether exploring formal philanthropy in rural India, China, Korea, etc may pose large logistical and interpretation challenges? Will respondents understand what is being asked? And how will respondents likely low participation rates be informative? I think Dr. Bekker hints at this challenge by mentioning that some samples (e.g., Africa) are too costly or impractical at this stage.

I also wasn’t certain whether the intention was to collect data from representative samples in each nation? That was suggested early on but not reiterated later (page 14 or 18).

Also, I was confused about the eventual reach of the Global Giving survey. The applicant criticized previous surveys for their limited scope and reliance on WEIRD samples, but page 18 of the application indicates that the Global Giving survey would be conducted in 31 countries focusing primarily on the EU and other wealth nations. If so, I think the novelty and impact of the project may be weaker than I anticipated.

At some points, the applicant mentions a greater interest in prosociality more broadly (e.g., blood donation or informal assistance), which I think is of great value, especially if these concepts will be assessed similarly in multiple far-reaching locations!

**Assessment of the knowledge utilisation**

**Explanation.** Criteria - knowledge utilisation (KU)

**Potential**

▷ contribution to society and/or other academic areas;
▷ disciplines and organisations that might benefit from the results.

**Implementation**

▷ action plan to allow the outcomes of the research project to benefit the potential knowledge users;
▷ if and how the potential knowledge users will be involved;
▷ (concrete) outcomes for society and/or other academic disciplines;
▷ the period over which knowledge utilisation is expected to occur.

The selection committee assesses:

▷ whether the applicant has given a realistic description of the potential for knowledge utilisation;
▷ and to what extent the applicant has presented a concrete and convincing plan for the implementation of the available potential.

If a researcher is of the opinion that the proposed research is not appropriate for KU then he/she should explain why he/she thinks that KU is not applicable. The selection committee will assess the arguments given for this

**Question a:** What is your opinion on the described potential for knowledge utilisation?

**Comments:** I agree with Dr. Bekkers that a deeper understanding of the various factors that shape human generosity (or formal philanthropy, which is more limited, but still of value) would be great value to charities and governments around the globe. I think the applicant’s plans to share this information sounds wise and ambitious, which I appreciate!

**Question b:** Please comment on the effectiveness and feasibility of the proposed approach for realizing KU (implementation).

**Comments:** I think Dr. Bekkers connections in the philanthropy science network and more broadly make his KU plan very likely to see realisation.

**Question c:** Only answer this question in case the applicant argued that KU is not to be expected given the nature of the research proposal! Does the applicant convincingly explain why KU is not applicable for his/her research project (see also the information under criterion 3 listed above)?

**Comments:** N/A

**Final assessment**

**Question a:** How do you assess the entire application? Please give your final scoring (A+/A/B/U/F/U).

**Comments referee 02, A High quality, significance and recommendation for funding**

**Question b:** Could you please summarize (point by point) the strengths and weaknesses of the grant application focussing on the candidate, proposal and knowledge utilisation? Strengths:

**Weaknesses:**

**Comments:** I listed several strengths and weaknesses above. One additional thought: The applicant indicates that the amassed data will allow for exploration into several competing/complimentary theories. Perhaps due to the space restrictions, I wasn’t clear on how these data would adjudicate between the various explanations.

**Datamanagement**

**Explanation.** Responsible data management is part of good research. For the collection/generation of data and the analysis of this data, timely measures need to be taken to ensure the storage and later reuse of the data. This means that prior to the start of the research project researchers must ascertain a) which data could be relevant and b) how these data could be stored. After a proposal has been awarded funding, the researcher will draw up a detailed data management plan in which the researcher explains how all relevant data will be made findable, accessible, interoperable and reusable (FAIR). The data management section( 2e) is not an assessment criterion for the research proposal. However for all the data management sections of these proposals, you can make suggestions and give advice that could be helpful for the researcher in drawing up the data management plan to be submitted after funding was awarded. See also www.nwo.nl/datamanagement

**Question Datamanagement: advice and/or suggestions**

**Comments:** I very much admire the applicant’s commitment to open science and how these data will be shared with the larger academic community. This reality makes the impact of the project much larger, as it allows others to learn and explore the wealth of information collected.

| Referee report of referee 3 |

**Assessment of the quality of the researcher**

**Explanation.** Explanation. Criteria - the quality of the researcher

▷ in terms of profile fit in the target group;
▷ in the top 10% of his/her international peer group;
academic excellence as demonstrated by numerous publications of international standing and/or other academic achievements;
> inspiring enthusiasm for research and/or technology;
> persuasiveness;
> demonstrably capable of successfully developing own innovative line of research;
> has both a national and international prominent position;
> demonstrable leadership and coaching skills.
The Vici scheme aims at outstanding researchers only: the top 10% of his/her international peer group.

Question a: What is your opinion on the past performance of the researcher (as demonstrated by publications and other relevant scientific achievements)?

