By our editors

Systemic Problems in Academia: The Positive Publication Bias and Solutions from a Human Factors Perspective
by Stefan Gaillard and Sean Devine

Empirical/Editorial

The Complexity of Joint Regeneration: How an Advanced Implant Could Fail by Its In Vivo Proven Bone Component
by Paweena Diloksumpan et al.

Ponies, Joints, Complexity, and the Method of Difference
by Hubertus Nederbragt

Rewilding Cognition: Complex Dynamics in Open Experimental Systems
by Wendy Ross & Frédéric Vallée-Tourangeau

Cognition stays wild: A commentary on Ross and Vallée-Tourangeau’s Rewilding Cognition
by Vlad P. Glăveanu & Alex Gillespie
Empirical

Real-effort survey designs: Open-ended questions to overcome the challenge of measuring behavior in surveys

by Caroline Fischer

The Gloria Adherence Subproject: Problems and Randomization Mistakes

by Linda Hartman et al.

The Unsuccessful Self-treatment of a Case of 'Writer's Block': A Replication in Science Education

by Georgios Ampatzidis

Classical Conditioning for Pain: The Development of a Customized Single-Case Experimental Design

by Tamal Kumar De, Victoria J. Madden, Johan W. S. Vlaeyen, and Patrick Onghena

Do Carryover Effects Influence Attentional Bias to Threat in the Dot-Probe Task?

by Joshua W. Maxwell, Lin Fang, and Joshua Carlson
Rejected Grant Applications

*Communication Circuits and Inequalities of Health: A Case of Greenlanders in Denmark*

by Daria Schwalbe

*Are We Meeting Best Practice Standards?: A Longitudinal Analysis of Mental Health Practices Within the Florida Child Welfare System with Implications for Child Well-being*

by Daniel Dunleavy

*Neurological Markers of Maladaptive Brain Activity in Fibromyalgia and their Relationship with Treatment Effectiveness*

by Elia Valentini
Systemic Problems in Academia: The Positive Publication Bias and Solutions from a Human Factors Perspective

Stefan Gaillard¹,², Sean Devine³

It will come as no surprise to readers of this journal that failure is systematically ignored in academia. The entire culture of academia revolves around a notion of linear progress made in leaps and bounds by great thinkers, rather than the more realistic image of science as an ongoing process of trial and error. For those already convinced that error deserves a place in scientific discussion, it is often tempting to treat the dismissal of failure as a moral failure on the part of researchers and publishers—if only we were more courageous to do the work that science demands of us, we would publish our failures all the time; if it was not for journals rejecting my work because it is not "innovative" enough, I would tell everyone of my important failures. This perspective, while understandable, misses the mark. Specifically, it neglects the structure of modern academia that perpetuates an ideal of "success-first" science on all fronts—a structure, like any structure, that severely limits researchers’ and journals’ ability to change it. This point is described superbly in a recent editorial piece by public health researchers Sindall and Barrington (2020), who point out the many barriers to sharing and promoting scholarly understandings of failure in the public health interventionist field of Water, Sanitation, and Hygiene (WASH), which they experienced personally as well.

In their in the Journal of Trial and Error's (JOTE) first issue, the authors argue not only for the role of researchers themselves in pursuing submission and publication of reports of what went wrong, but also of funding partners in sharing information about projects that did not "turn out as planned," and importantly, for the responsibility of journals in welcoming such sort of publications. Critically, they argue that the widespread fixation on "novelty" contributes to a "culture of success-ism" that is "highly unscientific". In short, the outward-facing sheen of scientific success—propagated by scientists, funders, and stakeholders alike—challenges the very foundations scientific development is built on: transparency, collaboration, and trial and error.

Sindall and Barrington (2020) paint a compelling picture of an academic system that struggles to find a place for failure and the "multitude of sins" such a system necessarily perpetuates. Rightly so, they place the responsibility for change not only on researchers, but also funders and publication bodies. But the buck doesn’t stop here. The same culture of success-ism that limits researchers limits individual publishers as well. The problem is diffuse and the blame cannot be laid on a single "bad actor". Accordingly, we feel it pertinent to extend Sindall & Barrington’s argument by highlighting the structural barriers that restrict the spread of failure in the sciences, even at the editorial level. To do so, we use systems theory to frame three limitations journals face when publishing failure and how these limitations cannot simply be solved by good agents engaging in good science.

Failure and Systems

In the field of Human Factors, and more specifically in Forensic Human Factors, it has long been recognized that failure usually occurs in the complex environment of a system. When accidents such as the Chernobyl nuclear disaster occur, it is easy to shift blame to a single individual or to a single team that was present when mistakes occurred. In response to these disasters, these employees are subsequently reprimanded and sometimes fired. A new person might fill their position and, in some cases, training protocols might be updated. However, these are often very ineffective measures. As previously indicated, mistakes are made in a complex environment. This means that mistakes can have occurred for any number of reasons, and often because of a combination of reasons. Accordingly, removing just one factor from the environment (i.e., the employee present during the disaster) is unlikely to significantly diminish the chances of similar mistakes happening in the future, because it was unlikely to be the determining factor (Holden, 2009). Fields such as the aviation industry have long recognized that there are better alternatives to simply replacing or retraining staff following a negative incident. For instance, at the beginning of the proliferation of commercial flights, it became clear that cockpits were often too complicated for pilots to understand, which lead to mistakes being made, and consequently accidents occurred. Instead of firing...
the pilots who made mistakes and/or making small changes to training protocols, the aviation industry started making changes to the cockpits themselves. In other words, the industry transformed the environment, not the agents.

By increasing the simplicity of the cockpit, pilots were less taxed and could invest more cognitive resources towards flying the plane. An example of such an environmental change is the adaptation of cockpit alarms. In older planes, cockpits would have the same audio-visual cues for every alarm. Thus, the pilot got the same feedback from the system when the fuel was slightly low as when an important engine was malfunctioning. Pilots were then mentally taxed on two levels: they were constantly confronted with distracting alarms, some of which might not need their immediate attention; and when attending to these alarms, they had to invest great effort in distinguishing the alarm and assessing its importance. As a result, pilots learned to ignore most alarms. A research article published in 1996 summarized problems such as these at the time as follows: “The proliferation of alarms has led to the increased likelihood of false or multiple alarm events and the consequent inability of the operator to determine the underlying cause of the alarm(s)” (Gilson et al., 1996, pp. 12). Under such a system, a situation such as the following might occur: During a flight, one of the important engines malfunctioned and needs immediate attention. At the same time, a number of other small, but non-urgent, malfunctions could occur. Because the very attentive, but overly taxed pilot cannot properly differentiate between the different alarms, they end up paying attention to the wrong malfunction. A short while later, the plane crashes. As recently as 2019, a plane crash occurred due to a poorly designed alarm system (Levin, 2019).

Faced with the issue of the inattentive pilot, an airline company is presented with two solutions. On the one hand, the company could fire the pilot, hire a new one, and train them for a longer period of time, with a more experienced instructor, and with more expensive training equipment. Besides the fact that this would be very expensive, it would also be unlikely to solve the problem. Humans’ cognitive abilities are limited. This is because, like all human systems, their underlying biological systems operate on limited resources. Similarly, in the realms of cognition, some cognitive scientists believe you can train people for things like working memory, sustained attention, selective attention, and other cognitive skills, but all of these will have a natural ceiling—a point beyond which it is almost impossible to progress (Kwok et al., 2011; Traut et al., 2021).

In short, simply replacing or retraining bad agents doesn’t work on its own; we cannot reduce airplane crashes by just getting new and better pilots. As in aviation, the same can be said for scientific publication. The difference, however, as Sindall and Barrington point out, is that we are already in a nosedive when it comes to recognizing and publishing failure in science.

Obstacles to Failure

As hinted at in the last section, we do not believe the issues Sindall and Barrington (2020) bring up can be solved by simply creating better, more failure-friendly, journals. Rather, we argue that these issues are inherent in the structure of modern scientific practice. By “structural”, we mean that science as it is currently understood is defined by success; failure is not scientific under the current paradigm. Even when scientific failures are recognized, they are framed within the larger context of long-run success—“failing now to succeed later”—rather than as an integral part of the scientific process in its own right. In this regard, “structural” is quite literal, such that to displace the structure of success that builds up the modern scientific apparatus would mean challenging its integrity to the point of collapse. But where one structure collapses, another is built. The question is whether we need to be so radical. Can journals not simply reform their practices to allow for more diverse results to be published? Do we risk throwing the baby out with the bathwater?

In this section, we raise the possibility that editorial reform may not be enough, or at least, that it may be very difficult. To do so, we discuss three issues that journals, who wish to publish scientific failure in the current landscape, face. Throughout, we use our experience at JOTE as a touchstone and example for some of these issues.

Funding

It is JOTE’s mission to promote and normalize the process of failure in science. However, we recognize that actually doing so comes at high costs for new and established journals. As Sindall & Barrington highlight, scientific funding is a scarce resource. This is true not only for scientists themselves, but for the journals that publish their work; and especially true for journals that are sympathetic to publishing scientific failures. Funding bodies want to support a compendium of amazing, world-changing findings, not a collection of brick-building errors.

As a result of this scarcity, journals may turn to article processing charges (APCs) as a way to maintain their services. This business-model is effective when the journal can provide services that financially-secure journals cannot (open access, equity, less emphasis on positive results), but becomes intractable when additional social costs are added on top of the economic ones. These social costs refer both to those experienced by the journal—loss of reputation (more on this below), difficulty finding reviewers and submissions, accusations of predation—and by...
the submitters, who bear the brunt of the cost and, for social reasons (e.g., working at a lesser-known institution, being from a non-Western background, etc.) may be restricted in their available funds. This issue is exacerbated for journals that encourage researchers to publish their failures, which, because of the “culture of success-ism” in academia, is already perceived as a socially costly endeavor. As a result, publishers that aim to counter this culture start off in a precarious position, limited in terms of traditional funding and at a disadvantage in asking would-be authors—who themselves experience the social and professional costs of publishing null results—for support.

Therefore, journals willing to publish and promote failure in science are left scrumming for funds, often relying on the goodwill of the editorial team to keep them afloat. Alternatively, journals can support the publicizing of failure through the publicizing of success, spinning the failure as an altruistic side-benefit of an otherwise traditional, positive-result-publishing, outlet for scientific advancement. While sometimes the only option, this “failure as an afterthought” mindset often contributes to the very culture of success-ism that the publication of failure in science aims to combat.

Prestige
Above and beyond financial stability, journals comfortable with the idea of publishing failure face an uphill battle with their own reputation. While science makes room for failure, a scientific method cannot be said to exist without people enacting it. This so-called scientific method is typically associated with academia, and while Sindall and Barrington are right to encourage people to be more “scientific” by sharing failure, the problem is an academic system that does not. Thus, many would-be authors are disincentivized from submitting their work to failure-friendly journals for fear that publishing their null, negative, or unclear results would harm their reputation, decrease their chances to acquire funding in the future, and reflect poorly upon them as scientists. This reluctance to submit is a unique challenge that journals who publish failure face. Even positive-results-publishing low impact-factor journals have the benefit—if it can be called one—of researchers submitting their work there after they have exhausted more well-established outlets. In contrast, publishers that want to promote failure find themselves in the awkward position of having to convince researchers to publish their work at all (i.e., as opposed to scrapping the whole project), not to mention in their journal, for their peers to see.

Increasing the prestige of failure would contribute to an increased willingness on the part of academics to send in their failures. Examples of initiatives and changes that can increase the prestige of failure within academia include brilliant failure awards and changes to the priority rule. Brilliant failure awards, like other awards, signal what the awardee considers to be important and/or valuable. The priority rule refers to the tendency within science to award scientists who are the first to make a discovery, usually by giving them credit—for example, in the form of naming a theory or subject matter after them. Implicitly, the priority rule indicates that only the final success is worthy of reward, although this success often hasn’t been possible without the failures of others along the way. Changing these incentives can motivate more researchers to publish their failure.

At the same time, those researchers that do already want to publish about failure often encounter unexpected hurdles in the publication process. If the journal is subscription-based, the readership of an already niche topic—failure within some field of science—is reduced further, rendering the whole exercise much less appealing to the open-minded researcher who wants their failure to be read and understood. Conversely, if the journal is open-access, it is likely to rely on APCs, requiring researchers that send in an article to a journal to pay a fee upon submission. As discussed in the previous section (3.1), with the increasing trend towards open science, APCs are sometimes seen as necessary by open access journals to cover costs. However, these charges selectively disadvantage researchers at less financially prosperous institutions. Often funding to cover high article processing charges is simply not available. This, in turn, prevents researchers in less financially secure situations from sharing their failures.

Sindall and Barrington (2020) provide several examples from humanitarian aid projects which went wrong. Two relevant cases in South Africa are mentioned, where failures led to a waste of money and physical injuries. Publishing about failure could prevent this from happening again in the future, but only if the relevant people are able to access these publications. In countries with a lower quality of life, the relative difference that can be made to improving quality of life by preventing waste of money and physical injuries is larger.

It is not only at the submission phase, however, that the culture of success-ism rears its head. Peer reviewers may also be reluctant to contribute their expertise to results that do not reveal some great discovery about their area of study. On the one hand, this is reasonable because experts are busy and unsolicited requests for reviews are often perceived as a nuisance. On the other hand, reviewers’ increased willingness to respond to larger-scale positive-publishing journals over smaller-scale negative-publishing ones reflects again the equating of experimental success with prestige—a relationship that undergirds the philosophy of success-ism that journals like JOTE seek to overturn.

Thus, in addition to battling for financial stability,
publishers that seek to promote failure in science struggle to meet publication quotas or rally enthusiasm for publishing in their journal. The culture of success-ism in academia does double duty in this regard, draining failure-friendly journals’ wallets and potential contributors.

Lack of Training

Despite the challenges presented so far in this article, it is clear that not all academics or funders share the same unwavering commitment to success-ism (otherwise we wouldn’t be able to publish this editorial). Still, when journals do promote failure and trial and error become operational, they face the stark realization that neither researchers nor editors nor reviewers are properly equipped to analyze and interpret failure or trial and error.

Academic training in the sciences, whether it be in experimental design or statistical inference, focuses little on the positive interpretation of negative results. That is, scientists today are unequipped to constructively answer the question “what went wrong?” While more and more spaces are opening up to discuss failure in science, such as the WASH Failures Team created by Sindall and Barrington and others, there remains a lack of knowledge about how to critically evaluate failure in the sciences. This is reflected at all levels of the editorial process, from the submissions themselves, wherein reviewers often struggle to quantify or explain the scientific failure in question, to the peer reviews, wherein peer reviewers are used to critiquing a piece based on its positive contribution to the field rather than its negative influence on well-held ideas or techniques.

This lack of knowledge about how to critically evaluate failure extends beyond the sciences to many other types of work such as so-called high-risk industries (Bevilacqua & Ciarpapica, 2018). This lack of knowledge is not due to a lack of literature on the topic. In philosophy, the process of what we have called “trial and error” has been analyzed in a wide range of contexts. In fields like evolutionary epistemology, agnotology, and within the thought experiments of intuition pumps, much attention is paid to the role of trial and error and failure in the process of gaining knowledge. Additionally, a significant part of the field of Human Factors is dedicated to the analysis and prevention of failure—taking into consideration processes of trial and error. The problem doesn’t seem to be what knowledge is available, but rather what knowledge is known by the relevant actors and put into practice.

In one sense, the novelty of interpreting failures in academic practice leads to exciting new horizons where experts are pushed outside their comfort zone and must engage in live, critical, evaluation of data. In another sense, however, this new frontier beckons for structure—structure that can only be achieved through the long-term, symbiotic, practice of conducting and interpreting failed research and the normalization of failure in science.

In the meantime however, the onus of structuring failure falls on the editorial team of any given failure-friendly journal—a difficult and daunting task that may further discourage the widespread publication of scientific failure. Although training can help, it is insufficient on its own: just like with pilots, there is only so much we can cognitively expect from academics.

Destined to Fail?

Sindall and Barrington (2020) highlight the need for “more structured ways to report on failures” in science. We could not agree more. However, there are very real, structural, limitations from a publisher’s perspective that makes this goal very difficult to achieve. Thus, we argue that it is not only that “[j]ournals want to publish novel research and failure is too often not seen as novel” (italics added), but that oftentimes doing elsewise is self-destructive.

That being said, JOTE is in a fortunate position where we have (and continue to) overcome many of the challenges discussed in this editorial. We do so by leaning into our unorthodoxy, embracing failure wholeheartedly and interdisciplinarily, and tirelessly working to draw informative conclusions from inconclusive results. Nevertheless, the question must be asked: “In the current scientific climate, is this model sustainable?” To that, we must answer with a clear “no.” The evidence is clear that failure is required for a healthy science to flourish. The fact that failure has not even yet budded in the sciences highlights the incompatibility of the current model of scientific success and the promotion of scientific failure.

Rather than aim for a reformed system that accommodates failure, our goal is to help usher in a new system wherein failure is a necessary component in scientific progress and discourse. In other words, JOTE’s goal is to become obsolete. That is, one day, we hope that our efforts contribute to the formation of a scientific mainstream that recognizes the integral role that failure plays in the scientific process and actively normalizes its publication.

But how to achieve this goal? This piece, alongside Sindall and Barrington’s, has emphasized the many obstacles that publishers and researchers alike face in the publication and promotion of failure. While it may seem like there is no way out of success-first science, we offer some suggestions here for changes in academia and publishing. In the aviation industry, a nosedived airplane leads to multiple thorough investigations into the circumstances that lead to the problem. This is then followed up with an integral solution, taking into account different factors and their relation to each other. Likewise, in academia and publishing, multiple studies have analyzed the conditions which gave rise to the prob-
lem of positive publication bias. Now is the time to start with an integral solution, where instead of tackling positive publication bias from just one angle, researchers, editors, publishers, and managers all work together to solve the problem. In an attempt to facilitate this process, JOTE is increasing its collaborations - for example, by setting up a brilliant failure award at Utrecht University together with Utrecht Young Academy (and hopefully elsewhere), and by providing lunch lectures about failure and uncertainty at University Medical Centers, which we are starting to do with the help of The New Utrecht School. To reflect this shift from mostly publishing to a broader focus, we will launch the Center of Trial and Error—a place where all our activities come together. The long-term goal is to create an academic culture where failure is an accepted part of the scientific process. Therefore, while we are flattered by Sindall’s and Barrington’s (2020) hope that the Journal of Trial and Error won’t be the last “outlet for failure”; we certainly hope that we will be the last of our kind. We look forward to the day where founding a journal based only on failure is regarded as just as inane as founding one based solely on success.

Acknowledgments

The authors would like to thank Max Bautista Pernyà for his substantial comments on the first draft, especially concerning the systemic approach, which led to a complete overhaul and ultimately improvement of the article.

References


Levin, A. (2019). Boeing failed to predict that slew of 737 max warning alarms would confuse pilots, investigators say.


The Complexity of Joint Regeneration: How an Advanced Implant Could Fail by Its In Vivo Proven Bone Component

Paweena Diloksumpan1†, Florencia Abinzano2†, Mylène de Ruijter2, Anneloes Mensinga1,2, Saskia Plomp1, Ilyas Khan3, Harold Brommer1, Ineke Smit1, Miguel Dias Castilho2, P. René van Weeren1, Jos Malda1,2, Riccardo Levato1,2

Articular cartilage damage is a major challenge in healthcare due to the lack of long-term repair options. There are several promising regenerative implant-based approaches for the treatment, but the fixation of the implant remains a significant challenge. This study evaluated the potential for repair of an osteochondral implant produced through a novel combined bioprinting-based chondral-bone integration, with and without cells, in an equine model. Implants consisted of a melt electrowritten polycaprolactone (PCL) framework for the chondral compartment, which was firmly integrated with a bone anchor. The bone anchor was produced by extrusion-based printing of a low-temperature setting bioceramic material that had been proven to be effective for osteo-regeneration in an orthotopic, non-load-bearing and non-articular site in the same species in an earlier in vivo study. Articular cartilage-derived progenitor cells were seeded into the PCL framework and cultured for 28 days in vitro in the presence of bone morphogenetic protein-9 (BMP-9), resulting in the formation of abundant extracellular matrix rich in glycosaminoglycans (GAGs) and type II collagen. The constructs were implanted in the stifle joints of Shetland ponies with cell-free scaffolds as controls. Clinical signs were monitored, and progression of healing was observed non-invasively through radiographic examinations and quantitative gait analysis. Biochemical and histological analyses 6 months after implantation revealed minimal deposition of GAGs and type II collagen in the chondral compartment of the defect site for both types of implants. Quantitative micro-computed tomography showed collapse of the bone anchor with low volume of mineralized neo-bone formation in both groups. Histology confirmed that the PCL framework within the chondral compartment was still present. It was concluded that the collapse of the osteal anchor, resulting in loss of the mechanical support of the chondral compartment, strongly affected overall outcome, precluding evaluation of the influence of BMP-9 stimulated cells on in vivo cartilage regeneration.

Keywords cartilage, 3D (bio)printing, equine model, osteochondral, articular cartilage-derived progenitor cells

Local cartilage damage is a major challenge in human healthcare since it leads to an increased risk of developing early osteoarthritis (Kloppenburg & Berenbaum, 2020). Most of the available repair approaches are palliative with limited alleviation time, generating fibrous tissue with reduced mechanical strength (Kwon et al., 2019). There are no effective treatments that can fully restore the anatomical structure and function of focal cartilage defects. This unmet clinical need drives the ongoing quest for regenerative medicine and tissue-engineering approaches for articular cartilage repair (Malda et al., 2019).

Many new promising technologies (Johnstone et al., 2019; Patel et al., 2019) are currently being developed and tested with the aim of finding an implant that is effective in facilitating regeneration of cartilage. Given the difficulties associated with the fixation of chondral constructs in the joint (Gotterbarm et al., 2008; Mancini et al., 2017), an alternative approach is the use of composite osteochondral constructs composed of distinct osteal and chondral compartments that can be surgically press-fitted into suitably prepared defects, thereby avoiding the risk of dislodgement (Martin et al., 2007). However,
Take-home Message

The collapse of the osteal anchor of these osteochondral implants affected the integration of the construct with the native tissue and seriously compromised the mechanical stability. This precluded drawing firm conclusions about the potential for in vivo cartilage repair of the chondral compartment, consisting of bone morphogenetic protein-9 (BMP-9) stimulated progenitor cells. Collapse occurred because bone regeneration competent materials, which had shown success in a different anatomical location, were unable to correctly anchor the implant in the complex biomechanical environment of the joint. The imperfect dimensional match, originating from the intrinsic variability of the 3D printing process, had a much bigger influence than in previous studies. When combined with the brittle nature of the material, the implant was unable to withstand the complex loading of the knee joint.

The latter approach still faces many challenges, including design and optimization of the osteal compartment to act as an anchor for the overlying chondral compartment, production of a firm and durable connection between osteal and chondral compartment (Boushell et al., 2017), and optimization of the composition and structure of the chondral compartment (Diekman & Guilak, 2013; Lee et al., 2014).

To address those challenges, biomaterials that hold the potential for facilitating osteoregeneration within the osteal compartment were recently developed and investigated. First, 3D printed brushite-based scaffolds have been shown to be effective in promoting new bone growth after 6 months in an equine model that used the tuber coxae as implantation site (Vindas Bolaños et al., 2020). However, these materials are usually processed using aggressive acidic treatments, precluding the direct incorporation of cells and/or some types of polymer during the fabrication phase. Therefore, an apatite-based scaffold that could harden under physiological conditions into a calcium-deficient hydroxyapatite (CDHA) was developed. This material had been shown to be effective in a 7-month long in vivo study, upon implantation in a critical size defect in the tuber coxae of horses. That study was performed to compare two sophisticated architectures with constant and gradient pore size, respectively. The material was shown to facilitate excellent new bone formation, particularly when using the scaffold with constant pore size (Diloksumpan, Vindas Bolaños, et al., 2020). This showed the potential for using such material as an osteal anchor of a tissue-engineered osteochondral graft.