Comments: Dr. Bekkers’s past performance is outstanding. His literature review of eight empirical mechanisms of philanthropy is cited and used by everyone in the field of charitable giving studies, and his other publications are in highly respected, highly ranked journals. He publishes in a breadth of journals, from political psychology to social psychology to sociology. His work is widely cited. His track record of gaining funding is impressive for any researcher, and particularly impressive given the difficulties of finding funding for charitable giving research.

Question b: Does the applicant belong to the top 10% of his/her international peer group (Please indicate which Group/Research Group or Peer Group you are referring to in this comparison.)? Which scientific achievements or talents of the applicant show that he/she belongs to this top?

Comments: Dr. Bekkers belongs to the top 10% - probably the top 1% - of the international peer group of scholars of nonprofits, charitable giving and prosocial behavior. The scientific achievements that justify this ranking are his extensive and high quality publications, the fact that his publications are widely cited, his ability to gain grant funding, and his extensive sharing of data with other researchers.

Question c: To what extent is there sufficient evidence that the applicant has the ability to lead and supervise a research group and support staff and to coach young researchers?

Comments: Dr. Bekkers is currently supervising a research team and has supervised numerous PhD students, postdocs and other researchers over the years. He has a proven track record of managing large grants. Assessment of the quality, innovative character and academic impact of the proposed research

Explanation. Criteria - the quality, innovative character and academic impact of the proposed research

> challenging content;
> originality of the research question;
> innovative scientific elements;
> aimed at building up a new line of research
> potential to make important contributions to science
> effectiveness of proposed methodology;
> international importance of the proposed research area.

Question a: Please comment on the relevance of the problem and on the originality and challenging content of the proposal.

Comments: Dr. Bekkers’ proposal takes on a long-standing and important set of problems in our discipline: why do levels of charitable giving vary so widely from country to country, and among individuals and groups within countries? This is an intellectually relevant problem in that it deals with cross-national variation in social capital and community participation, issues of concern to multiple disciplines, including sociology, psychology, political science and economics. It is a policy-relevant problem because charitable giving makes possible many social goods through the operation of nonprofit organizations. The proposal is original and challenging - details on what makes the proposal innovative are below.

Question b: What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?

Comments: There are many features that make the proposal innovative: 1) It derives theories about charitable giving from multiple social science disciplines. 2) It collects and combines data from many countries. 3) It measures influences on charitable giving behavior at three levels: national, group and individual. 4) It uses more than one method, relying mainly on survey data but building experiments into the survey methodology. 5) It collects and makes available data that will be widely available to other researchers. 6) It encourages cooperation among researchers from many different disciplines

Overall, the project tests existing theories but is innovative in that it combines many theories from many different disciplines, and tests them using innovative and extensive methods.

Question c: What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?

Comments: Dr. Bekkers’s research should make a major contribution to the advancement of scientific knowledge of charitable giving. It brings together the main theories of variation in charitable giving from a number of academic disciplines and tests them in a single multi-national dataset. It allows for tests of participation in other forms of prosocial behavior such as volunteering, blood and organ donation, and person to person helping. The data he collects will be widely used by scholars in economics, sociology, psychology, marketing, public administration and political science to solve the question of why people engage in prosocial behavior and why participation in prosocial behavior varies among countries, institutions and individuals.

The question of why people engage in prosocial behavior is a vital one both scientifically and practically, and this project will mark a major step towards answering that question.

Question d: To what extent is the proposed method effective? Please comment.

Comments: In this proposal, Dr. Bekkers uses the most up to date and effective methods in survey and experimental research on charitable giving. Many of these methods were developed by Dr. Bekkers himself in his excellent Giving in the Netherlands Panel Study. Others are adapted from new work by others. The most important aspect of the methods is the size of the data set collected, its diversity and breadth in selection of countries, and the use of macro level (countries), meso (organizations) and micro level (individuals) variables. The proposal anticipates many of the problems often encountered in philanthropy research, such as cross-cultural differences in the definition and conceptualization of charitable giving, and solves them using methods that Dr. Bekkers developed in his earlier survey projects.
Assessment of the knowledge utilisation

Explanation. Criteria - knowledge utilisation (KU)

Potential

➢ contribution to society and/or other academic areas;
➢ disciplines and organisations that might benefit from the results.

Implementation

➢ action plan to allow the outcomes of the research project to benefit the potential knowledge users
➢ if and how the potential knowledge users will be involved
➢ (concrete) outcomes for society and/or other academic disciplines;;
➢ the period over which knowledge utilisation is expected to occur.

The selection committee assesses:

➢ whether the applicant has given a realistic description of the potential for knowledge utilisation
➢ and to what extent the applicant has presented a concrete and convincing plan for the implementation of the available potential.