Another major challenge is the connection between the osteal and chondral compartments of the tissue-engineered osteochondral graft when using cell-friendly materials of strongly different mechanical characteristics. Recently, a technique for attaching the chondral compartment to the osteal compartment using melt electrowriting (MEW) was developed (Diloksumpan, de Ruijter, et al., 2020), in which MEW fibers of the chondral compartment were partially incorporated into the slowly setting apatite-based osteal compartment, thereby binding the two compartments together. This strategy allows both for optimizing the mechanical properties of the MEW-reinforced chondral compartment, and for integrating the chondral and osteal compartments.

Regarding the seeding of regeneration-competent cells within the chondral compartment, articular cartilage-derived progenitor cells have been recently identified and characterized in both humans and horses as a distinct cell population that has the potential for cartilage repair (McCarthy et al., 2012; Williams et al., 2010). This potential was further shown to be retained in combination with biomaterials (Frisbie et al., 2015; Levato et al., 2017), making this cell type a promising candidate for a comprehensive regenerative approach. Additionally, it was recently discovered that supplementation with BMP-9 during in vitro culture of ACPCs resulted in higher expression of gene-markers related with hyaline-like extracellular matrix production, compared to supplementation with transforming growth factor (TGF$\beta$), a more commonly used growth factor in cartilage tissue engineering (Morgan et al., 2020). This observation sparked interest in further investigation of the potential of BMP-9-stimulated ACPCs for cartilage repair in vivo.

The current study aimed at evaluating an osteochondral composite scaffold for cartilage repair. These constructs were composed of a combination of a previously proven osteogenic CDHA 3D printed scaffold for the osteal compartment, onto which a chondral compartment composed by MEW microfibrous meshes is tightly anchored. For the chondral compartment, an experimental group in which the MEW structure was seeded with ACPCs that had been stimulated for 28 days with BMP-9 before implantation (Abinzano et al., 2018), was compared with an implant featuring a non-filled, cell-free MEW cartilage scaffold as control. It was hypothesized that 1) the CDHA scaffold would show comparable performance in the horse when implanted in the subchondral bone as in the tuber coxae in terms of firmly anchoring in the surrounding tissue and inducing bone growth; 2) the novel interface would provide a lasting connection between the osteal and chondral compartments of osteochondral graft; and 3) the engineered chondral compartment of the osteochondral graft containing the stimulated ACPCs would outperform the cell-free structures in terms of in vivo cartilage matrix production (specifically, amount and density of type II collagen and GAGs, with resulting mechanical properties as close as pos-
Empirical

Materials and Methods

Experimental design

To assess the performance of integrated 3D printed osteochondral grafts that contained a cell-laden or a cell-free chondral compartment, the constructs were orthotopically implanted in a large animal model. Eight Shetland ponies (female, age 4 - 12 years, weight 149 - 217 kg (166 ± 29 kg)) were used and samples were implanted in the medial femoral ridge in the stifle joints. Healing was monitored for 6 months, after which the animals were humanely euthanized. The study was approved by the ethical and animal welfare body of the Utrecht University (Approval nr. AVD108002015307WP23).

Ponies were housed in individual boxes and fed a limited ration of concentrates together with hay for maintenance and free access to water. Quantitative gait analysis and radiographic examination of the stifle joints were performed before surgery for baseline values. Post-operatively, the animals were kept stable for 6 weeks with daily monitoring of vital signs, lameness checks at walk, and examination of the operated joints for swelling or other signs of inflammation. In week 5 and 6, they were hand-walked for 10 minutes twice daily and from week 7 on, they were kept at pasture. Quantitative gait analysis and radiographic exams were performed at 3 weeks, 3 months, and 6 months post-operatively. After 6 months, ponies were humanely euthanized for harvesting samples for both quantitative and qualitative analyses. The timeline of the experiment is represented in Figure 1.

Fabrication of construct

Microfiber meshes were produced from medical-grade PCL (Purasorb® PC 12 Corbion PURAC, The Netherlands) by using MEW technology as previously described (de Ruijter et al., 2019). The meshes were produced by horizontally patterning the microfiber (diameter = 10 μm) to form continuously uniform square spacing (300x300 μm) and vertically stacking the same pattern until reaching 1300 μm in total thickness. This structure was achieved by printing with a temperature of 90°C, a pressure of 1.25 bar, voltage of 10 kV, and collector velocity of 15 mm/s. Additionally, printing was performed at ambient temperature (22 - 24°C) with a humidity between 30 - 50%. Subsequently, PCL microfiber meshes were hydrolyzed by soaking them in sodium hydroxide (1M NaOH) for 15 minutes and washed in Milli Q water for 10 minutes 4 times. Finally, sterilization was carried out by immersion of the mesh in 70% ethanol for 15 minutes, followed by air-drying in a sterile cabinet until use.

Printable calcium phosphate (PCaP) paste was prepared as earlier described (Diloksumpan, de Ruijter, et al., 2020). In short, 2.2 g/ml of alpha-tricalcium phosphate (α-TCP, average particle size = 3.83 μm, Cambio Ceramics, Leiden, the Netherlands) and 0.13 g/ml of nano-hydroxyapatite (nano-HA, particle size = 200 nm, Ca₅(OH)(PO₄)₃, Sigma-Aldrich) were mixed with 40% w/v poloxamer solution (Pluronic® F-127, Sigma-Aldrich). α-TCP and nano-HA powder were disinfect with UV-light for 1 hour before mixing. The poloxamer solution was disinfected by filtration through a 0.22 μm sterile filter (Millex®-GS). This paste was loaded to a cartridge and kept at 4°C until use.

Osteochondral constructs were produced by combining the PCL microfiber mesh and the PCaP paste to form the reinforcement of the chondral compartment and the biomimetic bone compartment, respectively. Fabrication was performed by directly depositing the PCaP paste (approximated strand diameter = 250 μm) onto the hydrolyzed MEW mesh (Figure 2). Eighty percent of the mesh thickness was set as the initial height for depositing the first of non-macroporous PCaP layer, as this proved to
to their use for the experiment.

The constructs made of the combined CDHA and MEW meshes were disinfected in ethanol and exposed to UV-light for 1 hour as mentioned above. To avoid any pH changes that might affect the cells, the constructs were subsequently washed 3 times for 10 minutes with PBS and then immersed for 1 week in cell culture medium consisting of Dulbecco’s Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F-12, 11320033, Gibco, The Netherlands) supplemented with 10% fetal calf serum (FCS, Gibco, The Netherlands), 0.2 mM L-ascorbic acid 2-phosphate (Sigma), 1% MEM Non-Essential Amino Acids Solution (11140035, Gibco, The Netherlands) and 100 U mL⁻¹ penicillin with 100 μg mL⁻¹ streptomycin (Life Technologies, The Netherlands). Media were refreshed every 2-3 days.

On the day of seeding, medium was refreshed 2 hours before seeding and scaffolds were placed inside a custom-made polydimethylsiloxane (PDMS) ring (Figure 2) that prevented overflow of the cell suspension from the cartilage compartment to the bone scaffold. Ten million cells were suspended in 100 μl of medium and seeded on top of the constructs. The cell suspension was left to settle at the bottom of the cartilage part for 30 minutes. Afterwards, 2 ml of cartilage medium supplemented with 100 ng mL⁻¹ of BMP-9 (PeproTech, The Netherlands) was carefully added to the well. The seeded constructs were cultured for 4 weeks prior to implantation, refreshing the medium 3 times a week.

**Surgical procedure**

Ponies were premedicated with detomidine (intravenous (IV), 0.1 mg kg⁻¹) and morphine (IV, 0.1 mg kg⁻¹) and anesthesia was induced with midazolam (IV, 0.06 mg kg⁻¹) and ketamine (IV, 2.2 mg kg⁻¹). Anesthesia was maintained with isoflurane in oxygen together with continuous rate infusion of detomidine (IV, 0.075 mg kg⁻¹/h) and ketamine (IV, 0.5 mg kg⁻¹/h). Meloxicam (IV, 0.6 mg kg⁻¹), morphine (epidural injection, 0.1 – 0.2 mg kg⁻¹) and ampicillin (IV, 10 – 15 mg kg⁻¹) were administered preoperatively as analgesic medication and antibacterial preventative therapy, respectively.

The medial femoral ridge of the stifle joint was exposed by arthrotomy and an osteochondral lesion (diameter = 6 mm, depth = 6 mm) was surgically created using a power drill. The surgical area was flushed by saline for cooling and removal of debris. Cell-laden constructs were implanted press-fit in a randomly chosen hind limb, with the cell-free control being implanted in the contralateral limb. After closing the arthrotomy wound in four layers in routine fashion, procaine penicillin was administered (Procaped, intramuscular (IM), 20 mg kg⁻¹). Post-operatively, nonsteroidal anti-inflammatory medication (metacam, per os (PO), SID, 0.6 mg kg⁻¹) was administered for 5 days and opioids (tramadol, PO,
BID, 5mg \( \cdot \) kg\(^{-1} \)) were administered for 2 days.

Gait analysis
The ponies were trained on a treadmill prior to the study using a standard protocol for treadmill habituation. Twenty-eight spherical reflective markers with a diameter of 24\( mm \) (topline) and 19\( mm \) (elsewhere) were attached with double-sided tape and second glue to anatomical landmarks (Figure 3). Kinematic data were collected on a treadmill (Mustang, Fahrwangen, Switzerland) at trot using six infrared optical motion capture cameras (ProReflex, Qualisys, Gothenburg, Sweden) recording at a frame rate of 200\( Hz \) for 30 seconds at each session to obtain a sufficient number of strides.

To process the data, the reconstruction of three-dimensional coordinates of each marker was automatically calculated by Q-Track software (Qtrack, Qualisys, Gothenburg, Sweden). Each marker was identified and labelled using an automated model (AIM model) and manual tracking. Raw data of the designated markers were exported to Matlab (version 2018a, Niantics, California) for further analysis using custom written scripts. For each stride, two symmetry parameters were calculated using the vertical displacement of the head and pelvis (tubera sacrale) markers. For each stride, the differences between the two vertical displacement minima of the head (MinDiffhead) and pelvis (MinDiffpelvis) were calculated. Using the markers, limb-segments were formed and angles between these limb-segments were calculated. The difference between the maximal and minimal angle was defined as the range of motion (ROM) of a joint. For each timepoint, the mean value of all strides for each parameter was calculated.

Radiographic examination
Stifles were radiographed in 3 projections: lateromedial, cranialateral-caudomedial oblique and caudo-craniolateral projection using standard machine settings before surgery (baseline), at 3 weeks postoperatively and at 6 months, just before euthanasia.

Euthanasia and sample harvest
After 6 months, animals were euthanized by induction with Midazolam (IV, 0.06\( mg \cdot kg^{-1} \) body weight) with ketamine IV, (2.2\( mg \cdot kg^{-1} \) body weight) and subsequent administration of sodium pentobarbital (IV, 1400\( mg \cdot kg^{-1} \) body weight). Next, the stifle joint was exposed, and gross assessment of the medial trochlear ridge was performed, focusing on the degree of filling of the defect, the integration of repair tissue with the surrounding native tissue and the surface quality of the repair tissue. Subsequently, the entire osteochondral area containing the constructs was harvested for further analyses with the aid of a surgical bone saw. Harvested tissues were initially kept in sterilized PBS for micro-computed tomography (micro-CT) scanning, biomechanical analyses and for collecting tissue from the chondral compartment of the implant for biochemical analyses. After this, all tissues were fixed in 4% formaldehyde for subsequent histological processing.

Biomechanical evaluation
The compressive properties of the chondral compartment of the defect site, the adjacent surrounding native cartilage and the more distant surrounding native cartilage (5 - 10\( mm \)) from the boundary of the defect) (\( N = 7 \) for cell-laden constructs and \( N = 7 \) for cell-free constructs) were evaluated with a dynamic mechanical analyzer (DMA, DMA Q800, TA instrument) equipped with a custom-size compressing probe (diameter = 2\( mm \)). A ramp force of 0.250\( N \cdot min^{-1} \) was applied until reaching 2.0\( N \), to limit the deformation of sample to values below 200\( \mu m \). Compression modulus was calculated as the slope of the stress-strain curve in the range between 10 - 12\% strain.

Biochemical evaluation
Firstly, biochemical analyses were performed on supplemental pre-implantation constructs (\( N = 3 \)) that had been prepared in the same batch as the constructs that were later implanted. The chondral compartments of 28-day cultured constructs were removed and freeze-dried. Next, dry samples were digested in papain (Sigma Aldrich) at 60\°C overnight. DNA, sulphated glycosaminoglycan (sGAG), and alkaline phosphatase (ALP) content were quantified by performing the Quan-iT-PicoGreen-dsDNA-kit assay (Molecular Probes, Invitrogen, Carlsbad, USA), the dimethylmethylene blue assay (DMMB, Sigma-
Empirical

Diloksumpan et al.

Figure 4  Representative pictures of cell-laden and cell-free osteochondral constructs at the time of implantation. Cell-laden (A) and cell-free constructs (B) at the time of implantation. Positive safranin-O staining indicating the presence of glycosaminoglycans (pink = positive), positive type II collagen (brown = positive) and negative type I collagen (brown = positive) immunohistochemistry were observed in the chondral compartment of the cell-laden constructs before implantation (C).

Aldrich, The Netherlands) and the p-nitrophenyl phosphate assay (SIGMAFAST™, Sigma-Aldrich), respectively.

Secondly, tissue fractions that were collected from the chondral compartments of harvested implants (N = 6 for cell-laden constructs, N = 7 for cell-free constructs) were kept at -80°C, followed by lyophilization. Collagen content was quantified using an hydroxyproline assay (L-Hydroxyproline, Merck KGaA), and the sGAG and DNA quantification was performed as described above.

Microcomputed tomography

Microcomputed tomography was employed for the quantitative analysis of the bone compartments from the harvested osteochondral lesions (N = 7 for cell-laden constructs, N = 7 for cell-free constructs). Six freshly made osteochondral grafts were scanned in a micro-CT scanner (Quantum FX-Perkin Elmer) to quantify the initial volume of PCaP material, pre-operatively. The post-mortem harvested tissue containing the defect area and the surrounding native tissue were similarly scanned (voltage = 90kV, current = 200μA, voxel size = 30μm³ and total scanning time = 3 minutes). Subsequently, the 3D-reconstructed images were processed and analyzed using image J software (Schindelin et al., 2012) and Bone J plugin (Doube et al., 2010). Two-dimensional regions of interest (ROIs) were selected in an axial plane at the boundary between the defect and the surrounding native tissue and interpolated to form a three-dimensional volume of interest (VOI). Thresholding was performed to separately select areas of ceramics and newly formed bone respectively for further calculation. Then, the percentages of mineralized newly formed bone, of non-mineralized tissue and of remaining ceramics, including the percentage of ceramics volume loss, were quantified.

Histological evaluation

Firstly, supplemental pre-implantation constructs (N = 3) that had been prepared in the same batch as the ones that later were implanted were fixed in 4% formaldehyde. After decalcification in 0.5M Ethylenediaminetetraacetic acid (EDTA) disodium salt (pH = 8) for 1 day, tissues were dehydrated with graded ethanol series, cleared in xylene, and embedded in paraffin. Paraffin embedded tissues were sliced to 5μm sections. Histochemical evaluation of GAG was done by safranin-O / fast green staining. Type I collagen (primary antibody: monoclonal antibody EPR7785, 1.083 mg · ml¹, Abcam) and type II collagen (primary antibody: monoclonal antibody II-II6B3, 0.06 mg · ml¹, DSHB) were visualized by immunohistochemistry.

The tissues that were harvested after 6 months (N = 7 for cell-laden constructs, N = 7 for cell-free constructs) were kept in 4% formaldehyde and then decalcified in 0.5M EDTA disodium salt (pH = 8) for 24 weeks. Decalcified tissues were cut into two halves before processing to enable visual inspection of the center of the lesion. Tissues were dehydrated with graded ethanol series, cleared in xylene and finally embedded in paraffin. Paraffin embedded tissues were sliced to 5μm sections. For assessment of morphology and cell distribution, hematoxylin-eosin staining (Mayer's haematoxylin, Merck 109249 and eosin, Merck 115935) was performed. GAG and collagen alignment were assessed after safranin-O / fast green and picrosirius red staining, respectively. Types I collagen and type II collagen were visualized by immunohistochemistry, as described above. For immunohistochemistry, all samples were treated according to previously published protocols (Levato et al., 2017). Stained histological slides were imaged using a light microscope (Olympus BX51, Olympus Nederland B.V.), equipped with a digital camera (Olympus DP73, Olympus Nederland B.V.). To observe the picrosirius red stained slides, a polarizer was also mounted to the light microscope.

Statistical analysis

Normality of distribution of the data was assessed from skewness, kurtosis, and Q-Q plots. Results were reported as mean ± standard deviation. Wilcoxon signed rank tests were used to analyze
Figure 5  Representative pictures show white fragments of broken ceramic after press-fitting cell-laden and cell-free osteochondral constructs into the defect. Black arrows indicate the position of some visible bioceramic fragments. White arrows indicate protrusion of the chondral compartment

the biochemical, biomechanical, and micro-CT data. Statistical significance was set at \( p = 0.05 \). All tests were performed using Matlab (version R2018b, The MathWorks, Inc.).

To evaluate the gait parameters, stride-level data were analyzed with R software (version 3.6.0, R Core Team, 2019), using package NLME (version 3.1-137) for mixed modelling. Dependent variables were investigated for normality using normal probability plotting and examining for skewness and kurtosis. If not normally distributed, data were transformed to permit linear mixed modeling. The random effect was subject and timepoint was the fixed effect. Significance was set at \( p < 0.05 \) and p-values were corrected using the false discovery rate method. Residual plots were checked for heteroscedasticity versus the outcome, as well as for normality in Q-Q plots.

Results

In vitro

After 4 weeks of preculture, macroscopic characterization of tissue formation and hyaline-like extracellular matrix production were assessed both quantitatively and qualitatively within the chondral compartment of the cell-laden osteochondral constructs. The BMP-9 stimulated ACPCs meant to colonize the MEW scaffolds formed neo-tissue that had grown into a disc shape after 3 weeks of culture. During the 4th week of culture, outgrowth from the MEW meshes was observed (Figure 4A) from this construct. Cell-free constructs did not change after immersion in growth factor-free medium for 4 weeks (Figure 4B). Biochemical analyses of the chondral compartment of the cell-laden constructs were performed to quantify matrix production of stimulated cells toward chondrogenic lineage and osteogenic lineage, which revealed the presence of GAGs (GAGs-DNA\(^{-1}\) was \( 199.7 \pm 67.7 \mu g \cdot \mu g^{-1} \)) and ALP activity (ALP-DNA\(^{-1}\) was \( 3702 \pm 2111 U \cdot \mu g^{-1} \)), respectively. Safranin-O staining and type II collagen immunohistochemistry were also performed to visualize hyaline-like matrix production from stimulated cells, which revealed abundant deposition of GAGs and type II collagen within the constructs after 3 weeks of in vitro culture (Figure 4C), showing that the chondral compartments of the constructs (meant for subsequent implantation) were filled with a hyaline cartilage-like tissue. No preferential alignment of the collagen fibers could be observed.

Evaluation during surgical implantation

Both cell-laden and cell-free constructs were press-fit implanted into the surgically created defect sites. During this procedure, the slightly irregular outer edge of the osteal part of the construct hampered easy sliding of the construct down into the defect and some fragmentations of the edges of the bioceramic scaffold was observed during the procedure. This was similar for the cell-laden and cell-free constructs, which had identical osteal parts (Figure 5). Further, of some cell-laden constructs, the surface of the chondral compartment was not level over the entire circumference with the surrounding native cartilage after press-fitting into the defect site.

Post-operative clinical monitoring

After surgical implantation, the animals were checked clinically for physical appearance and vital signs on a daily basis. All ponies recovered well from anesthesia after surgery and passed uneventfully through the rehabilitation period without any abnormalities in body temperature or behavior, with good weight-bearing on all operated limbs and no clinical signs of lameness during the entire period, with the exception of a single pony that developed severe lameness at 10 weeks after surgery. This pony was treated with anti-inflammatory medication and examined radiographically, which revealed extensive osteolysis around the created lesion. Because of persistent discomfort, the pony was euthanized at 12 weeks after surgery. Therefore, it was excluded from all analyses.

Gait analysis

Objective gait analysis was used to check for lameness or other signs of dysfunction of the musculoskeletal system. Objective data retrieved before implantation and at the end of the experiment were assessed for relevant parameters, including symme-
Gait analysis: Symmetry parameters. Symmetry data of the head (A) and pelvis (B) show no consistent differences over time. However, pelvis pitch decreased consistently in all individuals (C).

Symmetry parameters
Front and hind limb lameness were analyzed through evaluation of the symmetry parameters of the head (MinDiff Head (Figure 6A)) and of the pelvis (MinDiff Pelvis (Figure 6B)). These values reflect the differences in minimal vertical displacement with a negative MinDiff indicating a left-sided asymmetry and a positive MinDiff a right-sided asymmetry. In the treated ponies (except for the case referred to above that was euthanized), for both the head and the pelvis, there was no clear pattern in the direction of the asymmetries between baseline and endpoint and those differences between baseline and endpoint were minimal and statistically not significant. Therefore, symmetry measures could not discriminate between cell-laden and cell-free constructs. Further, there was also no clear effect of timepoint on pelvis roll and pelvis yaw range of motion (Supplementary Figure 12), however, pelvis pitch range of motion (ROM) (Figure 6C) decreased for all subjects with almost 20% over time (Supplementary Table 1).

Limb parameters, effects of time
There was a significant effect of time for the height the toe was lifted from the surface during the swing phase of the limb that decreased significantly in the cell-free treated limbs, but not in the limbs treated with cell-laden constructs (Supplementary Table 1). The only other significant effect of time was a decrease in the extension of the metacarpophalangeal joint of the forelimb ipsilateral to the hind limb that had been treated with cell-laden constructs, indicating unloading of that forelimb (Supplementary Table 1).

Limb parameters, differences between cell-laden and cell-free at endpoint
There were no significant differences between any of the cell-laden and cell-free limb parameters at the end of the experiment. Results from the linear mixed model are shown in Supplementary Table 2.

Radiographic examination
Healing progression within the osteal compartment of the implanted osteochondral constructs was followed up non-invasively through radiographic examination. On the radiographs taken at baseline, 3, and 6 months, no obvious abnormalities in term of the architecture of the surrounding native tissue were detected, other than the defects that had been created. This was with the exception of the pony that developed severe lameness. In that animal, severe osteolysis was noted at the implantation site 3 months after the implantation (Supplementary Figure 13).

Post-mortem macroscopic evaluation of the repair tissue
Macroscopic characteristics, for instance, color, appearance, and filling level of the lesion, were ob-
Biochemical analyses of repair tissue within the chondral compartment

The deposition of GAGs and collagen, the two main elements that compose cartilage extracellular matrix, were quantified within the chondral compartment of the osteochondral graft 6 months after implantation. There were no significant differences in either GAGs (cell-laden: 30.46 ± 15.95 μg ⋅ g⁻¹, cell-free: 24.44 ± 15.31 μg ⋅ g⁻¹) or collagen expressed per DNA (cell-laden: 79.66 ± 91.21 μg ⋅ μg⁻¹, cell-free: 134.21 ± 153.73 μg ⋅ μg⁻¹) between the chondral compartments of the cell-laden and cell-free constructs (Figure 8A, 8B and Supplementary Figure 14). However, all values were substantially lower than those from native cartilage (Figure 8, grey dotted line) that was harvested distantly from the defect site.

Biomechanical properties of the repair tissue within the chondral compartment

Compressive strength of the chondral compartment was assessed and compared in three different locations: the defect site, adjacent surrounding native tissue, and distant surrounding native tissue (Figure 9A), the latter two as control measurements from healthy cartilage tissue. There were no significant differences in the Young’s modulus of the chondral compartment between cell-laden (0.31 ± 0.13 MPa) and cell-free (0.42 ± 0.19 MPa) constructs (Figure 9B). This was also true for two sites of the native cartilage, one close to the border of the defect (cell-laden: 1.75 ± 0.80 MPa, cell-free: 2.22 ± 0.48 MPa) and one at 5 − 10 mm from the defect boundary (cell-laden: 1.86 ± 0.78 MPa, cell-free: 2.19 ± 0.77 MPa) (Figure 9C, 9D). However, the compression modulus of the native tissue was substantially higher (approximately 5 − 6-fold) than inside the chondral compartment of the implant.