If a researcher is of the opinion that the proposed research is not appropriate for KU then he/she should explain why he/she thinks that KU is not applicable.

The selection committee will assess the arguments given for this

Question a: What is your opinion on the described potential for knowledge utilisation?

Comments: Dr. Bekkers is correct in asserting that this research will help solve challenges in other academic disciplines;. While his own degree is in sociology, the questions asked and theories used to answer them draw from and are relevant to law, psychology, economics, marketing, public administration and political science. The knowledge generated by this project will be widely utilized by academics working outside of Dr. Bekkers’ research team, as he is making the data publicly available. The data set that he collects will be far the most extensive data set on charitable giving in the world, and will also include data on volunteering, blood and organ donation, and informal person to person helping, making it useful to answer questions about other forms of prosocial behavior than charitable giving.

Question b: Please comment on the effectiveness and feasibility of the proposed approach for realizing KU (implementation).

Comments: The project will effectively realise knowledge utilisation by making the data, methods and analyses openly available to other researchers. The dataset will be the highest quality and most extensive data available on charitable giving, replicating and extending the methods of the successful and highly used Philanthropy Panel Study (US) and Giving in the Netherlands Panel Study, and extending them to many other countries. As philanthropy and prosocial behavior are major topics of interest in every social science discipline, researchers will be eager to use these high quality cross-national data. In addition, Dr. Bekkers has built into the project numerous mechanisms to ensure that other researchers take an interest in the data, including competitive grant proposals, conferences, a website and the presence of an extensive international review board.

Question c: Only answer this question in case the applicant argued that KU is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why KU is not applicable for his/her research project (see also the information under criterion 3 listed above)?

Comments: This question is not required.

Final assessment

Question a: How do you assess the entire application? Please give your final scoring (A+/A/B/U/F/U).

Comments referee 01, A+ Highest quality, significance and recommendation for funding

Question b: Could you please summarize (point by point) the strengths and weaknesses of the grant application focusing on the candidate, proposal and knowledge utilisation? Strengths:

Comments: Candidate: Dr. Bekkers is widely known and highly respected within the field of nonprofit and philanthropic studies. His prior work is widely cited and published in top quality journals. I judge him to be in the top 1% of academics in our field. Dr. Bekkers has long and wide experience managing large grants and Proposal: The proposal is exceptional because it draws upon theories from multiple academic disciplines, uses the best available knowledge on data collection to create a sound methodology for cross-cultural research, and will collect data in multiple countries and at multiple levels of analysis.

Knowledge utilisation: The proposal will create a publicly available dataset useful to researchers in a number of social science disciplines, and will answer a question of practical and social relevance and utility. Weaknesses: None. If funded, this project would be a dream come true for hundreds of researchers of philanthropy and prosocial behavior, answering some of the most important questions in the field and creating a publicly available dataset for others to work on their own research agendas.

Datamanagement

Explanation. Responsible data management is part of good research. For the collection/generation of data and the analysis of this data, timely measures need to be taken to ensure the storage and later reuse of the data. This means that prior to the start of the research project researchers must ascertain a) which data could be relevant and b) how these data could be stored. After a proposal has been awarded funding, the researcher will draw up a detailed data management plan in which the researcher explains how all relevant data will be made findable, accessible, interoperable and reusable (FAIR). The data management section( 2e) is not an assessment criterion for the research proposal. However for all the data management sections of these proposals, you can make suggestions and give advice that could be helpful for the researcher in drawing up the data management plan to be submitted after funding was awarded. See also www.nwo.nl/datamanagement

Question Datamanagement: advice and/or suggestions

Comments: Dr. Bekkers has extensive experience managing and making publicly available data from his Giving in the Netherlands Philanthropy Study. His proposal for managing and making available the data from this project is sound and I have no recommendations for data management outside of what he has already outlined.
Assessment of the quality of the researcher

Explanation. The applicant has demonstrated this ability widely in his past research including European FP7 projects, Research cooperation with Dutch and international teams in other projects, and activities such as co-founding the European Research Network on Philanthropy (ERNOP) in which he has also taken a lead in developing substantial international cooperation projects.

In addition the PhDs and Post-docs he supervised have themselves developed into esteemed and successful members of the European research community whom the reviewer has had an opportunity to be in contact with directly.

Assessment of the quality, innovative character and academic impact of the proposed research

Explanation. To what extent is there sufficient evidence that the applicant has the ability to lead and supervise a research group and support staff and to coach young researchers?

Comments. The applicant has demonstrated this ability widely in his past research including European FP7 projects, Research cooperation with Dutch and international teams in other projects, and activities such as co-founding the European Research Network on Philanthropy (ERNOP) in which he has also taken a lead in developing substantial international cooperation projects.

The Vici scheme aims at outstanding researchers only: the top 10% of his/her international peer group.

Question a: What is your opinion on the past performance of the researcher (as demonstrated by publications and other relevant scientific achievements)?

Comments. The applicant has an excellent track record of publications and academic achievements. He has advanced the knowledge on philanthropic giving systemically and has established with few colleagues this segment of research in Europe starting from a wstrong base of such work on the Netherlands. In addition he has contributed methodologically to position this research on Europe in a top tear compared international standards.

Pof. Bekkers has not only published widely and on both Dutch and international levels but has attracted very substantial amounts of grant funding to his field of research which has given additional momentum to research on giving, volunteering, the non-profit sector and social innovation.

Question b: Does the applicant belong to the top 10% of his/her international peer group (Please indicate which Group/Research Group or Peer Group you are referring to in this comparison.)? Which scientific achievements or talents of the applicant show that he/she belongs to this top?

Comments. The applicant certainly belongs to the top 10% of his international peer group of social scientist with an emphasis on philanthropy, the non-profit sector and social innovation.

In todays world private contributions and approaches towards social innovation and social problem-solving attract increasing attention first of all of society and public policy in general and secondly (and still with limited interest) the academic community in particular. In this context it is highly relevant to develop more profound knowledge on the circumstances and determinants which influence the behaviour of private individuals to contribute their resources to such activities. The current state of knowledge in these giving activities is still scattered and based on driven by very heterogenous theoretical assumptions. In addition data it at all are typically available for single countries and largely lack a comparative dimensions. As a consequence our comparative knowledge on levels of giving in different countries if fragmented and incomplete and merits substantial efforts to be developed.

This body of research has additional relevance because it sheds light on the forces of solidarity, care and altruism. In situations of alleged social disintegration a better understanding of influencing factors prompting or inhibiting these forces gains even more relevance.

The particular promise which this strand of research holds lies in its immediate relevance for policy makers and leadership of non-profit organizations alike who will be able to develop their resource strategies and their overall strategies based on more stringent evidence. The applicant has strong potential to contribute towards this due to his already existing involvement in practitioner and policy networks.

Question b: What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?

Comments. This pioneer proposal (the first comparative global study on giving) addresses innovation of several different levels: - Multinational: It breaks new ground in terms of internationally comparative knowledge in order to map and explain country differences in giving. - Multi-theoretical: It intends to integrate a currently heterogeneous base of theories explaining giving behaviour. - Multi-method: It builds its approach both on the “mega-analysis” of existing data from previous surveys which the applicant has already standardized to allow for secondary use and benefit and a new approach combining survey and experimental research methods. - Multi-regression modelling based on large
Questions on the proposal:

**Question a:** What is your opinion on the described potential for knowledge utilisation?

**Comments:** The project will make a very substantial contribution towards the advancement of interdisciplinary knowledge on giving in a global perspective. At the same time, it will contribute to the development of a much better theoretical understanding of the reasons why people give or behave in a pro-social way. The project will also have a profound methodological effect by developing a body of data open for further use (open-source approach). The structure of the project will also hold good promise for a flow of top publications both by the project team members themselves but also by international colleagues in the network who will be able to explore subquestions by utilizing the data provided by this project.

**Question b:** To what extent is the proposed method effective? Please comment.

**Comments:** The proposed method, in particular the combination of using existing data in a "mega-analysis" and producing new data in a mixed-method combination of survey and experiments promises to be very effective. In terms of connecting the work of the project to an international network of research colleagues, this will ensure that KU will not only be a hypothetical option but a realistic perspective. The method, data, and publications will therefore effectively influence the research landscape for a very long time.

**Question c:** What is your opinion on the proposed method for realizing KU (implementation)?

**Comments:** The provisions described are convincing both technically and in terms of connecting the work of the project to an international network of research colleagues form the outset. This will ensure that KU will not only be a hypothetical option but a realistic perspective. The method, data, and publications will therefore effectively influence the research landscape for a very long time.

**Question d:** To what extent is the proposed method effective? Please comment.

**Comments:** The proposed method, in particular the combination of using existing data in a "mega-analysis" and producing new data in a mixed-method combination of survey and experiments promises to be very effective. In terms of connecting the work of the project to an international network of research colleagues, this will ensure that KU will not only be a hypothetical option but a realistic perspective. The method, data, and publications will therefore effectively influence the research landscape for a very long time.

**Assessment of the knowledge utilisation**

**Explanation, Criteria - knowledge utilisation (KU)**

**Potential**
- contribution to society and/or other academic areas;
- disciplines and organisations that might benefit from the results.

**Implementation**
- action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- if and how the potential knowledge users will be involved;
- (concrete) outcomes for society and/or other academic disciplines;
- the period over which knowledge utilisation is expected to occur.

The selection committee assesses:
- whether the applicant has given a realistic description of the potential for knowledge utilisation;
- and to what extent the applicant has presented a concrete and convincing plan for the implementation of the available potential.