Micro-CT evaluation of repair tissue within the osteal compartment

Bone healing and integration after was assessed through micro-CT scanning 6 months after implantation. Micro-CT images showed significant bone loss surrounding the implant in both the cell-laden and the cell-free groups, which could be visualized as black areas between the porous bioceramic structure (white) and the surrounding native bone (grey). However, mineralized bone formation could be visualized in some scaffolds from both groups with an integration to neighboring native bone (Figure 10A). Statistically, there were no significant differences in mineralized bone formation (cell-laden: 6.14% ± 10.09%, cell-free: 4.73% ± 4.93%) and non-mineralized tissue (cell-laden: 81.38% ± 15.37%, cell-free: 74.71% ± 12.44%). However, there was a significant difference in the amount of remaining ceramics between the two groups (cell-laden: 12.48% ± 9.75%, cell-free: 20.56% ± 10.54% (p = 0.0313)) (Figure 10B). In line with this, there was a difference in the degra-
Figure 10  Representative micro-CT images from the middle of the sagittal plane of the constructs and quantification from 3D-reconstruction of micro-CT. Representative micro-CT images from the middle of the sagittal plane of the constructs (white = ceramics, grey = mineralized tissue, black = non-mineralized tissue) (A). Quantitative analysis from micro-CT reconstruction showing percentage of mineralized bone formation, non-mineralized tissue, and remaining ceramics (B). The volume loss of ceramics was slightly higher in the cell-laden constructs compared to the cell-free ones (C).

Histological evaluation of the osteochondral repair tissue

Histological slides were assessed to identify the composition of the repair tissue matrix deposited within the defect site. In the chondral compartment, the defect sites of both cell-laden and cell-free structures were filled with fibrous repair tissue with degenerated and necrotic superficial surface with minimal inflammatory reaction, as revealed by H&E and safranin-O staining (Figure 11, Supplementary Figure ??, Supplementary Figure 17, and Supplementary Figure 18). Integration at the boundary of the defect between chondral repair tissue and surrounding native cartilage was observed in both groups. The production of GAGs, type II collagen, and type I collagen was very limited in the repair tissue in both groups (Figure 11). The organization of the collagen fibrils in both groups seemed random, without any hierarchical pattern that could be identified by polarized light imaging of picrosirius red staining. Additionally, the special distribution of PCL-microfibers, which had disappeared because of the xylene treatment during sample preparation, was still traceable within the chondral compartment of both groups (1 out of 7 for cell-laden and 5 out of 7 for cell-free structure).

In the bone compartment, there was positive staining for type I collagen in some scaffolds from both groups at places where there were islands of new mineralized bone formation. There were multifocal coalescing spots of inflammatory reaction characterized by macrophages, multinucleated giant cells, lymphocytes, eosinophils, and plasma cells (Supplementary Figure 16, Supplementary Figure 17, Supplementary Figure 18).

Discussion

This study aimed to evaluate the efficacy of an engineered osteochondral composite scaffold that was fabricated by combining a proven osteogenic CDHA scaffold for the osteal compartment with a novel interface for the connection between the chondral and osteal compartments. For the chondral compartment, BMP-9 stimulated cell-laden and cell-free constructs were compared. The cell-laden constructs contained in vitro formed tissue that was rich in GAGs and type II collagen, obtained by seeding Articular Cartilage Progenitor Cells (ACPCs) and stimulating them with BMP-9 for 4 weeks prior to
implantation. After implantation in an equine osteochondral defect for 6 months, there was poor chondral repair tissue in both the cell-laden and cell-free implants. The repair tissue was akin to fibrocartilage and was characterized by the presence of fibrous tissue with low content of GAGs and type II collagen and a degenerated surface. The CDHA scaffold had failed to act as an osteal anchor, as evidenced by the radiographical images showing misalignment and partial collapse of the CDHA construct, the presence of CDHA fragments within the defect and in the surrounding tissues, and a limited volume of newly formed calcified bone in the pores of the ostal anchor.

In the quest for a method to achieve satisfactory and durable repair of articular cartilage, several osteochondral grafts that incorporate cells and that were manipulated to optimize biochemical and biomechanical properties, have been investigated in the past decades (Huang et al., 2016). Articular Cartilage Progenitor Cells have become a promising cell source due to their ability to retain their chondrogenicity after their expansion for several passages (Williams et al., 2010). Recently, growth factor BMP-9 was shown to be a potent stimulator of chondrogenic differentiation of this cell type in vitro (Morgan et al., 2020). This warranted further investigations to evaluate the use of BMP-9 stimulated ACPCs for cartilage repair in vivo. Indeed, the cell-laden chondral compartment showed a high presence of neo-cartilage extracellular matrix production after pre-culture at the time of implantation, yet the average GAG content decreased approximately 6.5-fold during the in vivo implantation period. The GAG content of cell-laden constructs in fact decreased to the level of the cell-free constructs, suggesting loss or disintegration of the in vitro formed tissue. Which factor initiated this loss of in vitro formed tissue remains unclear. A previous study from McCarthy et al. (2012) demonstrated superior results in using ACPCs for cartilage repair in an equine model, in comparison with mesenchymal stem cells. However, due to the use of different materials and cell culture protocols, it is impossible to directly compare those results with the ones from the current study. Several factors might have been involved in the deterioration of the chondral compartment in this study, most prominently mechanical stresses due to the partial failure of the ostal basis and the resulting poor osteointegration (Heuierjans et al., 2018).

The nature of the ostal anchor is an important factor when developing tissue-engineered osteochondral implants. Much work has been done on the development of several types of bone grafts and many of these are routinely used in clinical settings (Oryan et al., 2014), so of the various elements of an osteochondral implant, the bone part is seemingly the least difficult one. However, the relationship between the ostal anchor and the quantity and quality of the repair tissue in the chondral compartment has been the subject of debate (Bal et al., 2010) and it is still unclear what ostal anchor would form the best base for facilitating cartilage repair. The exact same bioceramic material tested in this long-term, orthotopic equine study, had previously been shown to successfully guide osteoregeneration in the same species, when implanted in the tuber coxae, an anatomical locus less subject to intense mechanical loads (Diloksumpan, de Ruijter, et al., 2020). Additionally, this previous study also focused on comparing different pore architec-

Figure 11  Representative histological images from the center part of cell-laden and cell-free structures after implantation for 6 months. Safranin-O/fast green (red color = positive) (A, E); Collagen type II (brown color = positive) (B, F); Picrosirius-red (C, G); collagen type I (brown color = positive) (D, H) of cell-laden (A-D) and cell-free structures (Scale bar = 1 mm).
The current study aimed at evaluating a cell-laden and a cell-free version of an osteochondral composite scaffold for cartilage repair that consisted of an in vivo proven osteogenic bone scaffold for the osteal compartment and made use of a novel interface for the connection of the chondral and osteal compartments.
was to be tested in the study, is not possible. Also, no conclusion could be reached about the interface between the osteal and chondral compartments that was used since delocalized MEW-mesh structures were observed in some scaffolds from both groups. This might be due to misalignments of the osteal compartment as discussed above, to shear forces during loading, or a combination of both.

During the in vivo post-operative monitoring of the animals, the clinical signs were very mild and far from alarming, except for the single pony that developed severe lameness. Clinical examinations were performed routinely by experienced veterinary specialists, however, assessment of locomotion through visual observation alone is subjective and known to have poor repeatability, especially in mild cases. This is partly due to the inability of human visual perception to properly distinguish, notice, and quantify differences in locomotion at high resolution (Serra Bragança et al., 2018). Therefore, quantitative gait analysis was employed as an objective and non-invasive assessment. The gait analysis data did not show many differences with respect to baseline. This may to a certain extent have been related to methodological factors. During the assessment, ponies were put on a treadmill and they were imposed the same belt velocity during both measurements. Therefore, the subjects were forced to trot at the same velocity, ensuring that stride length needed to be maintained. This might be the reason why there were no differences between timepoints for maximal protraction and retraction (the limb parameters). However, pelvis pitch range of motion (ROM) decreased for all subjects with almost 20% over time. This pattern is often seen in case of dysfunction of the back. The finding may thus be related to earlier observations that bilateral hindlimb lameness may induce back problems in horses (Alvarez et al., 2008; Alvarez et al., 2007; Greve et al., 2017). Toe dragging of the lame limb, in which the hoof is lifted less high off the ground, is another sign of pain (Buchner et al., 1995). Nevertheless, the overall impact of the bilateral lesions in the stifles joints was low, as evidenced by the fact that there was no sign of load redistribution from the hind to the front limbs. If that had been the case, the subjects would have compensated by displacing their center of mass more to the front, resulting in more negative angles for forelimb fetlock extension, as fetlock hyper extension correlates with peak ground reaction force (GRFPeak) (Crevier-Denoix et al., 2010), where less negative angles indicate a lowered GRFpeak. In fact, only the fetlock angles of the forelimb ipsilateral to cell-laden construct changed, becoming less negative, hence indicating unloading rather than additional loading (lower GRFpeak). The reason for this is not clear.

It can be concluded that even seemingly minor modifications of a successful implant may have grave consequences and extrapolation is dangerous in the complex in vivo situation. In this case, the failure of the osteal compartment of the construct, the use of which seemed well-backed by solid in vivo data, did not permit drawing conclusions about the original hypotheses. Given the relatively frequently occurring, rather disappointing results of in vivo orthotopic testing of promising techniques for joint repair, it may be wise to put more emphasis on performing pilot experiments before embarking on a full-scale in vivo study in a large animal experiment (Vindas Bolaños et al., 2017). Functional joint repair remains a huge challenge that has not been addressed to some satisfying extent during the last decades, despite many promising approaches. It is likely that the quest for a real solution will go on for some time by trial and error with more errors to come. Those errors are inevitable and need to be made but they should take the least possible toll on experimental animals.

Conclusion

This study presented the results from the evaluation of a cell-laden and cell-free versions of an osteochondral implant for cartilage repair in a challenging in vivo large animal model. The osteal anchor of this osteochondral implant, composed of a bioceramic material that had previously been proven to facilitate mineralized new bone formation in the same species, failed to perform as an effective fixation with sufficient stabilization for both cell-laden and cell-free osteochondral implants. This insufficient fixation was evidenced by the extensive osteolysis, the collapse and misalignment of the osteal anchor, and the limited volume of newly formed bone. The failure of the bone anchor hindered the evaluation of the two versions of the chondral compartment for cartilage repair. The study shows that, even after an equivalent ceramic bone component had shown very satisfactory results in the same species, minor differences in the implant and a change in testing condition proved to be enough to lead to completely different results, in this case precluding drawing conclusions about the effect of the principal variable. This outcome stresses the need of carrying out in vivo pilot studies under exactly the same conditions before moving into a larger in vivo study.
References


Greve, L., Dyson, S., & Pfau, T. (2017). Alterations in thoracolumbosacral movement when pain causing lameness has been improved by diagnostic analgesia. The Veterinary Journal, 224, 55–63. https://doi.org/10/gbr4


View interactive version here


Acknowledgments

The authors would like to thank Dr. J.P.A.M. (Thijs) van Loon and Dr. Janny de Grauw for assistance with anesthesia, Dr. S. Veraa for assistance with the radiographic examination, and Paulina Nunez Bernal for assistance with the biomechanical characterization. The authors also wish to acknowledge the funding support from Royal Thai government scholarship (Thailand, PD), the Dutch Arthritis Foundation (LLP-12 and LLP-22), the AO Foundation (Collaborative Research Center AO Foundation, 3D OC Constructs, FA), the European Research Council under grant agreement N0. 647426 (3D-JOINT, JM), and the UK Regenerative Medicine Platform (UKRMP, MR/L02280X/1). The primary antibodies against collagen type II (II-I6683) developed by T. F. Linsenmayer and E. S. Engvall, respectively, were obtained from the DSHB developed under the auspices of the NICHD and maintained by the University of Iowa, Department of Biology, Iowa City, IA, USA.
Supplementary information

Table 1  Symmetry parameters (differences between baseline (before implantation) and endpoint (6 months after implantation) of the study for all ponies). Values are given in estimated means (CI), significant results are bold.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Baseline</th>
<th>Endpoint</th>
<th>% Difference</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MinDiff Head</td>
<td>-1.00 (-5.08 - 3.08)</td>
<td>0.33 (-3.75 - 4.41)</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>Pelvis roll ROM</td>
<td>4.81 (3.67 - 5.94)</td>
<td>5.21 (4.08 - 6.35)</td>
<td>8.50</td>
<td>0.24</td>
</tr>
<tr>
<td>Pelvis pitch ROM</td>
<td>6.89 (5.38 - 8.39)</td>
<td>5.57 (4.06 - 7.07)</td>
<td>-19.16</td>
<td>0.00</td>
</tr>
<tr>
<td>Pelvis yaw ROM</td>
<td>3.52 (2.29 - 4.75)</td>
<td>3.70 (2.46 - 4.93)</td>
<td>5.07</td>
<td>0.56</td>
</tr>
<tr>
<td>MinDiff Pelvis</td>
<td>-1.36 (-4.61 - 1.89)</td>
<td>-2.52 (-5.77 - 0.73)</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>MaxDiff Pelvis</td>
<td>0.98 (-2.08 - 4.04)</td>
<td>1.68 (-1.39 - -4.74)</td>
<td>0.54</td>
<td></td>
</tr>
<tr>
<td>Fetlock hyperextension (cell-laden) Ipsilateral Front</td>
<td>-36.11 (-37.96 - -34.25)</td>
<td>-33.58 (-35.43 - -31.72)</td>
<td>-7.01</td>
<td>0.00</td>
</tr>
<tr>
<td>Fetlock hyperextension (cell-free) Ipsilateral Front</td>
<td>-36.41 (-39.02 - -33.80)</td>
<td>-34.61 (-37.22 - -32.00)</td>
<td>-4.93</td>
<td>0.19</td>
</tr>
<tr>
<td>Fetlock hyperextension (cell-laden)</td>
<td>-40.00 (-42.59 - -37.41)</td>
<td>-39.49 (-42.08 - -36.90)</td>
<td>-1.27</td>
<td>0.37</td>
</tr>
<tr>
<td>Fetlock hyperextension (cell-free)</td>
<td>-41.08 (-44.30 - -37.86)</td>
<td>-40.72 (-43.94 - -37.50)</td>
<td>-0.87</td>
<td>0.71</td>
</tr>
<tr>
<td>Limb Height (cell-laden)</td>
<td>88.55 (72.90 - 104.19)</td>
<td>81.70 (66.06 - 97.35)</td>
<td>-7.73</td>
<td>0.06</td>
</tr>
<tr>
<td>Limb Height (cell-free)</td>
<td>91.39 (73.26 - 109.51)</td>
<td>84.64 (66.51 - 102.76)</td>
<td>-7.39</td>
<td>0.01</td>
</tr>
<tr>
<td>Max Protraction (cell-laden)</td>
<td>19.16 (17.22 - 21.11)</td>
<td>19.65 (17.71 - 21.59)</td>
<td>2.55</td>
<td>0.34</td>
</tr>
<tr>
<td>Max Protraction (cell-free)</td>
<td>19.59 (16.93 - 22.26)</td>
<td>20.60 (17.94 - 23.27)</td>
<td>5.16</td>
<td>0.18</td>
</tr>
<tr>
<td>Max Retraction (cell-laden)</td>
<td>18.78 (16.99 - 20.58)</td>
<td>18.76 (16.97 - 20.56)</td>
<td>-0.11</td>
<td>0.97</td>
</tr>
<tr>
<td>Max Retraction (cell-free)</td>
<td>18.45 (15.48 - 21.42)</td>
<td>18.19 (15.22 - 21.16)</td>
<td>-1.39</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Table 2  Hind limb parameters (differences between cell-free and cell-laden constructs at 6 months after implantation). Values are given in estimated means (CI)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cell-free</th>
<th>Cell-laden</th>
<th>% Difference</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fetlock hyperextension (ROM)</td>
<td>-40.7 (-43.5 - -37.9)</td>
<td>-39.5 (-42.2 - -36.7)</td>
<td>3.12</td>
<td>0.19098</td>
</tr>
<tr>
<td>Maximal protraction (ROM)</td>
<td>20.6 (17.8 - 23.3)</td>
<td>19.7 (16.9 - 22.4)</td>
<td>4.85</td>
<td>0.39042</td>
</tr>
<tr>
<td>Maximal retraction (ROM)</td>
<td>18.2 (15.5 - 20.9)</td>
<td>18.8 (16.1 - 21.5)</td>
<td>-3.04</td>
<td>0.57655</td>
</tr>
<tr>
<td>Limb height (mm)</td>
<td>84.6 (67.4 - 101.9)</td>
<td>81.7 (64.4 - 98.9)</td>
<td>3.59</td>
<td>0.18918</td>
</tr>
</tbody>
</table>
Figure 12  Gait analysis: symmetry parameters. Pelvis roll and yaw joint angles (A, B) showed no significant differences between baseline and 6 months after implantation and neither did the kinematic hind limb parameters (C, D, E) between baseline and 6 months after induction. Limb height of the hind limbs (F) decreased for both hindlimbs, but only significantly for the cell free group, though there was no difference in limb height between cell-laden and cell-free groups at 6 months after implementation.

Figure 13  Representative radiographic images (lateral-medial and cranio-lateral-caudo-medial oblique projections) of the stifle of ponies before implantation, 3 months after implantation and 6 months after implantation. Red arrows indicate the implantation sites. No radiographic abnormalities were noted in any of the ponies (1st row), except for the pony that became severely lame at 10 weeks. In this animal extensive osteolysis was observed (2nd row).
Figure 14  Amount of GAG that was normalized with amount of DNA. Amount of GAG/DNA as absolute value (A) and as a ratio (B) in the individual animals. Amount of collagen/DNA as absolute value (C) and as a ratio (D) in the individual animals.

Figure 15  Representative Hematoxylin-Eosin (H&E) staining of 6-month harvested samples showing a degenerated and necrotic superficial layer at the surface of the chondral compartment featuring inflammatory cells. Black arrow indicates area of degenerated cells, P = plasma cells.
Figure 16  Representative immunohistochemistry of collagen type I staining of 6-month harvested samples from cell-laden and cell-free osteochondral structures.

Figure 17  Representative Hematoxylin-Eosin (H&E) staining of 6-month harvested samples from cell-laden osteochondral structures. CR = Remaining ceramic, NB = Newly formed bone, MEW = pattern of PCL-microfibers, * = Multifocal foci of inflammatory reaction, # = artifact from cutting, EO = Eosinophil, L = Lymphocyte.

Figure 18  Representative Hematoxylin-Eosin (H&E) staining of 6-month harvested samples from cell-free osteochondral structures. CR = Remaining ceramic, NB = Newly formed bone, MEW = pattern of PCL-microfibers, * = Multifocal foci of inflammatory reaction, # = artifact from cutting, EO = Eosinophil, L = Lymphocyte.
Ponies, joints, complexity, and the method of difference

Hubertus Nederbragt

An important topic in biomedical science is the repair of damaged organs by *in vitro* cultured differentiated stem cells. This article evaluates an article in this field, entitled “The complexity of joint regeneration”, by Diloksumpan et al. (2021), who described a regeneration experiment of artificial damage of the joint of ponies. The experiment failed and I describe the possible cause of this failure by discussing the design of the experiment in the light of J.S. Mill's Method of Difference, published in 1843. I continue with a discussion of the concept of complexity that was introduced by the authors of the paper, by pointing out that three types of complexity may be distinguished; one of these is complicatedness, which characterizes the assumed complexity of the joint experiment. I propose that this complicatedness can be solved by the use of the method of difference.

**Keywords** regeneration of joints, pilot experiment, method of difference, complexity, reductionism

**T**issue repair and regeneration is one of the hot topics in biomedical research. The most recent branch of this field is the repair of tissues and organs through the use of stem cells differentiated *in vitro*. These cells go on to become the source of new tissues after they have been implanted in the living body at the site of the defect. In the implant the cells may be combined with non-biological materials that can serve as a matrix on or in which the new tissue may take the necessary form.

The paper by Diloksumpan et al. (2021) in this issue of JOTE is a good example of this novel approach. The authors are involved in a study of the repair of damage of cartilage in bones and joints, using 3D-printed constructs; these serve as a scaffold on top of which new cartilage that has been grown in a cell culture system may be implanted inside the damaged site, leading to newly-formed cartilage growing from the implant.

**Experiment**

In their experiment Diloksumpan et al. (2021) artificially damaged cartilage by drilling a hole in the joints of ponies, filled the gap with the construct that bears the cultured cartilage, and watched the outcome. To evaluate the results an admirable battery of well-chosen methods was applied to monitor the changes in biomechanical properties, biochemical composition (analysis of collagen and glycosaminoglycans), and morphology of the joints (radiography, microcomputed tomography, histology, and immunohistochemistry), and their function was established by following the walking behavior (gait analysis) of the ponies.

However, in the end, the experiment failed. The main reason, according to the authors, was the failure of the surgical procedure: the construct collapsed because the materials could not be correctly implanted in the joint, whereas they, according to the authors, had been implanted successfully before in another test. The conclusion of the authors was that they should have performed a pilot experiment because, as they state in the “Take Home Message” paragraph, the expected success of the experiment was based on a previous success with experiments with the same materials but in another anatomical location. The cited literature of the authors suggests that they earlier implanted the construct in the tuber coxae of horses, whereas the experiment that failed was performed with implantation in the joint of ponies. I heartily agree with their conclusion of the missing pilot experiment. But what pilot experiment? A pilot experiment is an experiment that may be done to test an idea, to see if it is worthwhile to give that idea further attention. When the result of the pilot is negative, the idea is discarded and will not be tested again; when the result is positive, the idea will be further tested and developed.

**Companion Article**

Diloksumpan et al. (2021)

The Complexity of Joint Regeneration: How an Advanced Implant could Fail by Its *In Vivo* Proven Bone Component

DOI: 10.36850/e3
this may lead to a more definite experiment. The advantage of a pilot experiment is that it can be done on a small scale and takes less time, energy, material (and experimental animals), and money. The results of a pilot experiment are not meant to be published. But even though a pilot experiment may be just a preliminary and small trial it should, in its design, fulfill the criteria of the ‘method of difference’, otherwise it is not a pilot experiment but a useless experiment. To this method of difference I will now turn my attention.

Method of Difference

In 1843 John Stuart Mill (1806-1873) published a book, A system of logic. Ratiocinative and inductive, having as a subtitle Being a connected view of the principles of evidence and the methods of scientific investigation.

One of these principles is the Method of Difference that can be found in the Second Canon of the Four Methods of Experimental Inquiry of Mills, described in Chapter VIII, Book III: “Of Induction” (Mill, 1843/1884):

If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ is the effect, or the cause, or an indispensable part of the cause, of the phenomenon. (p. 256)

The text may be a little bit obscure for the modern reader, but it may help to consider the name Mill gave to this canon: The Method of Difference. When you study the effect or the possible cause of a factor in an experiment (be it in a laboratory, or in the field, or in epidemiology), one must make sure that the factor under investigation is the only variable, that there is only one factor that is different and that all the others are similar. After almost two centuries of scientific practice at universities and research institutes we may expect that these principles are part of the thinking and doing of the members of the scientific enterprise, that they belong to a set of reflexes of those working in the clinics, laboratories, and workplaces. But sometimes they are not.

Let us look at the experiment that was set up to test the efficacy of a construct for cartilage repair. The construct contained three parts: (1) a microfiber mesh with (2) a calcium paste that together mimicked the two layers of bone, (3) combined with in vitro cultured cartilage on top of the calcium layer. Since the cultured cartilage in the construct was supposed to do the job – the other two parts should fix the cartilage into the bone of the joint – a similar construct without cartilage was implanted in the joint at the opposite side of the animal. But it did not work according to the expectations.

An analysis of the earlier findings by the authors, presented in their introduction paragraph, may help us to find what kind of pilot experiment they should have performed given the method of difference discussed above. The following variables (or, in other words, the possible differences) taken from that introduction, should be considered:

- the construct, with two parts, each of them being a separate variable,
- the location of the implant, the variables being tuber coxae (hip bone) or joint,
- the animal: horse or pony,
- the surgical procedure (a variable as mentioned by the authors).