If a researcher is of the opinion that the proposed research is not appropriate for KU then he/she should explain why he/she thinks that KU is not applicable.

The selection committee will assess the arguments given for this
certain risk. However existing networks and trust relationships of the PI will most likely successfully mitigate this risk.

KU:

Strengths: As a reference project in methodology and data this project holds a very strong potential for KU both in longitudinal and cross-sectional levels. The provisions to secure KU are appropriate.

Weaknesses: Maintenance of the data repositories not fully under control of the PI/project, but the proposed choice of institutional open source repositories will be the best possible mitigation of this risk.

**Datamanagement**

**Explanation.** Responsible data management is part of good research. For the collection/generation of data and the analysis of this data, timely measures need to be taken to ensure the storage and later reuse of the data. This means that prior to the start of the research project researchers must ascertain a) which data could be relevant and b) how these data could be stored. After a proposal has been awarded funding, the researcher will draw up a detailed data management plan in which the researcher explains how all relevant data will be made findable, accessible, interoperable and reusable (FAIR). The data management section (2e) is not an assessment criterion for the research proposal. However for all the data management sections of these proposals, you can make suggestions and give advice that could be helpful for the researcher in drawing up the data management plan to be submitted after funding was awarded. See also www.nwo.nl/datamanagement

**Question Datamanagement: advice and/or suggestions**

**Comments:** Given the multi-country and multi-method approach of the project I suggest to consider/negotiate simultaneous data storage with repositories in different countries at the same time. This might help to ascertain longest possible and most widespread KU.

<table>
<thead>
<tr>
<th>Referee report of referee 5</th>
</tr>
</thead>
</table>

**Assessment of the quality of the researcher**

**Explanation.** Criteria - the quality of the researcher

▷ in terms of profile fit in the target group;
▷ in the top 10% of his/her international peer group;
▷ academic excellence as demonstrated by numerous publications of international standing and/or other academic achievements;
▷ inspiring enthusiasm for research and/or technology;
▷ persuasiveness;
▷ demonstrably capable of successfully developing own new innovative line of research;
▷ has both a national and international prominent position;
▷ demonstrable leadership and coaching skills.

The Vici scheme aims at outstanding researchers only: the top 10% of his/her international peer group.

**Question a:** What is your opinion on the past performance of the researcher (as demonstrated by publications and other relevant scientific achievements)?

**Comments:** Prof. Bekkers is in the top 1% of scholars in his field. He has an extraordinary array of publications and projects. His research is always creative, insightful, thorough and well-argued. It always includes a strong theoretical model and comprehensive literature review. He brings new insights to old research questions and creates new questions for the field. The fact that he started a panel study of philanthropy in the Netherlands was a huge undertaking, yet he did that, so it does not surprise me that he is "going global" in his research design.

I have not looked, but if somebody asked me who was the most highly cited scholar in philanthropy in the last decade, I would guess Dr. Bekkers.

**Question b: Does the applicant belong to the top 10% of his/her international peer group (Please indicate which Group/Research Group or Peer Group you are referring to in this comparison.)? Which scientific achievements or talents of the applicant show that he/she belongs to this top?**

**Comments:** Dr. Bekkers is clearly within the top 10% of scholars studying philanthropy and nonprofits (NPOs or NGOs). In fact, I would rank him in the top 1%. He has refereed papers in top journals on all aspects of giving, volunteering, government crowding out (or in) of charitable giving, as well as the effects of faith on philanthropy. He has created original data, analyzed secondary data and conducted experiments in key areas of philanthropy. His research is always well done and well regarded. He is widely cited by English-speaking authors despite the fact that a large share of his publications are in Dutch.

**Question c: To what extent is there sufficient evidence that the applicant has the ability to lead and supervise a research group and support staff and to coach young researchers?**

**Comments:** Dr. Bekkers has already supervised research groups, support staffs, and mentored/coached young researchers. I have tried to coax him here more than once, but he is very committed to his work in the Netherlands. My observations based on his interactions with our faculty, staff, students and his colleagues is that he would be a great mentor, coach, and leader.

**Assessment of the quality, innovative character and academic impact of the proposed research**

**Explanation.** Criteria - the quality, innovative character and academic impact of the proposed research

▷ challenging content;
▷ originality of the research question;
▷ innovative scientific elements;
▷ aimed at building up a new line of research
▷ potential to make important contributions to science
▷ effectiveness of proposed methodology;
▷ international importance of the proposed research area.

**Question a:** Please comment on the relevance of the problem and on the originality and challenging content of the proposal.