When one wants to test the effect of a certain factor, call it a variable or a difference maker, one should repeat the experiment that has been performed before, of which one knows the outcome, change only one variable of interest and look at the effect. Only then is knowledge obtained of what the variable or the factor of concern does. A second variable cannot be introduced simultaneously because then it becomes impossible to determine which variable is responsible for the resulting effect. This is the sort of experimental error the method of difference wants to prevent. Does this error occur in the joint repair experiment? I think it does.

In a paper from the Veterinary Faculty of Utrecht University, published almost a generation ago, it was shown that, with regard to wound healing of the skin, there is a difference in the extent and speed of healing between ponies and horses and between anatomical locations of the wound (Wilming et al., 1999). It is therefore important to realize that the behavior of the construct at the implantation site cannot be determined when the implantation in the hip bone of a horse is compared to the effect of implantation in the stifle joint of a pony. Second, it may be asked whether the surgical skills and procedure for implantation of the scaffold construct in bone are adequate for implantation in a joint (here a possible pilot experiment suggests itself). Third, it cannot be deduced from the introductory paragraph which of the elements of the construct, separate and in combination, have been tested to find out whether they work in the bones and joints of horses and in the bone and joints of ponies. One has to consult the cited literature to find that out. The tests of all these differences may have been important in designing such an experiment as ambitious as described in the paper.

The routine of applying the method of difference, of being intuitively aware of its importance, should be imprinted on students as part of their academic education. I do not try to blame anyone, and I am speculating here, but my feeling is that the implantation experiment might have benefited from the involvement of an experienced senior scientist in the design of the project, especially because the design may have been driven too much by the expectation...
of success.

Complexity

Next, I want to discuss the assumed complexity of the experimental system. The word complexity has a place in the title of the paper, suggesting its importance for the authors, but in the paper itself it seems to be of minor relevance only. Nevertheless, it is considered a phenomenon that has a central place in everyday life.

First, there is the use of the word complexity that occurs in everyday parlance (and possibly also in the paper we are talking about) and is an expression of the embarrassment of the speaker when confronted with a situation she cannot control or a problem that she cannot solve. In the words of Chambers (2015, p. 748): ‘[A] thing which is complex is just one that defies understanding and/or simple description at first inspection.’

Then there is the reductionist view of complexity. Phenomena, things, and organisms are considered complex whenever they are composed of a multitude of different components. In principle it should be possible to take the components apart and study them one by one. The system is essentially knowable, and its “behavior” is theoretically predictable. It may be referred to as complicated rather than complex (Bawden, 2007). I shall refer to this type of complexity as “complicatedness” (rather than complication).

Finally, there is the third and more holistic notion of complexity; this has to do with the contingent nature of the interrelationships of the components. These interrelationships enable the system to adapt and they often produce phenomena that are more than the parts. These phenomena are then called emergent (Page, 2011).

The difference between the two latter notions has been described by Radder (2011) in his introduction to a review of a book dedicated to complexity:

Reductionist explanations of higher-level systems in terms of their fundamental components, interpretations of the laws of nature as universal and necessary regularities, and philosophical theories of causation that see a linear and one-way relation between cause and effect fail to account for the complexity of nature. These approaches may be applicable to simple cases—for instance, in physics—but they are definitely unsuitable in the case of the unsimple truths of the sciences of complexity (p. 385-386).

In my view the complicatedness of systems may be analyzed or explained by the use of these reductionist approaches as given here.

We now turn to the ponies and their joints. What type of complexity do we have to deal with?

We follow the title of the paper. Regeneration, in general, is a complex process that proceeds through the mutual interactions of a host of cellular and non-cellular components. It fulfills the criteria of a complex in the holistic sense: emergent and in the possession of the possibility of adaptation. However, regeneration was not the object of the study we discuss here. Instead, the object of study was the finding of a method to enable regeneration of the joint.

The pony, as a biological organism, certainly is a complex system. The interactions of the components of the animal (organs, blood, brain, limbs, etc.) make a pony more than the sum of its parts, as a system it is unpredictable and able to adapt. However, it is not the object of investigation.

The joint is a complex system too, with regard to both morphology and function, as meant by McShea (1996). The complexity of the joint in the implantation experiment may be analyzed in a reductionist way; in theory the components can be taken apart to study them and new components may be added, such as an implantation construct. Following the distinction between complex and complicated systems, the joint in this study may be considered a complicated organ.

The effect of an implantation of a construct in the joint may thus be analyzed in a reductionist approach: through biochemistry, biomechanics, microscopy and, on a higher functional level, gait analysis. This analysis should be based on the temporary isolation of each of the components, how they interact and how they contribute to the morphology or the function of the system. The analysis is a step-by-step process with one step at a time only. It has therefore to fulfill the criterion of the method of difference, too. This makes the complicatedness of the joint implantation in principle a solvable problem.

References


Insight problems are sometimes designed to encourage an incorrect and misleading interpretation that veils a simple answer. The socks problem is one such problem: Given black socks and brown socks in a drawer mixed in a ratio of four to five, how many socks will you have to take out to make sure that you have a pair of the same color? The ratio information is misleading since, with only two colors, pulling three socks will guarantee a matching pair. Recently, Vallée-Tourangeau and March (2020) offered a distinction between first- and second-order problem-solving: The former proceed with and through a physical model of the problem, while the latter proceeds in the absence of such interactions with the world, in other words on the basis of mental processes alone. Vallée-Tourangeau and March also proposed a thought experiment, suggesting that the ratio information in the socks problem might be quickly abandoned in a first-order environment, that is, one where participants observe the results of drawing socks out of a bag rather than imagining themselves doing so. We tested this prediction by randomly allocating participants to a low- (second-order) or high- (first-order) interactivity condition. Marginally more participants announced the correct answer within a 5-minute period in the high than in the low condition, although the difference was not significant. Detailed analysis of the video recording revealed the challenges of operationalizing a second-order condition, as participants engaged in dialogical interactions with the experimenter. In addition, the manner in which the high-interactivity condition was designed appeared to encourage the physical reification of the misleading ratio, thus anchoring that information more firmly rather than defusing it through interactivity. We close the paper with some reflections on wide, or systemic, cognition in experimental research on creative problem-solving.

Keywords  
interactivity, insight, thought experiments, exaptative actions, mixed methods

The socks problem

Insight problems are a class of problems used to investigate non-analytical problem-solving. Some insight problems are formulated in a manner that misleads participants by directing them towards an initial incorrect interpretation, therefore resisting incremental solution strategies. The solution to these problems involves abandoning this erroneous interpretation. The sock problem is a classic insight problem. The participant is asked “If you have black socks and brown socks in a drawer mixed in a ratio of 4 to 5, how many socks will you have to take out to make sure that you have a pair of the same color?” The problem masquerades as a mathematical problem—the conversational pragmatics foregrounds the 4:5—but it is misleading information. The answer is three, no matter what the ratio is. It is possible to pull a pair with the first two socks but also to pull a brown and a black, however, the third sock would necessarily have to match with one or the other. The ratio of brown socks to black socks is immaterial and a distraction designed to set the participant on the wrong path.

Bowden et al. (2005) suggest that this problem is difficult if you approach it mathematically but easier if you use a “what if” strategy: “That is, if the solver asks, ‘What if I take out a black sock then a brown sock? I would only need one more sock of either color to have a pair of the same color’” (p. 323). Jones (2003) suggests that: “The insight here involves moving from a representation of the problem based on mathematics to a representation of the problem based on imagining yourself removing the socks” [emphasis added]” (p. 12). Again, Chu...
Take-home Message

This report describes a failed experimental manipulation in object-supported problem-solving. Participant footage allows a granular analysis of the reason for this failure. On the basis of this analysis, we conclude that sequestered experimental conditions are unstable and that the environmental resources can obstruct as well as support higher cognitive processes.

Empirical

Ross & Vallée-Tourangeau

Figure 1  The Experimental Setup
Participants were given a bag with 50 black socks and 40 brown socks (panel a). The researcher stayed in the room to give instructions and make notes (panel b).

and MacGregor (2011) suggest “[t]he problem solver may only have to run through a couple of trials in her head before arriving at the conclusion that, at most, you only need to draw three socks to match a pair” (p. 211). In a later description of Fleck and Weisberg (2013), Weisberg (2015) noticed some successful participants imagined “taking the socks out of the drawer, and considering the information available at each step” (p. 32).

In other words, while the problem invites an initial mathematical representation, the solution not only requires the participant to see the problem non-mathematically but is likely facilitated when they see themselves in a real-world environment where they can solve it in a trial-and-error way. Already there is a bridge between mental abstraction and embodied experience being posited here—the problem is hard to solve on the basis of “pure” mental processes. Therefore, it seems almost trivial to suggest that placing participants in such a real-world situation would augment problem-solving—rather than the problem-solver imagining herself drawing socks from a bag, she could actually do that. Indeed, this is what one of us predicted in Vallée-Tourangeau and March (2020).

Forty-one participants (a mix of undergraduate and postgraduate psychology students) were recruited in exchange for course credits. One participant was excluded for knowing the problem, one for not adhering to the stipulations of having English as a first language, and one because of data loss. This left 38 participants (31 women) with a mean age of 26.28 years (SD = 12.14). The participants were assigned in turn to the two conditions. Eighteen participants took part in the low-interactivity condition and 20 participants took part in the high-interactivity condition.

Procedure

Participants were filmed in a purposely-built lab with three in-built cameras to facilitate coding for the behavioral hypothesis. To minimize the differences between the conditions, both high- and low-interactivity conditions had the same initial setup: In both conditions, each participant was presented with a bag of 90 socks, among which were 50 black and 40 brown, and the only difference was that participants in the high-interactivity condition were allowed to pull socks from the bag (see Figure 1).

The original instructions ran thus: “If you have a drawer with brown socks and black socks mixed in a ratio of 4:5, how many would you have to pull out in order to guarantee a pair?” The instructions were read aloud to participants. However, after piloting, these needed to be refined and the final instructions ran thus: “In this bag are brown socks and black socks in a ratio of 4 brown socks for every 5 black socks… so the ratio of brown socks to black socks is 4 to 5, is that clear? What is the minimum number of socks you think you need to take out of the bag at random to be sure you had a pair of either brown sock or back socks, so that is if you were to pull socks out at random from the bag, what is the minimum number of socks that you would need to be sure that you would have one pair, that’s two socks which match in color, whether that’s one pair of brown socks or one pair of black socks.”

Participants were given 5 minutes to answer the problem and they were allowed to make multiple attempts. They were not prevented from speaking to the researcher nor, however, were they requested to follow a think-aloud protocol. The researcher would only answer questions by repeating the relevant part of the instructions. The researcher kept detailed notes while observing the participants; the participants were recorded throughout the study and the videos were later watched and coded. The participants’ performance was measured in terms of success and the latency to a correct solution.

Results

Performance was better in the high-interactivity condition in which 10 participants (or 50%) solved the problem compared to the low condition in which 7 participants (or 38%) did so. This difference was not significant, however $\chi^2(1, N = 38) = 0.47, p = .492$, Cramer’s V = 0.11. Participants were slower in the high-interactivity condition with an average latency of 116.20 seconds (SD = 78.67; although this was skewed by participant 41 who took 297 seconds) than in the low condition ($M = 103.57 \text{s, } SD = 48.32$, which was skewed by participant 18 who announced the answer immediately, hence with a latency of 0; see Figure 2); this difference was not significant, $t(15) = 0.38, p = .713$, Cohen’s d = 0.19. Removing these
two outlying participants, the mean latencies were 96.1 s (SD = 49.4) and 120.8 s (SD = 17.3) in the high- and low-interactivity conditions respectively (t(13) = 1.167, p = .264, Cohen’s d = 0.82).

Qualitative section

Method
Over the course of the research, it became clear that the experimental manipulation failed and that there were significant procedural and theoretical assumptions and errors. Some were failures in experimental procedures\(^1\) which could be rectified in a second study, but as we will see, others undermined the initial premise of the study and led to the decision to not repeat the study.

Additionally, the video data suggested that there were underlying patterns from which new knowledge about this problemsituation could be deduced beyond that contained within the structures of the hypothetico-deductive model. In other words, theoretically valid reasons for the practical and experimental failure of the experiment could be seen in the behavior of participants throughout the experimental situation. A decision was made to employ a modified version of Grounded Theory Method (GTM; Glaser & Strauss, 1967) with the video data to generate novel theories about problem-solving in this task. Thus, the focus of the research shifted from an experimental manipulation to an observational study assessing qualitatively how participants solve this problem. Consequently, we abandoned the original hypotheses.

Analytical process

The initial observations generated in vivo were followed up by a close examination of the video data with time-stamped qualitative memos. Conversa-\(^1\)tions were transcribed and time stamped. These observations were then grouped into conceptual categories before the videos were reviewed a second time to substantiate these categories. This iterative process continued through initial drafts of this paper. In tandem with the data analysis carried out through watching the videos, results were discussed between both authors and conceptual categories were refined.

Results
The initial hypothesis of an augmentative effect of a high-interactivity environment compared to a low-interactivity environment was not sustained. The qualitative analysis detailed below pinpoints why this might be so: Excess environmental information and the encouragement to interact with it reified unhelpful representations. Additionally, a form of social and linguistic interactivity which focused on the instructions and the researcher replaced the object-based interactivity. The iterative process of discovery emerged more often in this linguistic space than in the space between object and person. This was a form of interactivity that was not accounted for in the rather narrow (and, in hindsight, naïve) high- and low-interactivity conditions outlined above, and yet it occurred in both experimental conditions and was in part the reason for the inconclusive results and the failed manipulation.

Information in the environment
The socks problem is a hard one because participants get stuck on the unhelpful ratio. As suggested by Vallée-Tourangeau and March (2020), ignoring the ratio should lead to participants solving the problem more often and faster because they are not tempted to use inappropriate mathematical formulae. As we have addressed above, this is not an unreasonable assumption. However, the specific hypothesis under exploration, that a high-interactivity environment would increase the number of people who disregard the ratio, was not sustained by an analysis of what people did in this condition; it appeared that the opposite was the case, that a high-interactivity reinforced the unhelpful information.

There are two ways in which a problem-solver could approach solving a problem using interactivity: The first involves the sort of exaptative actions mentioned in the initial hypothesis, while the second involves using the world to scaffold existing representations. In the case of problems designed to elicit an unhelpful representation, it is plausible that such a representation, when reified, would be harder to disregard. A high-interactivity environment can scaffold unhelpful representations as well as helpful ones.

This reification was clearly in line with evidence in the study reported here: Interactivity allowed partic-
participants to represent the ratio in a solid form. Many of them pulled five brown socks from the bag and then four black socks and used these to structure their thoughts, so the ratio became harder to disregard. This was a sensible decision; it is not unreasonable to assume that the information in the problem is important (in terms of conversational pragmatics, it would make little sense to be informed of the ratio if it was not relevant; Grice, 1975). Therefore, the strategy of reifying the ratios is on the one hand sensible, and paradoxically, on the other hand, makes the problem harder to solve.

However, in addition to allowing the reification of the information in the problem, the presentation of actual socks expanded the amount of information and the number of possible unproductive pathways. Problem-solving in this less controlled environment did not scaffold thought but rather created an additional distraction. Indeed, the concrete anchoring of the problem led people to look at the distribution of socks in the actual bag that they had in front of them rather than considering it an abstracted entity. This was true even for those who were in the low-interactivity condition who tended to return their gaze to the bag when thinking. Take for example, participant 36: Figure 3 demonstrates the extent to which she was looking at and into the bag even while not interacting with the socks in it. Throughout the course of the experiment, she looked at the bag and the accompanying finger gestures and mouthing of numbers indicate that she was trying to count the socks that were in it.

So, rather than supporting problem-solving, the actual physical arrangements of socks added complexity. Participant 26 (low-interactivity) said “Well looking at this bag, all the black socks are at the top so it would take a while to get a brown pair” (01:36). Within a traditional problem-solving environment, participants know that they are solving a riddle and that they are in an artificial situation. In this experiment, the artificial, almost school-type problem collapsed with the far too mundane situation of finding a pair of socks. There simply is no “real-world” version of the hypothesized riddle. Somehow, the simplicity of this quotidian activity supported the intractability of the mathematical puzzle: If you want a pair of socks, you get them out of the bag.

Additional information was deeply unhelpful in this case. When the problem is presented as a verbal riddle, the number of socks is not given; not only is there no way of knowing it, it is also irrelevant. However, when the participants are presented with a bag full of socks and a problem masquerading as a mathematical one, the total number of socks appears to take on more importance, much like the ratio information in the verbal form of the puzzle. In addition, investigating this unproductive cul-de-sac was easier for those in the high-interactivity condition and therefore harder to disregard. Take for example participant 37: After asking for clarification of the problems at 33 seconds into the task, he asked: “How many were there in total?” (00:33). After 58 seconds elapsed, he decided to empty the bag and to count the socks. It is hard to overstate what a poor decision this is (illustrated in Figure 4 where a screenshot was taken every 10 seconds over the following 220 seconds). Perhaps like participant 25, who stated “Ah, that’s what I have to do, I have to know how many are in here” (02:15) before tipping the bag out, the participant thought he had solved the “trick”. Whatever the reason, the messiness of tipping out all the socks then counting them and the high time cost meant that he made no progress. This wrong route would not have been possible in a low-interactivity or second-order problem-solving environment, yet again it is important to stress that this is not an illogical decision.

The socks problem is a riddle commonly employed in research on insight problem-solving. The riddle goes: If you have black socks and brown socks in a drawer mixed in a ratio of four to five, how many socks will you have to take out to make sure that you have a pair of the same color? The ratio information is misleading since, numbers indicate that she was trying to count the socks that were in it.

So, rather than supporting problem-solving, the actual physical arrangements of socks added complexity. Participant 26 (low-interactivity) said “Well looking at this bag, all the black socks are at the top so it would take a while to get a brown pair” (01:36). Within a traditional problem-solving environment, participants know that they are solving a riddle and that they are in an artificial situation. In this experiment, the artificial, almost school-type problem collapsed with the far too mundane situation of finding a pair of socks. There simply is no “real-world” version of the hypothesized riddle. Somehow, the simplicity of this quotidian activity supported the intractability of the mathematical puzzle: If you want a pair of socks, you get them out of the bag.

Additional information was deeply unhelpful in this case. When the problem is presented as a verbal riddle, the number of socks is not given; not only is there no way of knowing it, it is also irrelevant. However, when the participants are presented with a bag full of socks and a problem masquerading as a mathematical one, the total number of socks appears to take on more importance, much like the ratio information in the verbal form of the puzzle. In addition, investigating this unproductive cul-de-sac was easier for those in the high-interactivity condition and therefore harder to disregard. Take for example participant 37: After asking for clarification of the problems at 33 seconds into the task, he asked: “How many were there in total?” (00:33). After 58 seconds elapsed, he decided to empty the bag and to count the socks. It is hard to overstate what a poor decision this is (illustrated in Figure 4 where a screenshot was taken every 10 seconds over the following 220 seconds). Perhaps like participant 25, who stated “Ah, that’s what I have to do, I have to know how many are in here” (02:15) before tipping the bag out, the participant thought he had solved the “trick”. Whatever the reason, the messiness of tipping out all the socks then counting them and the high time cost meant that he made no progress. This wrong route would not have been possible in a low-interactivity or second-order problem-solving environment, yet again it is important to stress that this is not an illogical decision.
Figure 4  Participant 37 tidying socks. Stills were taken at 60 seconds and every 10 seconds thereafter.
Porous cognitive boundaries

Additionally, the video data revealed versions of cognitive interactivity which were not related to the moving of the objects and hence did not feature in the original hypotheses but which scaffolded the pathway to solution. For instance, the instructions played a critical role in this scaffolding. This was an unintended artifact of the experimental situation and, albeit subsequently serendipitous, an inelgant one. The instructions were given verbally, and the participants were allowed to ask for clarification. The clarification took the form of the experimenter repeating the relevant set of instructions. Thus, what is often unspoken (checking instructions) became a traceable experimental artifact.

What this underscored was the importance of the instructions to the problem solution. Initially, there was a change made to the instructions, as outlined above, to ensure that all salient information was clearly presented, but this did not change the participants' need for clarification. This clarification revealed two things: First, a consistent epistemic state across participants despite consistent task instructions should not be assumed; second, it was possible to use these clarifications in a trial-and-error way to structure understanding and gain new knowledge. Indeed, the nature of the task requires restructuring the instructions rather than manipulating the visuospatial field. Therefore, it is not unreasonable that the interactivity coalesces on this linguistic plane.

The verbal nature of the instructions also increased interaction with the researcher. This was unanticipated in the original research design and invited the researcher into the experimental situation in a way normally avoided in experimental psychology. What these interactions revealed was the need for the participant to seek help from the outside world. Over the course of these interactions, the answer often revealed itself through a gradual recursive dialogical process. Taking for example participant 41 (Table 1), the last minute or so of her problem-solving involved a recursive, back-and-forth between extracting information from the socks that she had in her hand, the instructions, and the researcher through making guesses until she suggested the correct answer. Note how often she looked at the researcher and solicited feedback in the episodes selected. In this case, the playing with socks coupled with the information from the researcher and the redrawing of the epistemic map through the wrong guesses scaffolded the way to a solution.

This was even clearer in the low-interactivity condition where the participants sought out scaffolding more obviously. Take for example, participant 26 in Table 2. This was a low-interactivity participant who uncovered the answer through a series of guesses using the feedback as a scaffold. As she stated clearly when asked if she knows why: “When I got to the end I did...when I said two and then I realized...” (02:51). The act of saying the word alongside the feedback from the environment scaffolded her understanding. Knowledge was gained when the thought was in the world; even if that realization did not take material form, it was still generated in action, this time the action of guessing. To place the results from this participant in a low-interactivity environment would be misleading even if she did not carry out any actions on any objects.

Discussion

The quantitative results were inconclusive. It may be that if the study had reached the anticipated number of 60 participants, they would show a significant augmentative effect of interactivity, but that seems unlikely. Moreover, the cognitive messiness of the procedure would mean that it would be hard to trust that the results reflected the experimental manipulation, that is, one group of participants interacting with the physical model of the problem, and one group interacting with no features of the task environment, including with the researcher! Of course, the problems outlined above would cast the experiment as a failure by the standards of traditional experimental psychology. As argued by Gozli (2017), the experiment in psychology takes the form of rule-governed behavior. In the experiment reported here, the participants did not play by the rules of the experiment, namely to use only mental simulations in the low-interactivity, sequestered condition and to usefully employ the objects in the high-interactivity condition. Thus, the experimental manipulation and the experiment failed.

It would be possible, of course, to design a low-interactivity task environment where interaction with the experimenter was completely removed beyond the initial task instructions (e.g., the experimenter could walk away from the observation lab and take their notes from the control booth). Accordingly, the high interactivity condition could be designed to force participants to take one sock at a time, instruct them to reflect on what they see, and prompt them to provide an answer to the puzzle after each draw. Such task procedures might improve the operationalization of the levels of interactivity envisaged in Vallée-Tourangeau and March (2020)’s original thought experiment. Be that as it may, the data reported here unveil the affordances offered by the elements of the wider system, and the felicitious (such as soliciting feedback dialogically in the low-interactivity condition) and infelicitous outcomes (such as reifying the 4:5 ratio in the high-interactivity condition) that result from the participants’ actions. Such an analysis is likely to yield a greater understanding of how problem-solving unfolds in more complex situations.

The qualitative analysis undertaken here also
Table 1: Participant 41 solving the problem. The timestamps for Table 1 and Table 2 refer to the time from the start of the instructions.

<table>
<thead>
<tr>
<th>Time</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>04:53</td>
<td>Pulls out two more - this time both black</td>
</tr>
<tr>
<td>05:07</td>
<td>&quot;This is 8.....&quot;</td>
</tr>
<tr>
<td>05:08</td>
<td>Looks at researcher</td>
</tr>
<tr>
<td>05:14</td>
<td>Pulls through socks in hands</td>
</tr>
<tr>
<td>05:21</td>
<td>&quot;So I have to pull out a pair?&quot;</td>
</tr>
<tr>
<td>05:23</td>
<td>Researcher repeats instructions</td>
</tr>
<tr>
<td>05:35</td>
<td>&quot;So I could pull out four socks and then two would be a pair&quot;</td>
</tr>
<tr>
<td>05:42</td>
<td>Researcher repeats request for the minimum number</td>
</tr>
<tr>
<td>05:45</td>
<td>Plays with socks in her hand</td>
</tr>
<tr>
<td>05:50</td>
<td>&quot;Three&quot;</td>
</tr>
<tr>
<td>05:57</td>
<td>Correct explanation</td>
</tr>
</tbody>
</table>

offers an understanding of the detrimental role that interactivity can play alongside the traditionally-theorized augmentative role. In hindsight, such an effect is theoretically more plausible given an open cognitive system and a non-computational approach. The data presented here suggest that a
Table 2  Participant 26 solving the problem through incremental guessing

<table>
<thead>
<tr>
<th>Time</th>
<th>Audio Transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>01:18</td>
<td>“Oh what and if I get it right you’ll tell me?”</td>
</tr>
<tr>
<td>01:21</td>
<td>“Hang on, I don’t know how many socks are in the bag?”</td>
</tr>
<tr>
<td>01:24</td>
<td>Looks at researcher expectantly.</td>
</tr>
<tr>
<td>01:27</td>
<td>“So I don’t get how the ratio would matter if I don’t know how many socks were in the bag…oh. it’s just chance isn’t it?”</td>
</tr>
<tr>
<td>01:36</td>
<td>“Well, looking at the bag actually all the black ones are on top so it might be a while to get a pair of brown ones…”</td>
</tr>
<tr>
<td>01:45</td>
<td>“I’m so confused by the question” Looks at researcher again</td>
</tr>
<tr>
<td>01:47</td>
<td>Researcher repeats instructions</td>
</tr>
<tr>
<td>01:50</td>
<td>Starts to pull socks; Researcher stops her</td>
</tr>
<tr>
<td>02:03</td>
<td>Researcher ends instructions</td>
</tr>
<tr>
<td>02:08</td>
<td>“When I got to the end I did…when I said 2 and then I realized….”</td>
</tr>
<tr>
<td>02:11</td>
<td>“I’d have to pick 10 socks out”</td>
</tr>
<tr>
<td>02:17</td>
<td>“I’d have to pick 8 socks outs”</td>
</tr>
<tr>
<td>02:23</td>
<td>“I’d have to pick 20 socks out”</td>
</tr>
<tr>
<td>02:32</td>
<td>“I’d have to pick 18 socks out”</td>
</tr>
<tr>
<td>02:41</td>
<td>“I’d have to pick 4 socks out”. Said faster and more determined</td>
</tr>
<tr>
<td>02:44</td>
<td>“2 socks”</td>
</tr>
<tr>
<td>02:47</td>
<td>“3 socks”</td>
</tr>
<tr>
<td>02:49</td>
<td>Researcher confirms</td>
</tr>
<tr>
<td>02:50</td>
<td>Starts to laugh</td>
</tr>
</tbody>
</table>

cognitive system forms around any problem-solver and that once we approach cognition from this systemic perspective, we necessarily introduce multiple shifting parts. Here, when interactivity with the objects proved to be unhelpful, cognition seeped out into the wider system to encompass the interaction with the researcher via the instructions. These observations license two broader comments on the nature of experimental research in interactivity: (a) on problem-solver omniscience and (b) on the inherently contingent nature of wide cognition.

The omniscient problem-solver

Problem-solving is broadly the movement from a state of ignorance to a state of knowledge (Arfini, 2021). There are two forms this ignorance can take, ignorance of the process of problem solution or ignorance of the answer. The most efficient pathway to a solution can only be directly and easily enacted when the solution is already known and can be traced backward without the risk of a false start or a complete divergence from the path. This requires foreknowledge of the correct answer, which is a paradoxical situation and assumes an omniscient problem-solver.

For analytical problems, while the answer is unknown, the process of getting to the answer is clear and known. For example, for mental arithmetic, while the problem-solver may not know the solution, they will know the most efficient process and the simple operators to evince it. These steps will have been predetermined culturally and through personal experience. Therefore, it is likely that the agent will select the correct objects and the correct actions over those objects. This means that an agent-centric model of interactivity should yield supportive empirical data. However, it is unclear how far such a model would extend. In many ill-structured problems or non-analytical problems, the problem-solver is in double ignorance: They do not know the correct solution or the manner of approaching the correct solution. In terms of problem-solving with and through objects, this means that they will not necessarily select the correct objects or actions over them but merely those that satisfy the needs at the time (echoing the Criterion of Satisfactory Progress Theory; MacGregor et al., 2001; Ormerod et al., 2013). Without omniscience, the problem-solver can make logical and plausible moves that lead them further from the problem solution.

Wide cognition is contingent

The components in complex systems interact in complex ways and the results are emergent and contingent. Pre-specification of underlying causal mechanisms is difficult because adhering to a form of wide cognition (see Wilson & Clark, 2009) means that such pre-specification is necessarily an underspecification. It is impossible to predict in advance which solution routes will be useful, especially if that problem solution is generated by unplanned and exaptive gestures, which is more likely in materially rich environments. It suggests that in insight tasks where there is a double ignorance—ignorance of the solution and ignorance of the path to the solution—both the finding of the solution and the path taken become important.

Interactivity posits that a dynamic relationship with the external world is not only augmentative to cognition processes but necessary. However, experimental research in interactivity does so by paradoxically setting up a low-interactivity condition which admits non-interactive thinking. In this way, it assumes that there is an easy way to compare non-systemic and systemic cognition and that non-systemic cognition can be meaningfully isolated in an experimental psychologist’s lab. We suggest that the dualistic assumptions underlying this approach are misguided, as they assume an ideal type of experimental participant can be generated in sequestered experimental conditions. Rather, the evidence we present here suggests a cognitive leakage across both experimental conditions. In the process of operationalizing the so-called low-interactivity condition, the presence of the experimenter offered an external resource with which to scaffold the problem-solving process. The participant naturally transformed a solo and ostensibly unaided effort into a dyadic one. Clearly, the experimenter qua interlocutor was an unrequired conversational partner. Still, in the absence of hard physical or instructional constraints, the participant naturally availed themselves of this resource, bootstrapping the cognitive effort through a dialogue with their reluctant partner. And it is through this dialogue that the participant’s hunch underwent many different forms until it was eventually translated into the normative one. These ersatz dialogues offer a telling window into the distributed nature of thinking, in this instance, the distribution over time and the gradual sedimentation through iterative dialogical cycles. The reification of the solution was the product of interactivity despite the experimental procedure.

A systematic approach to cognition requires that every thoughtful (and arguably every non-thoughtful) encounter should be analyzed as part of a system; however, what constitutes part of the system is rarely specified. Vallée-Tourangeau and March (2020) suggest that a cognitive ecosystem is configured by “the reasoner, the physical reality of the problem, and the action possibilities offered by the external environment” (p. 826). We argue that the data here suggest “the action possibilities offered by the external environment” (p. 826) are broader than whether the problem is represented...
in movable artifacts or not. Rather, participants actively recruit beyond the resources of interest in the experimental conditions. This means a more granular approach needs to be considered.

Conclusion

The data presented here are not the clean data typically seen in psychological reports of experimental situations. Over the course of the study, a more ethnographic approach to the whole experimental situation provided rich data about different types of interactions that emerged in either a so-called low- or high-interactivity condition. Some aspects are unique to the situation of an exploratory, quasi-pilot study in problem-solving and others are broader reflections on the nature of research in psychology. As we have seen above, the transformation from a sequestered condition to a more “real-life” situation does not necessarily occur without cost and the nature of that cost deserves to be explored. Additionally, it casts doubt on the possibility of creating a fully sequestered condition. Rather, the data presented here suggest that both conditions are porous.

The failure of the experimental procedure here revealed the difficulty of doing research in open cognitive systems. Traditional methods of assessing embedded or extended cognition are still rooted in a cognitive and computational model of mind and contrast low- and high-interactivity environments to demonstrate an augmentative effect of interactivity. The research presented here suggests that the reality of experimenting in an open system is more complex than is believed. The target problem here was designed as a mental riddle and its transposition to an interactive space offered a telling window into the disconnect between first- and second-order problem-solving: The task only “works” as a second-order task.

It is tempting to question whether the mental tasks used in traditional cognitive psychology experiments are diagnostic of problem-solving outside of the sequestered laboratory environments in which they are commonly operationalized. We propose, rather, that a research program should develop in parallel to examine the solving of problems which are rendered complex not because of their structure but because of their situation. When it comes to problem-solving, we need a wider range of outcome measures to even begin to understand how it unfolds across a range of contexts, which requires a combination of qualitative and quantitative work. What we can see here is that the genesis of a new idea involves a transformation, which in turn involves resources; these resources come with a cost, and these transactions leave physical traces which can be mapped. To understand the transformations and transactions, that is, the process through which new ideas are constructed, the analysis must be local and granular. It is our belief that turning to the qualitative in cognitive psychology to complement the quantitative will yield great benefits.

References


Cognition stays wild: A commentary on Ross and Vallée-Tourangeau’s Rewilding Cognition

Vlad P. Glăveanu1,2, Alex Gillespie3,4

A “failed” experiment (Ross & Vallée-Tourangeau, 2021) tried to reveal the role played by materiality in solving an insight problem that made reference to embodied action, leading to valuable insights about the nature of cognition and the experimental method. In this commentary, we argue that this study reveals various forms of interactivity and brings new evidence against the idea that “pure” cognition can be isolated from either materiality or sociality. The question becomes, then, not whether the use of objects helps or hinders problem solving, but how objects, bodies, and other people participate in it, even in controlled lab settings, and to what effect. Reflections are offered on why and how cognition stays wild (i.e., embodied, dialogical, and surprising) and what this means for experimental work.

Keywords  cognition, problem-solving, interactivity, materiality, sociality, experiments

Ross and Vallée-Tourangeau, Ross, W., & Vallée-Tourangeau, F. (2021). Rewilding Cognition: Complex Dynamics in Open Experimental Systems. https://doi.org/10.36850/e4 (2021) distinguish first-order (hands-on and interactive) and second-order (cognitive and abstract) problem solving and set out to experimentally demonstrate the neglected superiority of first-order problem solving for certain creative tasks. To this end, they use the socks problem: “If you have a drawer with brown socks and black socks mixed in a ratio of 4:5, how many would you have to pull out in order to guarantee a pair?” They expected, with good reasons, that participants who interacted with socks would be better at re-solving the problem compared to participants who only imagined interacting with socks. However, their results revealed an inconclusively small difference between the conditions in terms of the number of participants who solved the problem and the time taken to do so.

The finding of this ostensibly failed experiment is insightful and richly developed by Ross and Vallée-Tourangeau. They note, with the power of hindsight, that they had been overly focused on the interactivity with the socks while failing to consider the multiple other dimensions of interactivity within the research design: interactivity with the instructions, internal dialogical interactivity, interactivity with the researcher, and interactivity with any available resources. Indeed, even the interactivity of participants who had socks was uncontrollable, with some upending the sock bag and getting distracted by counting the socks. They conclude that interactivity permeated both conditions, and this realization has broad implications for experimental research that tries to control interactivity. Interactivity, it seems, is so fundamental to humans that it cannot be experimentally isolated.

Pervasive Interactivity

To further illustrate the point, there is yet another layer of interactivity that was not considered by Ross and Vallée-Tourangeau, but which is evident in their data, namely the larger dialogical context of what it means for participants to be in a “psychology experiment.” Few contemporary participants are engaging in their first experiment, and few have not heard about experiments with confederates, deceptions, and creative twists. Thus, another dimension of interactivity is that participants know that the experiment is a game and that the experimenter may have hidden motives. They expect a trick, and this makes the seemingly simple instructions seem puzzling. We see this in their data. Participants 41, five minutes into the task asks: “so I have to pull out a pair?” (5:21). Participant 26, nearly two minutes into
the task states: “I’m so confused by the question” (1:45). Yet, the instructions given to participants were ostensibly straightforward. We suggest that what these participants are confused about is what they are “meant” to do (i.e., what is the experimental context around the text of the instruction). Participants are thus in a mental dialogue with the genre of the experiment. They are unsure what the experimenters were looking for and whether even the question itself was a trick. In short, the experimental situation has layers of interactivity that go far beyond the situation of waking up in the morning and trying to find a pair of matching socks in a sock drawer.

This pervasive interactivity within the experiment brings us to the “rewilding” metaphor used in the article. Their choice to introduce materiality (intentionally) and sociability (unintentionally) into the experimental design is the equivalent of opening the door to the “wilder”, harder to measure and control, elements of cognition. Reminiscent of older notions of “cognition in the wild” (Hutchins, 1995), where sanitized laboratories are replaced by messy, interactive settings of action and interaction, the use of “wild” here and elsewhere raises a legitimate question: did we ever manage to domesticate cognition in the first place? Did we achieve experimental designs that can, at will, cut off the mind from body, others, institutions, and culture (to name but a few “unruly” elements), in order to study “pure” thought (or what Vallée-Tourangeau and March in 2020 called second-order problem solving)? Decisively, no. We can certainly create methodologies that foreground abstract thought and place bodily movement, object manipulation, and dialogues with others into the background. Still, the influence of the latter can never be discounted. The fact that participants don’t visibly use their hands to tinker with objects doesn’t mean that they are any less embodied or that material forms of engagement have no part to play in their ongoing mental operations. Similarly, being left alone to solve a task doesn’t make that particular moment or context asocial. Inner forms of dialogicality are ever-present—the other is always there, just like bodies are. The question is, then, not to try and compare “wild” and “tame” forms of cognition, “muddied” and “pure”, distributed and internal, in terms of the participants’ performance in an experiment. A more interesting question is the one Ross and Vallée-Tourangeau raise: how exactly do materiality and sociability participate in cognition (here, problem-solving) in ways that either accelerate or hinder performance?

Socially, we suggest, contributes to creativity by expanding possibility through introducing alternative perspectives (Glăveanu, 2020; Zittoun & Gillespie, 2015). While creativity tasks are often based on divergent thinking, insight problems call for a better balance between divergent and convergent thinking. They also depend on social-psychological processes like perspective-taking (Glăveanu, 2015) that connect “creators” to their audiences. In the context of the study, the most direct audience was the experimenter and as such, it is not surprising that participants tried to propose, clarify, or specify their perspectives directly to her. This reaching out to the other not only serves the purpose of reaching the most suitable outcome, but it is also an engine for creative ideation. Creative insight is not the personal or internal moment most imagine it to be. It is, in fact, the result of past and present dialogical experiences in which the perspectives of the self are placed in dialogue with those of others (and the more diverse these others are, the more likely it is for participants to reach more creative solutions; Gassmann, 2001). The experiment reported by Ross and Vallée-Tourangeau was not designed to allow for creative outcomes but to end with a numeric answer, the number of pulls needed to reach a pair. It is interesting to imagine if allowing the participants to have a more playful and social approach to the problem (e.g., being in dialogue with other participants or the experimenter) might itself lead to more creatively productive aims.

The pervasive interactivity across both experimental conditions does not mean that Ross and Vallée-Tourangeau’s initial guiding research question was mistaken. There is a lot of evidence for humans having two types of cognition, as evident in the many “dual-process” models of cognition. It is evident in Vygotsky & Luria’s (1994) distinction between the mental functions humans share with other primates and the symbolically mediated mental functions that seem more peculiar to humans. And, more recently, it is evident in the distinction between thinking fast and slow (Kahneman, 2011). These two modes of thought seem pervasive in human cognition, not just creative problem solving. However, this does not mean they can be experimentally separated with ease. A mind capable of both first- and second-order problem solving is likely to leverage both within any real problem solving episode (why limit oneself to only one cognitive process?). The creative mind is characterized by movement within and between styles of thought (Gillespie & Zittoun, 2013). In the first-order condition, participants may close their eyes to conceptualize the problem abstractly. In the second-order condition, participants may simulate the socks using pen and paper or their mind’s eye. In short, in cognition, as it naturally occurs in the wild, humans likely oscillate between these two modes and perhaps even participate in both simultaneously—and it is this overlap of cognitive modes that holds the key to creative cognition.

Interestingly, this idea, that creative cognition stems from moving between concrete and abstract modes of thought, is evidenced by Ross and Vallée-Tourangeau’s own methodology. They begin with
a quantitative analysis that is abstracted, with conditions differentiated by numbers. However, in the qualitative analysis, they engage with the concrete particulars—just like participants who had access to the bag of socks, they opened up the experiment to have a look inside and see what was actually going on. This leads them to observe interactivity in the so-called non-interactive condition. Participants talked to the experimenter, asked questions about the question, and pitched answers to see the response. These concrete observations were then integrated back into a more abstract understanding of rewilding cognition not just in their own experiment, but in experiments more generally. Thus, Ross and Vallée-Tourangeau arguably, engaged in both first- and second-order problem solving in their own creative analysis.

One of the innovative aspects of the article, which Ross and Vallée-Tourangeau only make passing reference to, is the combination of qualitative analysis within an experimental design. It enables them to be both concrete about what went on in the experiment (analyzing videos) and also to be more abstract (assessing statistical differences). In the nomenclature of mixed methods research, they used a peculiar variant of an explanatory sequential design (Creswell & Creswell, 2018). These designs start with a surprising quantitative finding and then use qualitative research to generate plausible explanations. Ross and Vallée-Tourangeau’s variant of this design is peculiar because the quantitative and qualitative parts of the analysis pertain to the same events (participant behavior in the experiment). Normally, mixed methods sequential designs use separate data sources (e.g., a survey followed by an interview). Combining qualitative methods within the quantitative experiment provides two different lenses on the same behaviors—thus enabling each to elucidate the other. This qualitative-quantitative analysis of the behaviors within the experiment enabled them to zoom out to identify a surprising lack of difference between conditions and then zoom in to describe what participants were actually doing.

Conclusion
What the present studies and several others (e.g., Glăveanu et al., 2019; Hawlina, 2019) show us is that rethinking experiments in psychology, and other disciplines, remains an urgent task. The video recording of participants in experimental studies or conducting post-experiment interviews or focus groups should be the rule in this kind of research, not the exception. The possibility of collecting qualitative data within research designs that are meant to be quantitative and focused on measurement should be seen as a valuable opportunity by experimentalists. Detailed data about the actions, interactions, and beliefs of one’s participants are not meant to be analyzed mainly when the statistical analysis failed to offer the desired finding—it should become a source of deeper reflection about the phenomenon under study and about the participants being studied in most cases. Of course, there are pragmatic reasons for not overloading researchers with collecting and especially analyzing datasets that, on the surface, seem highly distinct. But, as Ross and Vallée-Tourangeau’s research shows, these are not even necessarily different datasets—they can be part of a single account of the course of action prompted by the experiment, the origin of both quantification and of qualitative forms of interpretation. Adding cameras or making room for interviews within a traditional experiment is not meant to disturb standardization or reduce control. On the contrary, these additions can offer precious insights into how things happened and what actually happened, above and beyond the narrow measurement of changes in a few pre-selected dependent variables (see also Glăveanu et al., 2019).

Moscovici (1991) lamented the separation between experiments and experience when experiments in psychology necessarily entail manipulations of participants’ experience. Too rarely have experimenters “opened the black box” of the experiment to examine what is actually going on (Corti et al., 2015; Psaltis, 2007). Ross and Vallée-Tourangeau’s use of qualitative methods within the experiment is an exciting illustration of the insights that can be obtained by studying what goes on in experiments, by moving between the abstract statistical type of analysis and the more concrete, grounded qualitative type of analysis. Ironically, their delving into the concrete particulars of their own experiment to examine what happened demonstrates the point that they failed to establish experimentally, namely, that first-order concrete thinking can pump insight and problem solving. It is also a vivid reminder that lab-based cognition doesn’t become “wild” whenever the door is opened for the physical manipulation of objects or the possibility to be in dialogue with the experimenter. It stays as wildly embodied, dialogical, and surprising as it ever was.

References
Gillespie, A., & Zittoun, T. (2013). Meaning making in motion: Bodies and minds moving through institutional and


Real-effort survey designs: Open-ended questions to overcome the challenge of measuring behavior in surveys

Caroline Fischer

Based on data triangulation, open-ended questions can be used to overcome a typical problem with data collection using surveys: Human behavior can only be captured as stated or intended, but not as real behavior. In this study on knowledge sharing in the workplace, a quantitative measure of behavioral intention was accompanied by such a qualitative, open-ended measure of behavior. The latter was used as a proxy for real instead of stated behavior. This item was coded according to the effort a participant made in answering. It is assumed that the greater the effort put into answering the open-ended question, the more likely it is that the described behavior will be performed in reality. A factorial experimental design was used to analyze the effect of rewards on employees' knowledge-sharing behavior. As a within-subject design was used, participants had to answer three open-ended questions referring to different vignettes. A strong order effect appeared, leading to longer answers on average for the first vignette (baseline) compared to subsequent vignettes, independent of treatment. Therefore, this approach to operationalizing behavior in surveys might not be useful in within-subject designs. However, it can be used in between-subject comparisons when participants are asked to answer to a single vignette.

Keywords survey design, survey experiment, real-effort design, human behavior, order effect

Measuring human behavior is an ongoing challenge for social scientists (Schwarz & Oyserman, 2001). In ideal circumstances, human behavior should be observed either in the field or in the lab. However, field trips and lab experiments are often both time- and resource-consuming. Observing behavior can also be prohibited in some cases, for instance if the behavior occurs too infrequently to be observed in a timely manner, or is not possible for others to observe (Schwarz & Oyserman, 2001, p. 128). Additionally, in the field of public administration and other fields using sensitive target groups, it is often difficult to recruit participants for a lab experiment or to even be allowed to observe them in the field. If lab experiments are conducted, they usually rely on students as participants. However, student samples are not externally valid, especially when a special population is being studied, such as public employees.

Therefore, self-reported behavior in surveys and survey experiments is an often-used alternative (James et al., 2017, p. 120). Surveys are “quick and cheap” (Friborg & Rosenvinge, 2013, p. 1398) and can easily reach an intended population. However, these approaches to data collection come with other disadvantages, such as not measuring real behavior. One solution could be to integrate open-ended questions as a measure of answering effort, and thus a proxy for real behavior, into quantitative surveys. This kind of method and data triangulation could help to overcome limitations of surveys related to the measurement of behavior.

Real-effort designs assume that greater effort in fulfilling a task makes it more likely that participants actually perform what they have stated (Dutcher et al., 2015). Fulfilling real-effort tasks are typically cognitively, creatively, or physically costly for participants (Charness et al., 2018). These costs are higher when more effort is put into fulfilling a task. In some studies, for example, participants are asked to count errors in real-effort tasks, thereby increasing the cognitive cost of the task (Andersen et al., 2018; Gneezy & List, 2006). Hence, it can be assumed that the more extensive, accurate, and fitting an answer to a question about future behavior is, the more likely it is that the described behavior will actually be performed in the future.

To test this approach to measuring behavior, a survey experiment was conducted on the effect of rewards on public employees' knowledge-sharing behavior. An open-ended measure of behavior accompanied a closed-ended question measuring behavioral intention. However, it turned out that differences in answers to the open-ended measure were determined by their order. Earlier answers were longer and more accurate, whereas subsequent answers were significantly shorter and, therefore, represented less effort. This unforeseen order ef-

License Creative Commons Attribution 4.0 This article is licensed under the Creative Commons Attribution 4.0 (CC-BY 4.0) license, which allows you to copy, redistribute, transform and build upon the article and all its contents in any medium or format for any purpose, provided that appropriate credit is given.

© Fischer 2021

Journal of Trial & Error, 2 (1) 2021 https://doi.org/10.36850/e5 | 44
effect made the qualitative data unusable for studying within-subject differences. Between-subject differences and differences between rotated questions did not suffer from this order effect.

This paper is organized as follows. Firstly, the state of research on the use of open-ended questions in surveys and real-effort experimental designs is described. Secondly, the results of the study are presented. Problems with the approach used to measure behavior with open-ended questions are highlighted. Subsequently, these issues are discussed, and recommendations are made for future research.

It is not possible to measure actual behavior in surveys and stated behavior is not always valid. Real-effort designs use real tasks in experiments to imitate real-world behavior and measure behavior more validly. This study combines both approaches, surveys and real-effort designs, and uses open-ended questions to measure effort in answering behavior as a proxy for actual instead of stated behavior. However, an order effect appears when this strategy is used to measure behavior in within-subject designs and, therefore, results are misleading.