**Comments:** The problem is quite relevant in helping policy makers, practitioners, and scholars understand the size and nature of philanthropy around the world. Perhaps more importantly, by testing several competing and complimentary theories to explain who gives (or not), to what types of charities (or not), the role of taxes (or not), and the role of government spending (or not), this project will inform the both the scholarly fields and the practitioner/public policy fields of best strategies to pursue to maximize the pro social behaviors in ways to increase total philanthropic giving and perhaps to do so in a manner to maximize the impact of these gifts.
A challenge of this study of course is its broad vision to work across several countries and continents to gather parallel data from each to address the same theoretical and empirical/practical concerns around philanthropy and public policy.

**Question b:** What are the innovative aspects of the proposal? Will the research break new ground by generating new concepts, a deeper understanding, new methods, etc.?

**Comments:** 1. Maps the philanthropic landscape in 145 countries on every continent. 2. Uses meaningful sample sizes (n>1,000) in each country, which enhances the types of analyses and the accuracy of each that can be conducted. 3. Tests multiple types of philanthropic behaviors both formal (giving money and volunteering to charities) and informally (donating blood, helping friends, relatives, and others not via charities). 4. Test Macro variables, Meso variables, and Micro variables and their impact on household giving. 5. Tests competing theories of giving. 6. Tests the effects of government policies such as taxes, tax deductions, but also government spending. 7. Understand who gives, why and where?

**Question c:** What is your opinion on its potential to make a major contribution to the advancement of scholarship, science or technology (academic impact)?

**Comments:** The academic impact will be path breaking. Many of these issues in this project have been studied separately in one country or another at a time. However, this is the first to study all of the key theories simultaneously and in so many countries. This will provide much greater understanding about which variables affect philanthropy and precision around how large or small the effects are for each variable, as well as how national and organizational variables may provide contextual effects.

**Question d:** To what extent is the proposed method effective? Please comment.

**Comments:** I have used some of these same effects in studying philanthropy in the USA and found them to be highly effective. Using an intentional theoretical set of nested models will enhance the value of the methods and outcomes.

**Assessment of the knowledge utilisation**

**Explanation.** Criteria - knowledge utilisation (KU)

**Potential**
- contribution to society and/or other academic areas;
- disciplines and organisations that might benefit from the results.

**Implementation**
- action plan to allow the outcomes of the research project to benefit the potential knowledge users;
- if and how the potential knowledge users will be involved;
- (concrete) outcomes for society and/or other academic disciplines;
- the period over which knowledge utilisation is expected to occur.

The selection committee assesses:
- whether the applicant has given a realistic description of the potential for knowledge utilisation;
- and to what extent the applicant has presented a concrete and convincing plan for the implementation of the available potential.

If a researcher is of the opinion that the proposed research is not appropriate for KU then he/she should explain why he/she thinks that KU is not applicable.

The selection committee will assess the arguments given for this

**Question a:** What is your opinion on the described potential for knowledge utilisation?

**Comments:** The proposed research is both relevant and doable. If carried out as propose, it will utilizes the extent knowledge in the filed but also the expertise from an active advisory board of scholars in this space from around the world.

Given the changing values around the world away from public funding and towards private philanthropic support the scientific value of this research will only be eclipsed by its applied value.

Given its multi-disciplinary approach, this work would enhance the understanding of philanthropy overall and through the lens of many specific disciplines ranging from sociology, Economics, political science, geography, public policy, etc.

**Question b:** Please comment on the effectiveness and feasibility of the proposed approach for realizing KU (implementation).

**Comments:** I think this is very achievable with adequate funding. I have no knowledge of the budget other than what is delineated (pp 29-30), but this would be a great return on these investments if made.

**Question c:** Only answer this question in case the applicant argued that KU is not to be expected given the nature of the research proposal: Does the applicant convincingly explain why KU is not applicable for his/her research project (see also the information under criterion 3 listed above)?

**Comments:** N/A

**Final assessment**

**Question a:** How do you assess the entire application? Please give your final scoring (A+/A/B/UF/U).

**Comments referee 01, A+** Highest quality, significance and recommendation for funding

**Question b:** Could you please summarize (point by point) the strengths and weaknesses of the grant application focussing on the candidate, proposal and knowledge utilisation? Strengths:

**Weaknesses:**

**Comments:** I think I have outlined these above, but I think the project does a great job of being under-girded by theory, built on a model of testable hypotheses, and would have tremendous academic, practitioner, and public policy benefits.

I wish I had thought of it first!!!