Open-ended questions in quantitative surveys as a measure of effort

Combining open-ended and closed-ended questions in a survey requires a pragmatic approach of method and data triangulation (Rossman & Wilson, 1985, p. 631). Quantitative and qualitative methods not only succeed one another, but are integrated within a single study. Due to this triangulation, "researchers are allowed to improve the accuracy of conclusions by relying on data from more than one method" (Rossman & Wilson, 1985, p. 632). Hence, the combination of qualitative and quantitative data is used here to offset weaknesses of the latter, draw on the strengths of the former, and enhance the validity of the findings (Bryman, 2006; Dewasiri et al., 2018, p. 105).

Several studies use this approach and compare answers to open-ended and closed-ended questions. Most of these studies report substantial discrepancies regarding answering behavior (Converse, 1984; Schwarz, 1999). For example, open-ended questions usually produce a more diverse set of answers. In a study reported by Schuman and Presser (1979), 60% of answers to an open question fell outside of the pre-coded answer categories of a related closed-ended question in the control group. Additionally, new types of responses occurred in a frequency that justified additional answer categories. In contrast, some answer options were less likely to be spontaneously produced in open-ended questions than they were to be chosen when explicitly offered as an option in closed questions (Schuman & Presser, 1979, p. 365). Similarly, in a study on the quantity of alcohol consumption, Greenfield et al. (2006) found that the answers on the quantity of consumed alcohol of around 30% of respondents differed significantly between answers on open-ended and closed questions. They found that the combination of both an open-ended and a closed measure of the quantity of alcohol consumption was a stronger predictor of alcohol-related consequences than the individual measures (Greenfield et al., 2006).

The aim of studies using open-ended and closed-ended questions in combination is usually to measure attitudes, feelings, or past behavior. However, self-reports, especially on past behavior, are a complex cognitive task that are highly context-dependent, and data are often seen as unreliable (Schwarz & Oyserman, 2001, p. 128). Some authors state that it is easier for participants to understand questions and recall relevant behavior when closed-ended questions are used (Schwarz & Oyserman, 2001, p. 131). In contrast, it is usually agreed that open-ended questions are more suited to asking about sensitive or stigmatizing information or when more in-depth information is needed (Friborg & Rosenvinge, 2013).

However, I argue that the advantages of closed-ended questions only apply to questions on past instead of future behavior, where more emphasis lies on "editing" the answer. Hence, it is more important to get an accurate answer than a correct recall of past behavior. Accordingly, Singer and Couper (2017) suggest using open-ended questions in quantitative surveys more often to "encourage more truthful answers [...] [and use them] as an indicator of response quality" (p. 3).

Real-effort experimental designs rely on the assumption that greater effort in fulfilling a task makes it more likely that participants actually perform what they have stated (Dutcher et al., 2015). This assumption is based on the costs connected to fulfilling a real-effort task: The higher the cost an individual is willing to invest in answering, the more likely it is that these costs will also be invested in performing the actual behavior. Results based on real-effort designs are more externally valid when "the cost function of that task shares important characteristics with the field task" (Dutcher et al., 2015, p. 3). In surveys and survey experiments, open-ended questions can be used to measure effort in answering a question and, ultimately, we can use this as a proxy to measure real instead of stated behavior. Hence, the effort is for the most part cognitive and creative, but typing answers might also cost physical effort (Charness & Greico, 2018, p. 75). Such a real-effort design might be a better estimator of actual behavior because "simply choosing a number may not capture the field environment and the psychological forces involved in putting forth actual effort" (Charness et al., 2018, p. 75). Greater effort can be operationalized with
an extensive, accurate, and fitting answer. This is based on different assumptions:

1. One dimension of effort is the length of time during which cognitive resources are used (Christensen-Szalanski, 1980). Accordingly, the length of a written answer is an approximator of the duration of the effort. Similarly, the length of answers to open-ended questions are often used as indicators of data quality (Galesic & Bosnjak, 2009, p. 350).

2. Accuracy and errors can also be used to capture effort (Charness et al., 2018, p. 78). Such a procedure was used by Gneezy and List (2006), who conducted an experiment in which participants had to enter library data into a database. They counted the number of errors in these entries as a measure of effort. Similarly, Andersen et al. (2018) registered an experiment in public administration research using this approach to compare work effort between public and private organizations. They asked participants to transcribe handwritten timesheets and checked their accuracy.

3. Furthermore, participants may not only be asked to transfer existing information into a database but also to be creative. Similarly, Charness et al. (2018) asked participants in an experiment to write a story about a predefined topic or by using specified words, and used this story as a measure for the real effort put forth in creative tasks. As it is difficult to rate creativity, the details provided in such an answer can be used to measure effort.

In such settings, participants' skills and abilities may strongly confound the results (Charness et al., 2018, p. 82). Longer answers, for example, may indicate that some participants can write quickly. Unfitting answers may indicate that some participants are unable to understand the question or articulate themselves. Charness et al. (2018) recommend using larger samples to capture this treatment effect. Furthermore, within-subjects designs might be useful to overcome these between-subjects differences.

**Putting a real-effort survey design into practice: The effect of rewards on knowledge-sharing behavior**

A survey experiment on knowledge-sharing behavior was conducted to analyze whether tangible or intangible incentives can foster that behavior. Knowledge sharing is the exchange of knowledge among individuals, teams, units, or organizations (Paulin & Suneson, 2012). It is the basis of and a subprocess in an organization's knowledge management. In this article, the term "knowledge sharing" describes the behavior of donating knowledge from one person to another or to a medium. Multiple determinants influence knowledge sharing. Among others, tangible and intangible rewards are considered to foster knowledge-sharing behavior. Tangible rewards are material incentives, such as financial bonuses, and usually enhance extrinsic motivation. In contrast, intangible rewards, such as praise from a colleague or supervisor, usually influences intrinsic motivation.

On the one hand, it has been shown that monitoring (e.g., controlling with performance measures) and rewards increase the knowledge-sharing activity of employees in an organization (Wang et al., 2011; Witherspoon et al., 2013). On the other hand, Bock and Kim (2002)’s results suggest that expected rewards do not affect knowledge sharing.

A 2x3 factorial survey experiment was designed to observe the within-subjects and between-subjects effects of the rewards offered. The research design was preregistered on the Open Science Framework (Fischer, 2018). Data were collected from German public employees in the core administration and health sector ($N = 623$) in 2018 using a self-administered questionnaire. As can be seen from Appendix A, most participants were female (61%). The mean age was 45 years, and participants had a tenure of 20 years on average.

Each participant was randomly assigned a set of three vignettes from a pool of six. Randomization was done after participants started to answer the survey, based on their respondent ID. The vignettes incorporated all independent variables (Appendix B). Each set contained vignettes on either explicit or implicit knowledge (between-subject design). The first vignette in each set was a baseline vignette, while the two following vignettes presented a tangible and an intangible reward for knowledge sharing in a randomly assigned order (within-subjects design).

Performance appraisals were used as the tangible reward (treatment) (participants were given this reminder: “You know that all shared information improves your performance appraisal”). Performance appraisals served as a proxy for later rewards because it seemed unrealistic in the public sector to offer bonuses or other tangible rewards directly based on a person's knowledge-sharing behavior. This decision should address the problem that vignette experiments are frequently criticized for being unrealistic and lacking external validity (Aguinis & Bradley, 2014, p. 361).

The intangible reward treatment was operationalized by offering explicit appreciation from co-workers without further illustrating how this appreciation would occur (participants were given this reminder: “You know, your co-workers appreciate your knowledge sharing”). A rather superficial description was chosen because every team shows appreciation in a different way (e.g., good team climate, respectful interactions) and the vignette was thought to be not too restrictive.

Knowledge-sharing behavior (KSB) was measured with an open-ended question as a proxy for real
behavior. The item reads: “If you decided to share your knowledge, please briefly describe how exactly you will share your knowledge.” Thus, participants were not asked to actually share their knowledge, but it was assumed that the cost function of answering the open-ended question on KSB was similar to the cost function of performing the described behavior in reality, for instance, when individuals share knowledge by writing an e-mail to a co-worker.

Additionally, knowledge-sharing intention (KSI) was measured in a closed-ended way to compare results to the verbatim responses. A measure of KSI was formed by taking the mean of two items, which were rated on 5-point Likert scales. The items read: “I will share this knowledge with my co-workers” and “I would like to share this knowledge with my co-workers.” The former item represents the future-oriented and behavioral part of intention and the latter represents the motivational part of intention. Both items were based on scales on KSI used by, for example, Bock and Kim (2002) and Lin and Suneson (2007). Both were adapted to the experimental context. The items correlated highly (depending on the vignette: $b = .66-.96$, $p < .001$) and were therefore compiled into one index for further analysis.

To code the open-ended answers, three dimensions were defined preliminarily, based on the literature, to rate the quality of the answer: the description’s length, accuracy, and fit with the vignette. Two dimensions were added inductively during the coding process. Firstly, some participants answered the question with few words and therefore did not score in the quantity dimension, but were considered separately from participants who did not answer because they at least put some effort into answering the question. Secondly, some participants answered very elaborately in terms of length and content, and this effort was assigned a bonus point. Each dimension was coded as a dummy variable and an additive index was formed, ranging from zero to five.

Coding was done by two independent raters. In­terrater reliability was calculated using Gwet’s Agreement Coefficient (AC; Gwet, 2014). Due to the fact that there was only partial agreement in the ratings of some dimensions (see Appendix C), the raters discussed differences in coding and specified coding rules. In a second step, both raters agreed on a rating. This “negotiated” rating was used for further analysis.

Results and discussion of issues emerging from the open-ended question approach

Table 1 and Table 2 give a short overview of the descriptive statistics for the dependent variables. Examples of verbatim answers to the open-ended question on KSB and their coding are provided in Appendix D.

The descriptive statistics already show that when open-ended questions (knowledge-sharing behavior) were presented later in the survey, there were more missing answers, whereas such a high rate of missing answers was not observed with the closed-ended question (knowledge-sharing intention). Furthermore, in the second and third open-ended questions, participants often referred to their previous answers (e.g., “see above”, “as just described”) or answered in a significantly shorter way. Means for knowledge-sharing behavior were, therefore, significantly smaller in the second and third vignettes than in the first, which did not measure knowledge-sharing intention.

Data were analyzed using Wilcoxon signed-rank

---

1 Want to thank Lisa-Sophia Preller for serving as the second coder of this study. Both raters were asked to use different coding strategies, to avoid effects of the coding procedure: (1) one rater coded the cases one by one in order to have an eye on the complete answer, (2) another rater coded each dimension across all cases one by one in order to have an eye on the comparability of the ratings. The fifth dimension (high effort in answering) was not coded by both raters but added afterward and coded by a single rater, therefore no interrater reliability was calculated here.
This study aimed at testing a more valid way to study human behavior in surveys. I tried to overcome the limited validity of self-reported behavior in surveys. I wanted to adapt experimental real-effort tasks to a survey design by using open-ended questions and analyzing the effort participants put into answering them.

Table 3  Correlation of KSI and KSB. Note. N baseline = 547, N intang. Reward = 501, N tang. Reward = 500.

<table>
<thead>
<tr>
<th></th>
<th>KSI without treatment</th>
<th>KSI appreciation</th>
<th>KSI achievement</th>
</tr>
</thead>
<tbody>
<tr>
<td>KSI without treatment</td>
<td>1.20**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KSB without treatment</td>
<td></td>
<td>0.052</td>
<td></td>
</tr>
<tr>
<td>KSB intang. reward</td>
<td></td>
<td>0.083</td>
<td></td>
</tr>
<tr>
<td>KSB tang. reward</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

tests (within-analysis) and Wilcoxon rank-sum tests (between-analysis) because the data were not normally distributed. Using the closed-ended measure of knowledge-sharing intention as the dependent variable led to a slightly greater intention to share when explicit knowledge was shared and a benefit was offered (M = 4.31, SD = 0.80) than without an offered benefit (M = 4.23, SD = 0.79; z = 3.23, p < .001, d = 0.09, bootstrapped 95% CIs [0.03, 0.16]).

However, using knowledge-sharing behavior as the dependent variable led to contradictory results, which might have been caused by a methodological problem. This issue was already identified while coding the data and inspecting the descriptive statistics: There was an order effect leading to longer answers on average for the first presented vignette compared to subsequent vignettes, independent of treatment. This result fits with the literature, as, for example, Galesic & Bosnjak (2009, p. 357) showed that open-ended questions asked later in a questionnaire were associated with shorter answers.

Accordingly, correlations between the stated knowledge-sharing intention and the coded answers related to knowledge-sharing behavior were not pronounced (Table 3). While knowledge-sharing intention and behavior were significantly but still only moderately correlated without an incentive treatment, the correlation was weaker and insignificant when incentives were induced. As the baseline vignette was always presented as the first vignette, this result supports the assumption of an order effect.

Analyzing knowledge-sharing behavior as the dependent variable, Wilcoxon signed-rank tests indicate significant differences between the treatment groups and the control group (explicit knowledge: appreciation vs. control group: z = -5.86, p < .01; achievement vs. control group: z = -4.40, p < .01; implicit knowledge: appreciation vs. control group: z = -6.83, p < .01; achievement vs. control group: z = -4.89, p < .001). Contradicting the hypothesized relationships, the data show that knowledge-sharing behavior is significantly higher in the control group, and thus in the answer to the first vignette (baseline without treatment).

However, the comparison of the two treatments might not suffer from this methodological problem as they were presented in a randomly rotated order, thus either as the second or third vignette. Comparing the two treatments regardless of the kind of knowledge shared yields a significant difference. The tangible reward treatment (M = 2.47, SD = 0.74) triggered more knowledge-sharing behavior than the intangible reward treatment (M = 2.38, SD = 0.69; z = 2.40, p = .017, d = 0.10, bootstrapped 95% CIs [0.02, 0.18]).

These results show that open-ended measures in surveys are exposed to strong order and fatigue effects. Open-ended questions presented later in the survey result in more missing answers, shorter answers, or answers referring to previous statements. Therefore, behavior reflecting minimal instead of satisfying answers was observed. Hence, these results could not be used in this study to analyze within-subject differences. However, the open-ended measure could be used as a proxy for real behavior in analyzing between-subjects differences and within-subjects comparisons between rotated vignettes (tangible and intangible rewards).

Conclusion: Recommendations for future research

Using open-ended questions to measure effort as a proxy for real instead of stated behavior was not useful in a within-subjects research design. Due to an order effect, open answers could not serve as a reliable estimator of the likelihood of actual behavior. However, an advantage of verbatim answers is that they can still be used to identify behavioral patterns and modes of KSB. In further research, this strength of qualitative data should be taken into account more instead of merely quantifying qualitative data.

While this approach toward operationalizing behavior in surveys might not be useful in within-subject designs, it can be used in between-subject comparisons if participants are asked to answer to a single vignette. However, as participants’ knowledge and competencies might influence answers on questions designed as real-effort tasks, attention must be paid to the sampling strategy when solely using between-subjects designs. When a baseline answer for an individual is missing or not taken into account, which is the idea behind a within-subjects design,
an individual cannot be matched to their own standard. The order effect that occurred in this study was not expected by the author in that magnitude. It is known in the literature that, especially regarding open-ended questions, questions asked later in a questionnaire are associated with fatigue and shorter answers (Galesic & Bosnjak, 2009). However, it was expected that questions directly succeeding each other within a survey would not suffer that strongly from such an order effect. Additionally, the survey was rather short and took participants between 10 and 15 minutes to complete. The author expected that fatigue effects were more likely to occur in longer surveys. Also, the study’s pretest with students and public sector professionals did not reveal such order and fatigue effects. Hence, this outcome might be due to characteristics of participants from online panels, who might be motivated to answer a questionnaire in an efficient and quick manner.

Due to the state of research on order effects in surveys, the treatment vignettes were randomized. However, I decided to exclude the baseline vignette from this randomization to avoid misunderstandings among the participants and to ensure the vignettes and questions were presented in a logical order. Future studies should, however, be aware of the occurrence of order effects in such an experimental setting, especially when data are collected with a sample of merely extrinsically motivated participants. To prevent order effects, between-subjects designs could be used instead of relying on within-subjects analyses. Apart from that, a combination of open-ended and closed questions allows one to detect order effects and should be included in case of uncertainty. Using both measures together also gives the option to analyze at least a part of the dependent variable and, therefore, minimizes risks of failure based on poor research designs.

Further research in this area may want to further consider the above-mentioned order and fatigue effects in answering open questions in survey experiments and how these might be prevented. Studies could, for example, experiment with rotating all vignettes instead of only rotating those with treatments to avoid an order effect. However, if the baseline vignette is not presented first, participants may have difficulties understanding the vignettes.

Furthermore, open-ended questions might be distributed throughout the survey instead of directly after one another in order to prevent participants from referring to earlier answers. However, questions asked in between may induce other treatments and thereby affect the answers. By reporting the failure of the research strategy used in this study, this article intends to serve as a step in further testing the possibility of using open-ended questions as valid measures of effort in surveys and survey experiments.

References


Andersen, S., James, O., & Jilke, S. (2018). Does work effort for public versus private organizations differ? Evidence from an online work task experiment: Pre-registration [Retrieved from]: https://www.socialsciencescrereg.org/trials/3361


Appendix A

Table 4  Sample Description

<table>
<thead>
<tr>
<th>Variable</th>
<th>N</th>
<th>M</th>
<th>SD</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>age</td>
<td>609</td>
<td>45.10</td>
<td>10.29</td>
<td>22</td>
<td>80</td>
</tr>
<tr>
<td>female</td>
<td>618</td>
<td>0.61</td>
<td>0.49</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>supervisor</td>
<td>615</td>
<td>0.29</td>
<td>0.45</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>tenure</td>
<td>615</td>
<td>20.20</td>
<td>11.23</td>
<td>0</td>
<td>47</td>
</tr>
<tr>
<td>fixed-term empolyym.</td>
<td>613</td>
<td>0.05</td>
<td>0.23</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Appendix B

Table 5  Vignette plan

<table>
<thead>
<tr>
<th></th>
<th>Explicit knowledge</th>
<th>Implicit knowledge</th>
</tr>
</thead>
<tbody>
<tr>
<td>Without treatment</td>
<td><strong>Vignette 1</strong>  During a daily routine at your workplace, you gathered information from several sources. This could also improve the work of your co-workers. Please decide whether you will share this information with your co-workers. (N = 319)</td>
<td><strong>Vignette 4</strong>  During a daily routine at your workplace, you had an experience that improved your work process. This experience could also help your co-workers. Please decide whether you will share your knowledge with your co-workers. (N = 310)</td>
</tr>
<tr>
<td>Intangible reward</td>
<td><strong>Vignette 2</strong>  During a daily routine at your workplace, you gathered information from several sources. This could also improve the work of your co-workers. You know that your co-workers appreciate your knowledge sharing. Please decide whether you will share this information with your co-workers. (N = 319)</td>
<td><strong>Vignette 5</strong>  During a daily routine at your workplace, you had an experience that improved your work process. This experience could also help your co-workers. You know that your co-workers appreciate your knowledge sharing. Please decide whether you will share your knowledge with your co-workers. (N = 307)</td>
</tr>
<tr>
<td>Tangible reward</td>
<td><strong>Vignette 3</strong>  During a daily routine at your workplace, you gathered information from several sources. This could also improve the work of your co-workers. You know that all shared information improves your performance appraisal. Please decide whether you will share this information with your co-workers. (N = 317)</td>
<td><strong>Vignette 6</strong>  During a daily routine at your workplace, you had an experience that improved your work process. This experience could also help your co-workers. You know that all shared information improves your performance appraisal. Please decide whether you will share your knowledge with your co-workers. (N = 309)</td>
</tr>
</tbody>
</table>
Appendix C

Table 6  Initial Interrater Reliability Open Answers on Knowledge-Sharing Behavior (KSB)

<table>
<thead>
<tr>
<th>Vignette order</th>
<th>Gwet’s AC</th>
<th>Cohen’s Kappa</th>
<th>Extent of Agreement</th>
</tr>
</thead>
<tbody>
<tr>
<td>Vignette 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First vignette (v28)</td>
<td>.9856***</td>
<td>.7430***</td>
<td>substantial</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.4674***</td>
<td>.4586***</td>
<td>moderate</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.7522***</td>
<td>.6225***</td>
<td>moderate</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.9090***</td>
<td>.5805***</td>
<td>moderate</td>
</tr>
<tr>
<td>Vignette 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First vignette (v34)</td>
<td>1.0000***</td>
<td>1.0000***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.4943***</td>
<td>.3995***</td>
<td>fair</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.8426***</td>
<td>.6480***</td>
<td>substantial</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.8726***</td>
<td>.5010***</td>
<td>perfect</td>
</tr>
<tr>
<td>Vignette 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second vignette (v30)</td>
<td>.9753***</td>
<td>.8308***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.5041***</td>
<td>.3075***</td>
<td>fair</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.7646***</td>
<td>.3530***</td>
<td>substantial</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.9170***</td>
<td>.7075***</td>
<td>perfect</td>
</tr>
<tr>
<td>Vignette 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third vignette (v38)</td>
<td>.9918***</td>
<td>.9332***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.7475***</td>
<td>.4645***</td>
<td>substantial</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.9087***</td>
<td>.3802*</td>
<td>substantial</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.7888***</td>
<td>.4652***</td>
<td>substantial</td>
</tr>
<tr>
<td>Vignette 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third vignette (v32)</td>
<td>.9642***</td>
<td>.8022***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.5584***</td>
<td>.3438***</td>
<td>fair</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.7813***</td>
<td>.4196***</td>
<td>substantial</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.8627***</td>
<td>.6546***</td>
<td>substantial</td>
</tr>
<tr>
<td>Vignette 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second vignette (v36)</td>
<td>1.0000 ***</td>
<td>1.0000 ***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.5302***</td>
<td>.2667***</td>
<td>moderate</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.7887***</td>
<td>.3119**</td>
<td>substantial</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.8417***</td>
<td>.5005***</td>
<td>substantial</td>
</tr>
<tr>
<td>Vignette 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First vignette (v16)</td>
<td>1.0000 ***</td>
<td>1.0000 ***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.6097***</td>
<td>.5835***</td>
<td>moderate</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.7829***</td>
<td>.6836***</td>
<td>substantial</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.9360***</td>
<td>.6932***</td>
<td>perfect</td>
</tr>
<tr>
<td>Vignette 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>First vignette (v22)</td>
<td>1.0000***</td>
<td>1.0000***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.5520***</td>
<td>.5479***</td>
<td>moderate</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.7015***</td>
<td>.5903***</td>
<td>moderate</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.9149***</td>
<td>.6533***</td>
<td>perfect</td>
</tr>
<tr>
<td>Vignette 5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second vignette (v18)</td>
<td>.9830***</td>
<td>.7921***</td>
<td>perfect</td>
</tr>
<tr>
<td>answer provided</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>quantity of the answer</td>
<td>.6968***</td>
<td>.5625***</td>
<td>moderate</td>
</tr>
<tr>
<td>accuracy of the answer</td>
<td>.7793***</td>
<td>.4512***</td>
<td>substantial</td>
</tr>
<tr>
<td>fit with vignette</td>
<td>.8637***</td>
<td>.5912***</td>
<td>substantial</td>
</tr>
</tbody>
</table>

**p < 0.001, ***p < 0.01, *p < 0.05**
<table>
<thead>
<tr>
<th>Vignette</th>
<th>Type of Vignette</th>
<th>Answer Provided</th>
<th>Quantity of the Answer</th>
<th>Accuracy of the Answer</th>
<th>Fit with Vignette</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vignette 5</strong></td>
<td>Third vignette v26</td>
<td>.9727***</td>
<td>.6613***</td>
<td>.8328***</td>
<td>.8888***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>moderate</td>
</tr>
<tr>
<td><strong>Vignette 6</strong></td>
<td>Third vignette v20</td>
<td>1.0000***</td>
<td>.6357***</td>
<td>.5394***</td>
<td>.8653***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>substantial</td>
</tr>
<tr>
<td><strong>Vignette 6</strong></td>
<td>Second vignette v24</td>
<td>1.0000***</td>
<td>.6873***</td>
<td>.7614***</td>
<td>.9132***</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>moderate</td>
</tr>
</tbody>
</table>

**p < 0.001, **p < 0.01, *p < 0.05
Appendix D

Selected open answers on knowledge-sharing behavior (KSB)

Note: Own translation of German answers

Core public administration

KSB-Index = 1 (does not fit the question)

- Everybody benefits from it.
- In my job, information and collegiality are the foundations of our conduct. It wouldn’t work any other way.
- It pushes the team forward.

KSB-Index = 2

- In a one-on-one conversation
- Very precise
- 1: jour fixe 2: mailing list

KSB-Index = 3

- New information is transferred personally.
- A written report or e-mail
- Via e-mail or during coffee break time

KSB-Index = 4

- A group e-mail addressed to the department to inform about the outcome of the issue.
- Orally in conversation, and in some circumstances by looking at the file.
- I announce this experience within the team, so everyone can decide for himself or herself whether they can or want to use this information.

KSB-Index = 5

- If it is something very interesting in my point of view, I would forward an e-mail with the necessary documents to my colleagues. If it is an issue under the category “...just so you have heard about it...”, I would seek direct talks and elaborate on the case.
- An electronic submission such as an appendix via e-mail, or a written submission such as copies distributed or an offer to approach me if required, depending on relevance.
- I prepare to address the issue in the next meeting and to share the new information with my co-workers.

Health sector

KSB-Index = 1 (does not fit the question)

- Leads to shorter working hours
- Because the simplification of work is good
- Improvement within the team

KSB-Index = 2

- In a team meeting
- In conversations
- By demonstration

KSB-Index = 3

- Not sure, it depends on the co-workers and their mood.
- I show them my approach.
- I address it briefly within the team and wait to see if somebody is interested.

KSB-Index = 4

- In a conversation/team meeting I make suggestions to restructure work processes and ask for opinions of others, after which the team has to decide.
- I would post it in our company network or speak about it in a team meeting.
- I gather co-workers, who work in a similar task field, and tell them about my experiences.

KSB-Index = 5

- I will share this experience in personal conversation and will passionately talk about my findings. Hopefully, they will keep this knowledge in mind when they are doing their job.
- I would share this with my co-workers in the first situation possible and not wait for the next team meeting. I always get excited when I discover something new that simplifies my work and I want to share this joy with others instantly.
- I inform my manager and ask for permission to change something actively. When it comes to random things, which don't require official procedure instructions, I would train the new co-workers directly, which is the better way of doing things.
The Gloria Adherence Subproject: Problems and Randomization Mistakes

Linda Hartman1,2, Marc R. Kok3, Esmeralda Molenaar4, Ed N. Griep5, Jaap M. Van Laar6, Jan Maarten Van Woerkom7, Cornelia F. Allaart8, Hennie G. Raterman9, Yvonne P. M. Ruiterman10, Marieke J. H. Voshaar11, João Redol12, Rui M. A. Pinto13, L. Thomas Klausch2, Willem F. Lems1, Maarten Boers1,2

Medication adherence, which is the extent to which patients take their medication as prescribed, is essential in treating chronic inflammatory diseases such as rheumatoid arthritis (RA). Therefore, we nested a subproject in the two-year multicenter Glucocorticoid Low-dose Outcome in Rheumatoid Arthritis (GLORIA) trial to add a low-dose prednisolone (5 mg/day) or placebo to the standard care in older people (65+ years) with RA. Adherence was measured with an electronic monitoring cap that recorded bottle openings in all patients. In the subproject, we performed an adherence intervention with an advanced cap that could communicate with an application on the smart device via Bluetooth. We randomized patients with a smart device to receive or not to receive adherence reminders on the smart device for three months. Multiple problems emerged that precluded an answer to the research question: sample size (overly optimistic estimates of older patients with a smart device), logistic issues (availability of smartcaps, data extraction), randomization and treatment allocation errors (despite training of personnel), and low quality of the data in the intervention group (hardware failure, discovered too late because data was read in batches). For future trials planning to include a subproject, we recommend keeping it simple, starting with a field test before the actual study starts, and monitoring data from the beginning of the study.

Keywords medication adherence, electronic caps, rheumatoid arthritis, nested trial, randomization

Medication adherence, which is the extent to which patients take their medication as prescribed (Brown & Bussell, 2011), is essential to treat chronic inflammatory diseases, such as rheumatoid arthritis (RA; Jungst et al., 2019). Almost half of patients with chronic diseases are non-adherent (Jungst et al., 2019; Kini, 2018). Non-adherence substantially increases the risk of a disease flare in RA (Jungst et al., 2019). In the two-year multicenter Glucocorticoid Low-dose Outcome in Rheumatoid Arthritis (GLORIA) trial, a low-dose prednisolone (5 mg/day) or placebo was added to the standard care in older people (65+ years) with RA. We measured adherence with electronic monitoring and pill count in all patients. In a subproject nested within the trial, we assessed the efficacy of an adherence intervention for the patients with a smart device.

Methods

The eligibility criteria for the subproject were to: be in possession of a smart device (smartphone or tablet) and be an active participant in the main GLORIA trial for at least three months. We planned to include patients with a smart device from all seven participating countries of the main trial because a pilot survey indicated that a sufficient number of patients (20%) possessed a smart device in all countries. Based on this feasibility assessment, we expected to include 50 patients. Before the training of the study personnel, it appeared that only Dutch patients had smart devices. Therefore, the training of the study personnel was only provided in the Netherlands.

Training of study personnel

The study personnel of all participating Dutch sites received an e-mail with information about the objective of the subproject, eligibility criteria, signing of the informed consent in duplicate, and an explanation about the randomization step. In addition, the study personnel received a manual that explained the subproject procedures (assembling of the cap, installation of application) that were needed before the patient could start with the subproject. Finally, the subproject and the related procedures were explained and demonstrated during the Dutch GLORIA researchers meeting, and an English instructional movie was sent to all participating centers. One of
Take-home Message

It proved difficult to perform a subproject nested within a large investigator-initiated trial, although we expected it to be easy. We recommend keeping a nested subproject simple, maintaining active contact with study centers, and monitoring data from the beginning to resolve emerging problems immediately.

The activated app was connected to the smart adherence cap of the medication bottle via Bluetooth. For patients, no additional actions were required. The patient only had to open the medication bottle once per day, preferably in the morning. Both the regular adherence caps in the control group and the smart adherence caps in the intervention group recorded each opening and closing event of the medication bottle. The smart adherence cap repeatedly attempted to open a connection with the GLORIA app via Bluetooth to report whether the bottle had been opened. If the bottle was not opened between 08:00 a.m. and 02:00 p.m., the app posted a notification on the opening screen to remind the patient to open the medication bottle. The patients in the control group did not receive an activated smart cap nor the reminder app; they were instructed to open the medication bottle once per day, as in the main trial. At the end of the three-month period, the bottle was returned, and the remaining capsules were counted. For patients that started the trial at six months or later, the return of the bottle was delayed by three months, because clinic visits occurred every six months. The adherence of both groups based on returned pills was calculated as follows: If the number of pills dispensed is D, the treatment period in number of days is P, and the number of pills returned is R, medication adherence (%) is calculated as: \((D-R)/P\times100\). The adherence of the two groups was compared. Social desirability could have influenced the adherence measurements with the caps and the number of returned pills. Patients might have returned the expected number of pills or opened the medication bottle as would be expected. However, we assume that most patients were as adherent as non-trial patients, because it would be difficult to maintain social desirability during the full two years of the trial.

Data registration and extraction
Apart from the bottle closing and opening events, which were always registered if the medication bottle was opened, the smart adherence caps also registered “pairing success” events, “no communication” events, “keep alive” events and “reset” events. The “pairing success” events guaranteed that the pairing of the adherence cap with the app before the study started was successful. Non-registration of this event indicated that the pairing of the cap and app was unsuccessful and that the cap was not able to communicate with the smart device during the entire three months of the study (reported by “no communication” events). Therefore, caps with no “pairing success” event at the start of the study were excluded.

Opening events
“No communication” events were registered at 02:00 p.m. if the smart adherence cap had been unable
to communicate with the app on the smart device between 08:00 a.m. and 02:00 p.m. on a given day. The time window from 08:00 a.m. until 02:00 p.m. was chosen to match the advice to take prednisolone in the morning. The registration of "no communication" events was independent of the successful or unsuccessful pairing between the app and the cap. The "keep alive" event was used to check if the cap still worked and if the battery was not empty. "Keep alive" events were registered one or more times per morning, indicating that the cap was functioning even if a patient did not open the medication bottle. A "reset" event was registered whenever a reset on the system was detected. This could be related to battery power being too low, or bad contacts on the battery power supply caused by faulty cables due to falls or abrupt movements of the cap. If more than one "reset" event was registered, the cap was excluded. The patients returned the regular caps and adherence caps to the research nurse, and the caps of each site were sent to the owner of the caps (BeyonDevices LDA, Sobral de Monte Agraço, Portugal). The data of the caps were read in batches and were uploaded in data files. The data files contained the registration of the events described above. The app generated a reminder message if the cap did not transmit an opening event between 08:00 a.m. and 02:00 p.m., or if no communication was established on that day (registered in the cap).

Discussion

Problems in the preparation and execution of the trial
We had multiple misfortunes related to the preparations of the subproject, technical issues with the app and smart adherence cap, and the randomization step. Therefore, it was not possible to test the effectiveness of adherence reminders delivered via the smart device during the three month trial period. We want to discuss and reflect on the problems and misfortunes that prevented us from answering our research question. All aspects that influenced the success of the subproject were discussed, and we reflected on how each aspect contributed to the subproject's failure. The discussions and reflections are mainly related to the researchers from the Netherlands, because only Dutch patients entered the subproject. The discussions were supported by the scientific advisory committee, which was set up especially for the GLORIA trial and by the developers of the adherence caps and the GLORIA app.

Preparations
We had a contract with a company that would deliver the adherence caps and the application for the subproject, but this company went bankrupt a few months before the study would start. We had to choose another company in a very short time frame because the main trial was almost ready to start. This choice was based on expertise, the projected time for development and implementation of the smart device reminder app, and an offer with limited costs because the budget of this investigator-initiated study was limited. Unfortunately, the chosen company needed more time than expected to develop suitable smart caps. In total, this caused a delay of 1.5 years in the delivery of the smart caps.

Sample size
The number of potential patients for the subproject proved much lower than expected. From initial feasibility assessments among the trial centers, it appeared that around 20% of the elderly patients possessed a smart device. Assuming that patients with a smart device are technically interested, we expected that about 80% of them would be willing to participate. Thus, with a sample size of 450 patients in the main study, we expected around 72 patients in the subproject. As it turned out, smart device owners in this senior population were only located in the Netherlands. This resulted in a significant waste of paid and unpaid work hours: The app and instruction manual were translated from English into seven localized versions with the help of the patient panel of the main trial. To include enough patients in the subproject, we amended the protocol so that patients could start with the subproject at any moment during the trial instead of only at the three-months visit.

Complexities for the participating sites
The changed inclusion moment for the subproject (as described above) made the inclusion procedure more complex for the participating sites. Another potential source of confusion was that we now had two types of similar-looking caps: the regular adherence caps used by all patients throughout the whole main trial, and the smart adherence caps only intended for patients in the intervention group of the subproject. The only difference between the caps was the color of the magnetic label on the cap, which had to be removed to activate the cap. It proved difficult for the research nurses to remember the differences and objectives of the regular and smart adherence caps. Also, despite the instructions outlined above, for some study personnel, it was not clear that patients had to be separately randomized for participation in the subproject and that the preparations before participation were different for the patients in the intervention and control group. For the smart adherence caps belonging to the intervention group, several manual actions were needed before the cap could be used. This involved a lot of work if a patient was randomized to the intervention group compared to control patients, where no additional actions were required.
Randomization and treatment allocation mistakes
These complexities resulted in mistakes in the randomization and treatment allocation, despite training of the sites. A total of 126 patients in the main trial were assessed for eligibility to participate in the subproject (Figure 1). Most patients were not eligible because they were not in possession of a smart device. From the non-eligible patients, three were nevertheless randomized for the subproject by mistake. The adherence data of these patients could not be used because the patients did not meet the eligibility criteria, did not give informed consent to use their data, or both.

In addition, randomization and treatment allocation failed for many patients because of mistakes at the participating sites and technical problems related to the smart adherence cap. A total of 23 patients received the intervention (smart cap). This included 16 patients that were randomized and properly allocated to the treatment, three patients originally randomized to the control group (contamination), and four patients who received the intervention without being randomized. A total of 19 patients received the control intervention (no reminder). This included 15 patients that were randomized to the control group and were properly allocated to the treatment, one patient originally randomized to the intervention (contamination, the smart adherence cap failed to activate), and three patients who received the control intervention but were not randomized.

The reasons for the randomization and treatment allocation mistakes are unknown. However, we have some suggestions for changes to the randomization procedure, which might decrease the chance of mistakes. In our study, patients could have been randomized easily to the subproject by first having the patient answer a few questions in a conversation (among others about the possession of a smart device), followed by one click on a button in the eCRF. Entry of ineligible patients could probably have been prevented if the answers to the questions needed to be entered correctly on the screen before randomization and allocation could proceed. The problem of contamination, i.e., patients randomized to one group but receiving the intervention of the other group, was probably caused by confusion among the research nurses performing the allocation procedure. It is likely that the explanation given to the research nurses about the subproject was insufficient. On-site training could have been better, but this is time-consuming and costly.

Functioning of smart adherence caps in the intervention group
Upon delivery, we found the smart adherence caps had technical limitations, but there was no possibility to resolve this limitation by the company. The medication bottle and smart device had to be close to each other to connect and exchange information about the opening events. If the medication bottle and smart device were more than a few meters apart, they did not connect and the patient received a reminder to open the bottle as a consequence. Of the 23 smart adherence caps applied, only two caps functioned as expected. The other adherence caps had technical errors from the start of the subproject or within 1-4 weeks after the start, or the caps were not returned (Table 1). The malfunctioning of the smart adherence caps in the intervention group was also reflected in the satisfaction rates reported by 18 of the 23 patients. On a scale of 0-10, the mean reported satisfaction rate was 5.6 (range 0-10). Multiple patients gave a low satisfaction rate because they received a reminder despite taking their medication (i.e., opened the bottle). On the other hand, six patients reported a grade of 8 or higher. The malfunctioning of the caps could explain these high rates. The caps did not or rarely communicated with the app and, therefore, the app never sent reminders. This meant that these patients did not receive unnecessary reminders when they opened the bottle.

Original Purpose
The ongoing two-year Glucocorticoid Low-dose Outcome in Rheumatoid Arthritis (GLORIA) trial examines the addition of daily low-dose (5 mg) prednisolone or placebo to the standard care in elderly patients (65+ years) with rheumatoid arthritis; adherence is measured with electronic caps and counting returned pills. A subproject was nested within the main GLORIA trial, limited to patients with a smart device who had taken the study treatment for at least three months. The objective was to test the effectiveness of adherence reminders delivered to the smart device through Bluetooth by a special cap during a three-month period. Eligible patients were randomized to the intervention group (receiving reminders) or the control group (no reminders). In the intervention group, an application was loaded onto the smart device that communicated with the adherence-monitoring device loaded into the cap of the drug bottle. The app sent a reminder message if the patient had forgotten to take the study medication. The hypothesis was that after three months, patients in the intervention group would be more adherent compared to the patients in the control group. Multiple problems prevented us from answering our research question. This paper reflects on these problems.

The research group was a part of the GLORIA consortium and was comprised of rheumatologists and researchers from the Netherlands, Germany, Italy, and Portugal, and the owner of the caps and the pharmaceutical manufacturer of the study medication from Portugal.
Table 1  Functioning of the smart adherence caps in the intervention group of the GLORIA subproject.

<table>
<thead>
<tr>
<th>Functioning of smart adherence cap</th>
<th>Number of patients (% of patients) (n = 23)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good functioning</td>
<td></td>
</tr>
<tr>
<td>full 12 weeks</td>
<td>2 (9)</td>
</tr>
<tr>
<td>6 weeks</td>
<td>1 (4)</td>
</tr>
<tr>
<td>3-4 weeks</td>
<td>1 (4)</td>
</tr>
<tr>
<td>2-3 weeks</td>
<td>1 (4)</td>
</tr>
<tr>
<td>1-2 weeks</td>
<td>2 (9)</td>
</tr>
<tr>
<td>&lt; 1 week</td>
<td>9 (39)</td>
</tr>
<tr>
<td>Pairing error (no pairing success event registered)</td>
<td>2 (9)</td>
</tr>
<tr>
<td>Cap technical errors (reset event registered)</td>
<td>1 (4)</td>
</tr>
<tr>
<td>Cap not returned</td>
<td>4 (17)</td>
</tr>
</tbody>
</table>

Data extraction and interpretation
The data extraction and interpretation had limitations. The data were read in batches when the study was almost finished because it was cheaper to send the caps in batches instead of each single cap from sites in the Netherlands to the company in Portugal. This prevented early detection of errors and malfunctioning. The data interpretation was a complicated process. It was difficult to interpret the different events described in the methods section*, and to unravel when reminders were sent by the app. This increased the risk of errors in the data interpretation. It would have been easier if the data file had only the opening and closing events, and the exact moments when reminders were sent. Additionally, we had to link the adherence data to the patient data in the eCRF. The kit number belonging to the patient had to be looked up in the eCRF, which also increased the risk of errors.

Conclusion
We were unable to answer our research question because of a multitude of problems. Even with the limited sample size and randomization and allocation errors, we were initially hopeful that we could compare the adherence of the two groups as a controlled trial instead of a randomized controlled trial. However, the quality of the data in the intervention group was unusable due to almost universal failure of the cap-app dyad. It turned out that it was difficult to perform a subproject nested within a large pragmatic trial, although we expected it to be easy. For future trials that test new devices or applications, we strongly recommend a clear plan regarding the requirements. For example, we recommend conducting preliminary field tests to resolve technical problems of the materials, optimize study procedures, and ensure that the plans are clear for all involved parties. The study monitor should also be in active contact with the study centers when the first patients start, to resolve emerging problems in real time and to emphasize the aspects that will help the study perform well. In addition, data stored in the device should be read during rather than after the study. It is also important to be aware that most seemingly “simple” studies are typically complex due to issues with people in different roles, technology, and processes.

References

Funding
The GLORIA project was funded by the European Union’s Horizon 2020 Research and Innovation Programme under the topic ‘Personalizing Health and Care’ (grant number 634886).

Acknowledgments
J. W. J. Bijlsma, R. Christensen, Y. M. Smulders and S. H. Ralston are members of the scientific advisory committee of the GLORIA trial

Appendix
Figure 1  Flow diagram
The Unsuccessful Self-treatment of a Case of ‘Writer’s Block’: A Replication in Science Education

Georgios Ampatzidis

1Hellenic Open University, Parodos Aristotelous 18, 26335, Patras, Greece

Received 28 August, 2021
Accepted 8 October, 2021
Published 24 November, 2021
Issued 3 March, 2022

Correspondence ampatzidis.georgios@ac.eap.gr

License This article is licensed under the Creative Commons Attribution 4.0 (CC BY 4.0) license, which allows you to copy, redistribute, transform and build upon the article and all its contents in any medium or format for any purpose, provided that appropriate credit is given.

© Ampatzidis 2021
References


Classical Conditioning for Pain: The Development of a Customized Single-Case Experimental Design

Tamal Kumar De¹, Victoria J. Madden¹, Johan W. S. Vlaeyen¹, Patrick Onghena¹

Single-case experiments are increasingly popular in the behavioral sciences. Due to their flexibility, single-case designs can be customized to test a variety of experimental hypotheses. We were interested in using a single-case experimental approach to test whether pain thresholds can be influenced by Pavlovian classical conditioning. Following the example of earlier studies into this topic, we planned to measure whether participants would more frequently report specific electrocutaneous stimuli as painful when they were presented with specific vibrotactile stimuli that had previously been associated with painful electrocutaneous stimuli. First, we decided on a mean difference effect size measure derived from the Sensation and Pain Rating Scale ratings for the electrocutaneous stimuli provided by the participants. Next, we discussed several possible single-case designs and evaluated their benefits and shortcomings. Then, we ran pilot tests with a few participants based on the possible single-case designs. We also conducted a simulation study to estimate the power of a randomization test to test our hypothesis using different values for effect size, number of participants, and number of measurements. Finally, we decided on a sequentially replicated AB phase design with 30 participants based on the results from the pilot tests and the power study. We plan to implement this single-case design in a future experiment to test our hypothesis.

Keywords single-case, classical conditioning, pain, randomization test, statistical power

Classical Conditioning

Persistent pain—pain that is still felt after bodily tissue has healed—is a major healthcare problem that is poorly understood and, consequently, difficult to treat effectively. While there are key models that go some way towards explaining how pain can persist without active tissue damage (Vlaeyen & Linton, 2012, 2000; Woolf, 1983, 2011), certain pain presentations remain unexplained. The imprecision hypothesis was proposed to address these unexplained presentations, and is founded on the idea that pain can be modulated by pain-associated cues—specifically, via “classical conditioning” (Moseley & Vlaeyen, 2015).

Our study design builds on two previous experiments (Madden et al., 2016; Traxler et al., 2019), which followed a Pavlovian or “classical” conditioning design (Pavlov, 1928) to test whether neutral but pain-associated cues can bias a participant’s decision about whether a stimulus is painful or non-painful. In the differential classical conditioning design used here, one neutral cue is paired with a painful electrocutaneous stimulus, while another neutral cue is paired with a non-painful electrocutaneous stimulus. It is expected that participants form associations between each cue and painfulness, such that subsequent perception of an ambiguous electrical stimulus is modified, dependent on the simultaneous presentation of one of the neutral cues.

We set out to design an SCE to test this hypothesis. Due to the before (conditioning) and after (conditioning) structure of multiple observations required for each participant, SCE designs are perfectly suited for testing this hypothesis.

Single-Case Experiments

SCEs are experiments in which the effect of manipulating an independent variable is observed in a single entity (Barlow et al., 2009; Kazdin, 2011). Variables of interest in the entity, which is often a single participant, are measured repeatedly over a period of time with the purpose of establishing a causal relationship between the independent variable, commonly referred to as the treatment condition, and the observed entity. SCEs are increasingly popular in behavioral and educational sciences due to their flexibility, low cost, and focus on effects of the intervention in individual participants (Heyvaert & Onghena, 2014b). SCEs may be preferable to group-based designs, where the existence or magnitude of the effect within an individual is more relevant than
We wanted to design a single-case experiment to test whether classical conditioning can influence pain thresholds. We settled on a customized AB phase design for which we ran pilot tests and a power study. The results indicated that this single-case design can be used to test our hypothesis.

Advantages of SCEs for Classical Conditioning
SCEs are well suited to research questions about within-individual processes. Although group designs are historically respected, they are typically analyzed in a way that aggregates data from all individuals to estimate an effect at the group level—an effect that may be non-existent in any of the individuals within the group (Molenaar & Campbell, 2009). This aggregation approach obscures the true, within-individual effect, and the between-individual variability of that effect (which, itself, is typically worthy of attention; Fisher et al., 2018). Of course, it is possible to consider within-individual changes in some group designs, but an SCE offers a superior approach to achieving a sufficiently powered (>80%) examination of within-subject effects (Ferron & Onghena, 1996; Ferron & Ware, 1995). Whereas traditional group-based designs generate knowledge on the population level only, SCEs generate knowledge at the level of the individual. Finally, findings from group-based designs cannot be generalized towards individuals, while SCEs can be aggregated for insights on populations.

Statistical Power of SCE RTs
Power analysis is an essential part of statistical hypothesis testing, particularly for determining sample size (Cohen, 1988). Several guidelines recommend a power analysis or at least a methodical description of how sample size was determined (Vohra et al., 2015; Wilkinson, 1999). In the context of SCEs, power analysis can be used to determine the number of measurements necessary, and for studies with replication, the number of participants.

For RTs, the randomization distribution is calculated empirically, and hence statistical power can only be estimated using computer-intensive simulations (Ferron & Ware, 1995; Onghena, 1992). Several studies have estimated the power of SCE RTs for various design conditions (Bouwmeester & Jongerling, 2020; De, Michiels, Taniaous, et al., 2020; Ferron & Onghena, 1996; Ferron & Sentovich, 2002; Ferron & Ware, 1995; Michiels et al., 2018). Although the results from these studies can be used as a reference, they are true only for the design parameters and simulated data distributions considered. Since we acquired unusual ordinal observed data, power ideally needed to be estimated using simulated data of the same type as this study would generate.