**Datamanagement**

**Explanation.** Responsible data management is part of good research. For the collection/generation of data and the analysis of this data, timely measures need to be taken to ensure the storage and later reuse of the data. This means that prior to the start of the research project researchers must ascertain 1) which data could be relevant and b) how these data could be stored. After a proposal has been awarded funding, the researcher will draw up a detailed data management plan in which the researcher explains how all relevant data
With respect to the criterion of Applicant Quality the reviewers a) qualify my
viability with such positive qualifications indicates a high reliability. In this letter I will
The five reviewers expressed great enthusiasm for my proposal. The reviews
Journal of Trial and Error
Center for Philanthropic Studies at Vrije Universiteit Amsterdam. Though there
continue to acquire them is not standard.
mentored 25 researchers in various stages which does indicate a longer history
correctly reflects that the number of supervision may seem low but I have
or 1% in my field, and c) express confidence in my ability to lead and supervise
researchers.
As R3 and R4 point out I have a proven track record in managing large
grants and have supervised numerous early stage researchers. However, R2
correctly reflects that the number of supervision may seem low but I have
mentored 25 researchers in various stages which does indicate a longer history
of supervision. I have acquired good supervision skills, though the path I took
to acquire them is not standard.
In 2008, I made a conscious decision to focus more strongly on philanthropy and I
moved from the Department of Sociology at Utrecht University to the
Center for Philanthropic Studies at Vrije Universiteit Amsterdam. Though there
were no graduate students that I could work with and the VU did not support
the Center with funding for research time, I have been able to acquire large
external grants. With these grants, and the work I have done in the European
Research Network on Philanthropy, I have demonstrated entrepreneurship and
vision in building a new field of research in Europe that is very much alive and
still growing today. The Center I lead is an attractive place also for researchers,
evidenced by the recent integration of a new ERC funded research project
“DONORS”, on blood donation including an associate professor, two graduate
students and a postdoc. In that perspective I appreciate the assessment by
R3 that my track record in gaining funding is ‘particularly impressive given
the difficulties of finding funding for charitable giving research’. Despite
unfavorable circumstances, I have been able to supervise three PhD students
and three postdocs. Having met them, R4 is convinced that I do this well, as
they have ‘developed into esteemed and successful members of the European
research community’.
In addition, I have directed the other staff employed by the Center for Phil-
anthropic Studies that I am leading since 2013. Also I have mentored dozens
of foreign graduate students in PhD seminars at the international ERNOP,
ARNOVA and ISTR conferences that I have co-organized and attended. People
in the philanthropy research community who know me will tell you that I am
very supportive of students and early career researchers. R2 says I have ‘strong
leadership and mentoring skills’. R3 notes that I have a ‘proven track record of
managing large grants’.

Rebuttal
File number
VLC.191.063
Name of candidate
René Bekkers
Title
Global Giving
General evaluation
The five reviewers expressed great enthusiasm for my proposal. The reviews are
A+ (R1), A (R2), A+ (R3), A+ (R4) and A+ (R5).
The ratings of the current reviewers are highly consistent with qualifications in
the previous two rounds for previous versions of this proposal – for previous
evaluations, see (https://renebekkers.wordpress.com/global-giving/). The fact
that this is the third time that a set of experts in the field evaluate the proposal
with such positive qualifications indicates a high reliability. In this letter I will
address the issues raised by the reviewers. I have addressed the issues that
you have raised with respect to the preproposal in a response posted at my
blog, (https://renebekkers.files.wordpress.com/2019/08/responses-to-reviews-
of-the-pre-proposal.pdf)

Quality of the researcher
With respect to the criterion of Applicant Quality the reviewers a) qualify my
research as innovative and leading internationally, b) place me in the top 10%
or 1% in my field, and c) express confidence in my ability to lead and supervise
researchers.
As R3 and R4 point out I have a proven track record in managing large
grants and have supervised numerous early stage researchers. However, R2
correctly reflects that the number of supervision may seem low but I have
mentored 25 researchers in various stages which does indicate a longer history
of supervision. I have acquired good supervision skills, though the path I took
to acquire them is not standard.
In 2008, I made a conscious decision to focus more strongly on philanthropy and I
moved from the Department of Sociology at Utrecht University to the
Center for Philanthropic Studies at Vrije Universiteit Amsterdam. Though there
will be made findable, accessible, interoperable and reusable (FAIR). The
data management section (2e) is not an assessment criterion for the research
proposal. However for all the data management sections of these proposals, you
can make suggestions and give advice that could be helpful for the researcher
in drawing up the data management plan to be submitted after funding was
awarded. See also www.nwo.nl/datamanagement

Question. Datamanagement: advice and/or suggestions
Comments: I admire and respect the commitment to making the data avail-
able and using an open science approach!!! I am confident the team will ensure
all appropriate privacy concerns are addressed, but just to acknowledge that
these become bigger issues when going global and when using an open science
approach.

Quality, innovative character and academic impact of the proposed research
With respect to the Proposal Quality, the reviewers state that a) the project is
‘extremely novel’ (R2), and will have a substantial impact not only ‘for the
entire field of research on philanthropy’ (R1), but also ‘in economics, sociology,
.... public administration and political science’ (R3) and ‘psychology, law,
marketing and communication science’ (R5). The research is b) ambitious in its
scope, c) with ‘greater potential than any project seen in this field in the last
10 years’ (R1), and d) using appropriate and innovative methods.
R2 raises questions on the project’s focus on formal philanthropy, particu-
larly in relation to the global scope of the research. Perhaps formal philanthropy
is ‘an odd feature of developed countries’. R2 also suggests that the measure-
ment of formal philanthropy will be difficult, especially in rural areas and when
participants do not have extensive experience with it.
Giving a fair description of the level and nature of giving, without a cultural
bias favoring western traditions is crucial for the success of Global Giving. The
comments voice a well-known concern with the validity of the measurement
of philanthropy. As explained on page 14 of the proposal, we will not impose
a specific (western) wording and avoid moral terms, but instead define giving
in behavioral terms. We will ask survey participants whether they have spent
money on others directly or indirectly through organizations. We will work with
an international network of country experts to develop a country specific set of
questions that are attuned to the local context, using examples of charitable
giving that are well known in the country, such as giving through WeChat
in China and zakat in Indonesia. In this way, we avoid erroneously giving
the impression that a country is not generous simply because its traditions
of helping are different. R3 states that ‘the proposal anticipates many of
With respect to Knowledge Utilisation, the reviewers state that the project will reach but realistic way. We will ask country experts to suggest policy issues that can be studied with the data we collect. Reviewers express confidence in my ability to connect these networks: ‘no one has broader international contacts or is more respected among both practitioners and academics’ (R1).

Furthermore, reviewers applaud the efficient use of ‘funds for research by re-using existing data’ (R4) and the open science approach, which make ‘the impact of the project much larger’ (R2). R3: ‘if funded, this project will be a dream come true for hundreds of researchers’.

In sum, the reviewers agree that this an excellent proposal worth funding. I am more than happy to answer any questions you may have during an interview.

**Decision**

**Project number**  
VI.C.191.063

**Applicant**  
Prof. dr. R.H.F.P. Bekkers

**Title**  
Global Giving

**Scores**

<table>
<thead>
<tr>
<th>Criteria 1</th>
<th>Criteria 2</th>
<th>Criteria 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0</td>
<td>2.4</td>
<td>3.2</td>
</tr>
</tbody>
</table>

Final qualification: Very good

**Quality of the researcher**

The reviewers assess the quality of the researcher as very good to excellent. All reviewers state that the researcher is within the top 10%, or even in the top 1% (R1, R3, R5) of his peer group. Reviewer 1 and 2 indicate that he is a leader in the field of philanthropy research. The researcher is cited and used by everyone in the field of charitable giving studies (R3), and he can be counted among the handful of peer leaders who have really advanced this field of research on a European level (R4). Besides, he is probably the most highly cited scholar in philanthropy in the last decade (R5). The committee agrees that the applicant is a leader in this particular field. The committee agrees with R2 that the productivity of his output is very good, but not exceptional for his status, level, or career stage. Furthermore, the committee agrees with R2 and R3 that the number of PhDs that he supervised is modest. The committee also notes that most of his top-publications are not from within the last five years. The committee is not completely convinced of the impact of the research of the applicant in other subfields and the scholarly contribution of this research outside the subfield. The committee considers the applicant as a well-recognized international expert on philanthropy and that he is a leading researcher in his niche focus area.

**Quality, innovative character and scientific impact of the research proposal**

The applicant proposes to examine different backgrounds for the generosity of citizens in 145 countries. The referees were very positive about the research proposal and call it a potentially revolutionary project (R1), the potential...
to reveal new, multi-level understanding of the variables that shape human philanthropy and generosity. The committee also considers the project to be very practical; in theoretical terms the proposal shows limited ambition and in that sense is less innovative. During the interview the applicant was asked if he had considered to include other factors like time and volunteering, or children, as the next generation for philanthropy. When answering these questions the applicant explained that other aspects are beyond the scope of the proposed project. Therefore, the committee concludes that the applicant stays within the rather specific niche of philanthropy and did not showed sufficiently the ambition to broaden his project to other disciplines. The committee assessed the quality, innovative character and scientific impact of the research proposal as very good. But still has reservations with regards to the research design.

Knowledge utilisation
The research proposal focuses on deeper understanding of global giving. The results of the study will be used for communication and dissemination of the outcomes to the target groups. The researcher will make all data open access and available to others. The applicant has a very strong track record. He is well connected to stakeholders in the charity business and he has an outstanding network, such as The European Research Network on Philanthropy. This ensures that the data will be shared within the focus field. All reviewers were positive about the plans for knowledge utilisation, because of the applicant’s track record and career. However, the committee was not completely convinced by the dissemination plans, as they find the niche-focus to be limited without indication of an active link to other scientific areas. The committee assessed the knowledge utilisation plan as limited, albeit very good.

References


affect donations to different types of charitable organizations. Social Science Research, 39(6), 1073–1087. https://doi.org/10.1016/j.ssrresearch.2010.06.005


View interactive version here.


**License**

Open Access. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit https://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020