Methods
Experimental Setup
In this experiment, participants receive two different types of stimuli: a vibrotactile stimulation as the conditioned stimulus (CS) and an electrocutaneous stimulation as the unconditioned stimulus (US). At first, one neutral stimulus (a vibrotactile stimulus,
Figure 1. Arrangement of Tactors and Electrodes on the Back (Traxler et al., 2019, Adapted from)

Laboratory Setup

The participants receive two types of stimuli at the same time, replicating the procedure in Traxler et al. (Traxler et al., 2019). Stimulus onset and timing are controlled using Affect 4.0 (Spruyt et al., 2009). A vibrotactile stimulus is delivered to the skin using tactors manufactured by Dancer Design taped to the participant’s skin (Dancer Design, n.d.). This stimulus is of fixed duration and a clearly perceptible intensity. Three tactors are used in the arrangement as the source of a CS (see Figure 1). The SIDE tactor is allocated to CS\textsubscript{neutral}, whereas the ABOVE and BELOW tactors are assigned to CS\textsuperscript{+} and CS\textsuperscript{−} in a counterbalanced way across participants.

An electrocutaneous stimulation is delivered by passing a current across two surface electrodes located at the midpoint between the tactors to serve as the US (Figure 1). The current is delivered using a DS7A constant current stimulator (Digitimer Limited, n.d.). The stimulation intensity is calibrated individually. The number of pulses delivered on each stimulation occasion is varied to provide a US\textsubscript{high} that is usually painful, a US\textsubscript{low} that is usually non-painful, and a US\textsubscript{test} that is calibrated to lie close to the pain threshold, the boundary between non-painful and painful.

Tactor and Electrode Preparation

On arrival, participants are seated straddling a chair in front of a desk with a computer monitor, mouse, and keyboard. Participants are asked to bend backwards to identify the point at which the greatest bend is seen. A point in the upper lumbar region is marked on the back, 2cm to the left of the spine, where the electrodes are placed such that the mark lies exactly between them. Three other points—4cm above, below, and to the left side of the electrodes—are marked, and the tactors are taped in place such that the closest border of each tactor lies at a tactor mark (Figure 1). Calibration is performed by the procedure described in Traxler et al. (Traxler et al., 2019), with the CS\textsubscript{neutral} paired with each electrocutaneous stimulus. The CS\textsubscript{neutral} is used to provide vibratory stimulation consistent with the CS\textsuperscript{+}/CS\textsuperscript{−} stimuli that are later presented in the experimental trials to ensure that any modulation of pain by the vibration itself is consistent across calibration and experimental trials. Ratings of CS\textsubscript{neutral} trials from the experimental phase are not relevant to the research question and are therefore not analyzed.

Experimental Trials

Based on the experimental setup discussed previously, and given the pairing of CSs and USs, the experiment for a participant can theoretically consist of five different types of trials:

1. **Training trials:** Each training trial consists of one CS, with each of the three CS types used in different
trials. There is no electrocutaneous stimulation (i.e., US) presented. These trials are conducted to familiarize the participants with the locations of the CS tactors.

Baseline trials: The baseline trials consist of two combinations of stimulations: a CS+ with a simultaneous US\textsubscript{test} and a CS- with a simultaneous US\textsubscript{test}. These trials are used to record baseline pain ratings of participants before they are conditioned to associate the painful US with a particular CS.

Acquisition or learning trials: These trials consist of two combinations of stimulations: a CS+ with an US\textsubscript{high} and a CS- with an US\textsubscript{low}, and are meant to condition participants to associate painful US to CS+ and vice versa.

Test trials: These trials are the same as the two baseline trials: a CS+ with an US\textsubscript{test}, and a CS- with an US\textsubscript{test}, and are meant to record pain ratings of participants to test whether they have associated the painful US with CS+ and vice versa.

Irrelevant trials: These trials consist of a CS\textsubscript{neutral} combined with an US\textsubscript{test} Stimuli. These trials are noninformative and are included to satisfy the assumptions of the analytical approach; they are included only to achieve consistency in the time gaps between baseline or test trials.

Observed Variables

The participants are requested to provide two reports on each trial. First, they rate the stimulation event on the Sensation and Pain Rating Scale (SPARS; previously known as FESTNRS; Madden et al., 2019; Madden et al., 2017). The SPARS is anchored at -50 (no sensation), 0 (the exact point at which the feeling transitions to pain), and 50 (worst pain imaginable). The two distinct ranges for non-painful (-50 to -1) and painful (1 to 50) are clearly marked and explained to the participants. Participants are explicitly advised to make an initial decision about whether a trial was non-painful or painful before assigning a rating from the appropriate side of the scale, without selecting 0 on SPARS.

Second, the participant indicates the location at which they feel the vibrotactile stimulus from the three options “ABOVE”, “BELOW”, and “SIDE”. This is only used to confirm that the participants are identifying the location of the stimuli correctly. Given that discrimination of the vibrotactile stimuli is necessary for differential learning, participants who fail to identify the location of the stimulus correctly in at least 75% of the trials are excluded and replaced by additional participants.

Effect Size Measure and Test Statistic

We are interested in calculating whether the participants judged trials of CS+/US\textsubscript{test} and CS-/US\textsubscript{test} as painful or non-painful differently for baseline and test trials. For this purpose, we first convert the SPARS ratings to a binary variable: 0 for non-painful (-50 to -1 on SPARS) and 1 for painful (1 to 50 on SPARS). A rating of 0 on SPARS can neither be classified as painful nor as non-painful, therefore these are considered indeterminate and were marked as missing. Then, we pair two trials, a CS+/US\textsubscript{test} and a CS-/US\textsubscript{test}, and calculate the difference between the corresponding binary variables. The resulting difference indicates whether the participant is rating CS+/US\textsubscript{test} and CS-/US\textsubscript{test} differently. The difference can result in three different values: 1 when the CS+/US\textsubscript{test} trial is rated as painful while the corresponding CS-/US\textsubscript{test} trial is rated as non-painful; -1 when the CS+/US\textsubscript{test} trial is rated as non-painful while the corresponding CS-/US\textsubscript{test} trial is rated as painful; and finally, 0 when both CS+/US\textsubscript{test} and CS-/US\textsubscript{test} trials are rated as painful or both are rated as non-painful. This difference value, calculated from a pair of trials, constitutes one measurement in our SCE. A simple mean difference (MD) effect size measure is calculated as the difference between the means of the difference values derived from the test trial pairs and the difference values derived from the baseline trial pairs. Due to the flexibility of RTs, this effect size measure can also be used as the test statistic (Heyvaert & Onghena, 2014a).

Single-Case Design

In this section, we describe our considerations regarding which SCE design to use in our experiment. We discuss several initial design possibilities that were chosen by trial and error and conclude with a final design.

Initial Design Options

Randomized Block Design. Since each of our measurements requires a pair of trials, we immediately considered a randomized block design (RBD) for the experiment. In an RBD SCE, similar to an RBD in traditional group designs, the measurement occasions are divided into small blocks, with each block containing administrations of all treatment conditions (Onghena, 2005). For our study, a possible block consists of the two baseline (or test) trials, a CS+/US\textsubscript{test} and a CS-/US\textsubscript{test}, in random order. Hence, the experiment for a participant can consist of several blocks of baseline trial pairs, followed by a period of acquisition trials to condition the participants, followed by several blocks of test trial pairs. However, this design is immediately rejected, as an RBD SCE requires all treatment conditions to be applied inside a block. Since the effects of conditioning are acquired over a period of time and are expected to carry over to future trials, true alternation of treatment conditions is not possible in this experiment.

AB Phase Design with In-Phase Acquisition. When we rejected the RBD, we realized that the experiment should consist of periods of baseline trial pairs and periods of test trial pairs. The simplest method to achieve this is using an AB phase design, with...
an A phase as the baseline condition and a subsequent B phase as the test condition (Barlow et al., 2009). In favor of internal validity, the phase transition is preferably randomized. Each phase consists of several blocks of trials, each of which yields one measurement. Each measurement block in the baseline phase consists of $k$ irrelevant trial pairs (two $CS_{neutral}/US_{test}$) and a baseline trial pair (a $CS+/US_{test}$ and a $CS-/US_{test}$ in random order). Similarly, each measurement block in the test phase consists of $k$ acquisition trial pairs (a $CS+/US_{high}$ and a $CS-/US_{low}$ in random order) followed by a test trial pair (a $CS+/US_{test}$ and a $CS-/US_{test}$ in random order). The number of irrelevant or acquisition trial pairs in each block, or $k$, is set based on the required level of acquisition. A lower $k$ would mean we would need a lower number of trials to achieve a certain number of measurements, but would also typically result in weaker acquisition of associations, and vice versa for a higher $k$. Unfortunately, this design has low acquisition of the associations as there is no learning period. Additionally, RTs for AB phase designs require a large number of measurements to achieve sufficient statistical power (Ferron & Ware, 1995; Michiels & Onghena, 2019). As a result, the experiment for a participant would need a very large number of trials over a significantly long time to ensure both high acquisition and enough measurements. As a result, this design is rejected in favor of a slightly modified AB phase design. A similar ABAB phase design is also rejected due to the potentially large number of trials required and possibility of carry-over effects between phases.

**Final Design**

The final design combines elements from both previous designs: an AB phase design with both out-of-phase and in-phase learning. The baseline phase for this design contains several pairs of baseline trial pairs (a $CS+/US_{test}$ and a $CS-/US_{test}$ in random order), followed by a period of learning with only acquisition trial pairs (a $CS+/US_{high}$ and a $CS-/US_{low}$ in random order). Finally, the test phase includes test trial pairs (a $CS+/US_{test}$ and a $CS-/US_{test}$ in random order) with a few acquisition trial pairs in between at regular intervals to reinforce the associations and hence prevent extinction. The time of onset of the test phase is randomized. The experiment starts with a few training trials before the baseline phase.

The number of trials and the number of participants are to be decided based on a power study and some pilot experiments. However, the guidelines for phase designs by Kratochwill et al. (Kratochwill et al., 2010) recommend at least five measurements in both baseline and test phases. Since we are also curious about any increase or decrease in acquisition over time, we decide to include at least 15 measurements in the test phase.

Due to the use of specialized equipment and individual calibration, multiple participants cannot be tested at the same time. Hence, simultaneous replication using a multiple baseline design (MBD), which is able to control for environmental confounding factors, is not possible (Barlow et al., 2009; Kazdin, 2011). Instead, the experiment is to be run as a sequentially replicated AB phase design experiment.

**Randomization Scheme and RT**

The design is randomized using intervention start point randomization for each participant (Edgington, 1975). In this scheme, the total number of measurements and the minimum number of measurements in each phase is decided beforehand, and the intervention can start at any time point that satisfies these restrictions. Therefore, if the number of measurements is $N$, and the minimum number of measurements in A and B phase are $a$ and $b$ respectively, the number of possible randomizations is $r = N - a - b + 1$. For $P$ participants with the same randomization scheme, the total number of randomizations for the replicated experiment is $r^P$.

We conduct a single RT for all participants combined based on these $r^P$ randomizations. The null hypothesis for this RT is that no difference exists between the baseline phase measurements and test phase measurements for all participants. As discussed previously, the measurements represent whether participants rate $CS+/US_{test}$ trials as painful more frequently than $CS-/US_{test}$ trials. Since the Pavlovian conditioning procedure is aimed at participants perceiving the $CS+/US_{test}$ as more painful, the RT is one-sided with the alternate hypothesis being that the test phase measurements are higher than the baseline phase measurements for at least one participant. This RT can be conducted with the average (across participants) of the MD effect size measure discussed previously as the combined test statistic.

We use the Shiny SCDA (Single-Case Data Analysis) web app for SCEs to randomize and analyze the experiment (De, Michiels, Vlaeyen, et al., 2020). Whereas this web app does not include an option for simultaneously replicated AB phase design, it does include an MBD option. The MBD option in Shiny SCDA uses the Koehler-Levin regulated randomization procedure (Levin et al., 2018), which, when using an identical set of possible start points for all participants, is analogous to our randomization scheme. Hence, we can both randomly select test phase start points and run the RT in Shiny SCDA. Additionally, we can plot the observed data for visual analysis in Shiny SCDA.

**Pilot Tests**

We ran several pilot tests to estimate possible effect sizes and to test possible design choices. We first tested nine participants under slight variations of our initial design choice. Later, after a simulated
Table 1  Estimated Power Simulated Using Different Values for Effect Size, Number of Measurements, and Number of Participants.

<table>
<thead>
<tr>
<th>No. of measurements</th>
<th>No. of participants</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>0.1</td>
</tr>
<tr>
<td>25</td>
<td>10</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>13.4</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>16.4</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>18.9</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>22.2</td>
</tr>
<tr>
<td>30</td>
<td>10</td>
<td>13.0</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>17.9</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>21.9</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>26.4</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>30.0</td>
</tr>
<tr>
<td>35</td>
<td>10</td>
<td>14.3</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>21.0</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>27.4</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>32.8</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>37.4</td>
</tr>
<tr>
<td>40</td>
<td>10</td>
<td>15.4</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>23.6</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>30.7</td>
</tr>
<tr>
<td></td>
<td>40</td>
<td>37.0</td>
</tr>
<tr>
<td></td>
<td>50</td>
<td>42.8</td>
</tr>
</tbody>
</table>

power study and adjustments to the design, we tested another four participants. We analyzed these four tests using visual analysis and an RT in Shiny SCDA. These four tests were, however, not randomized, and the test phase for all four participants started at the tenth measurement occasion. However, for the purposes of demonstration, we ignored the assumption of randomization. We will discuss the results from these pilot tests, and the final four tests in particular in the Results section.

There were two important takeaways from the first nine pilot tests that were very useful for the design of the power study. First, the observed MD effect sizes were extremely small. The average effect size from the first nine tests was slightly negative at -0.037, with a maximum of 0.143 and a minimum of -0.400. Second, the pilot tests also revealed difficulties in running the experiment for more than 30-35 minutes for a participant. Considering each trial took around 15 seconds, this gave us a maximum of around 140 trials.

Power Study
To ensure sufficient statistical power for our RT, we needed to first estimate power for different values of number of measurements (N) and number of participants (P), and then choose sufficiently high values of N and P for the experiment.

Ideally, we would want to select the maximum N possible within the limits of how long an experiment can be reasonably run. This strategy presents us with two advantages. First, increasing N should result in reduction in the variability of the estimate of effect size. Second, maximizing N should allow us to achieve sufficient power with lower P, which directly equates to lower cost for the experiment.

We used the Monte Carlo method used by Ferron and Onghena (Ferron & Onghena, 1996) to estimate power. In this method, several datasets are simulated under a set of simulation conditions. The proportion of these simulated datasets in which the RT leads to a rejection of the null hypothesis gives an estimate of statistical power for the given set of simulation conditions. A simulation study of such complexity is both difficult to program and computationally intensive to execute. Fortunately, we were able to modify and repurpose R code used by De, Michiels, Taniaou, et al. (2020) for this study.

Simulation Conditions
The following three simulation conditions were varied for this power study:

Effect size: Based on the low effect sizes observed in the pilot tests, we simulated MD effect sizes of 0.1, 0.2, 0.3, 0.4, and 0.5 for the simulated observed data.

Number of measurements: We decided on a minimum of five measurements in the baseline phase and a minimum of 15 measurements in the test phase. As a result, the experiment needs more than 20 measurements. However, from the pilot tests it was evident that an experiment with more than 140 trials (70 trial pairs) was not feasible. Considering around 30 trials are required for the training and learning periods, and a few more acquisition trials for reinforcement during test phase, it was difficult to have more than 40 measurements (80 trials). Hence, we used 25, 30, 35, and 40 measurements for simulated data.

Number of participants: Since the pilot tests revealed a small effect size, and the number of measurements is also limited, to ensure high power we considered a large range of participant count at 10, 20, 30, 40, and 50.

Other Simulation Parameters
For the power study, we wanted to simulate observed values that were similar to the experiment. For the baseline, we assumed a symmetric distribution around 0 for the SPARS rating. Therefore, the painful/non-painful ratings were expected to be 0s and 1s equally for both CS+/US<sub>test</sub> and CS-/US<sub>test</sub> trials. Hence, an observed value corresponding to a trial pair in the baseline was expected to be 1 with 25% probability, -1 with 25% probability, and 0 with
50% probability. We simulated observed values as 1, -1, and 0 with these probabilities. For the test phase, we increased the probability of 1 by half the selected effect size and decreased the probability of -1 by half the selected effect size. This resulted in an expected MD effect size equal to the selected effect size.

Since the number of randomizations for the RT was huge, we simulated Monte Carlo RTs with 1000 randomizations (Edgington, 1969). For estimating power, we simulated 10000 datasets for each set of simulation conditions. Finally, we used a 5% level of significance for the simulated RTs.

The simulations were run on supercomputer nodes at the Flemish Supercomputer Center (Leuven, Belgium). This allowed testing more simulation conditions and achieve high accuracy simulating a large number of datasets for each simulation condition; however, the simulations can also be run at a smaller scale on a personal computer.

Results

Power Analysis

The results from the power study (Table 1) revealed that due to the relatively small effect sizes and restricted number of measurements, the number of participants need to be high to achieve 80% power. If we restrict the number of measurements to 30, which allows sufficient acquisition trials for learning and reinforcement within 30 minutes, 30 participants result in sufficient power even with a moderate effect size.

Based on these results, we decided to run the final set of pilot tests with 30 measurements. We decided on six training trials (two trials for each type of CS) before the baseline phase, 24 acquisition trials between the baseline and test phases, and additionally one acquisition trial pair for reinforcement after two test trial pairs during the test phase. With a minimum of five measurements in the baseline phase, and a minimum of 15 measurements in the test phase, this design allows for 11 possible randomizations. Depending on the number of measurements in the test phase which will vary based on the test phase start point selected, the experiment could theoretically consist of 104 to 114 trials. This is the design we intend to follow in the final experiment with 30 participants (Figure 2).

Pilot Tests

As mentioned previously, the initial nine pilot tests resulted in extremely small (and a few negative) effect sizes. The average MD effect size was -0.037, with a maximum of 0.143 and a minimum of -0.400. The low effect sizes indicated that the final design would need a large number of measurements and participants to achieve sufficient power in the RT. The first seven of these pilot tests consisted of 10-20

![Figure 2](image-url)
Figure 3  Plot of observed scores obtained from the final four participants in the pilot tests.

measurements each, which seemed too few. On the other hand, the last two pilot tests out of these consisted of 238 trials each. The feedback from both the experimenters and participants was that these tests were too long. These results influenced the decision to limit the number of trails to 140, remove irrelevant trial pairs from the baseline phase, and lower the number of acquisition trail pairs in the treatment phase.

The later four pilot tests were run using the design parameters we decided on after the power study. The average MD effect size for these participants was 0.057, with a maximum of 0.476 and a minimum of -0.250. Visual analysis (Figure 4) revealed that the first, third, and fourth participant did not seem to show any sign of conditioning. However, the second participant seemed to show significant conditioning effects. This is also confirmed by the MD effect size for the second participant (0.476). Finally, a Monte Carlo RT with 1000 randomizations resulted in a p-value of 0.112. Hence, the null hypothesis of the RT could not be rejected at a 5% level of significance.

Discussion

The power analysis and pilot test results confirm that the SCE design developed for this study can be effectively used to test the effect of classical conditioning on pain thresholds. The power study provides strong evidence that this design results in sufficient power if the number of measurements and participants are chosen correctly. The results from the final four pilot tests were encouraging. Even though the RT lacked power due to the small number of participants, the p-value was low. The visual analysis also seemed to suggest a large effect for at least one participant.

The final study using this protocol was not conducted immediately, due to lack of resources at the time. However, we hope to conduct this study as soon as the opportunity arises. Meanwhile, it seemed that the discussions regarding the SCE design and preparations for the study might be useful for other researchers developing similar protocols. Therefore, we decided to prepare this manuscript.

The final design suffers from a few limitations. The first limitation is the observed variable defined by us, which only yields values of -1, 0, and 1. Unfortunately, this restricts variation in observed data and can cause duplicates in the randomization distribution, which can affect power. An alternative is to use the difference between SPARS ratings from CS+/US test and CS-/US test trial pairs. However, this would require a slightly different hypothesis, which would not clarify whether classical conditioning can affect pain thresholds specifically.

The second limitation is due to the properties of the AB phase design. RTs using randomization of intervention start points in AB phase designs are known to lack power at lower sample sizes (Michiels & Onghena, 2019). AB phase designs do not satisfy the guidelines set by Kratochwill et al. (Kratochwill et al., 2010) without multiple replications. Since we are not sure how quickly the effect of our conditioning is reversed, we cannot use phase designs with more phase changes, such as the ABAB phase design. Due to the requirement of specialized equipment, we cannot use an MBD. Critically, we cannot increase the number of measurements due to time constraints. Therefore, we have to rely on sequential replications for both validity and statistical power.

The power study for this design also presents certain limitations. We used only one possible distribution of the SPARS ratings to simulate our observed data. While the assumption of a symmetric distribution around 0 is not unreasonable, with more pilot data, it might be possible to simulate using a distribution that resembles the observed data. We also did not account for any variability in effect size across participants. Instead, we simulated an equal effect size for all participants. However, we believe these are reasonable assumptions given the scope of our study.

Finally, we conducted Monte Carlo RTs for both the power study and the pilot data. As discussed previously, the number of possible randomizations for this design is $r^P$, where $r$ denotes the number of possible randomizations for one participant, and $P$ denotes the number of participants. Even for the smaller scale of the pilot data, computing $11^4$ pos-
sible randomizations would have been extremely costly. Monte Carlo RTs present a simple alternative to this computation cost while maintaining sufficient statistical power (Edgington, 1969; Hope, 1968). Hence, we intend to conduct a Monte Carlo RT in the final experiment.

As introduced earlier, this design presents certain advantages over traditional group study designs. Single-case designs are a better match for studying within-person changes, because they allow detailed insight into within-individual processes rather than assuming that the response of all individuals in a group is consistent. Additionally, these designs typically require a smaller number of participants. This allows for a lower overall cost and is particularly important for pain studies because it can be difficult to recruit participants. Tamal Kumar De, Victoria J. Madden, Johan W. S. Vlaeyen, and Patrick Onghena

Conclusion

In this manuscript, we described how we used trial and error to design an SCE testing whether classical conditioning can affect pain thresholds. We used a sequentially replicated AB phase design and conducted a simulated power study to determine sample size. We decided on 30 participants and 30 measurements per participant. Finally, we ran some pilot tests using our design. While the results from the pilot tests were inconclusive, they were sufficiently encouraging that we plan to conduct the full study in the near future.

Funding

This research is supported by the Asthenes long-term structural funding - Methusalem grant (nr. METH/15/011) by the Flemish Government, Belgium. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Centre), funded by the Research Foundation - Flanders (FWO) and the Flemish Government. Research collaboration supported by the IASP Developing Countries Collaborative Research Grant. VJM is supported by the Fogarty International Center of the National Institutes of Health (award K43TW011442). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Data Availability

The data and R code used in this study are openly available on Open Science Foundation

References


Table 2  Observed scores for the initial nine pilot tests.

<table>
<thead>
<tr>
<th>P1</th>
<th>V1</th>
<th>P2</th>
<th>V2</th>
<th>P3</th>
<th>V3</th>
<th>P4</th>
<th>V4</th>
<th>P5</th>
<th>V5</th>
<th>P6</th>
<th>V6</th>
<th>P7</th>
<th>V7</th>
<th>P8</th>
<th>V8</th>
<th>P9</th>
<th>V9</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>NA</td>
<td>A</td>
<td>-1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>-1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>-1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>-1</td>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>A</td>
<td>NA</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>-1</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>-1</td>
<td>A</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>-1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>NA</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>-1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>NA</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>-1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>1</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Appendix
Table 3  Observed scores from the final four pilot tests formatted for Shiny SCDA.

<table>
<thead>
<tr>
<th></th>
<th>P10</th>
<th>V10</th>
<th>P11</th>
<th>V11</th>
<th>P12</th>
<th>V12</th>
<th>P13</th>
<th>V13</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>NA</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>A</td>
<td>-1</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
<td>0</td>
<td>A</td>
</tr>
<tr>
<td>B</td>
<td>NA</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>NA</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>NA</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>1</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
<td>0</td>
<td>B</td>
</tr>
</tbody>
</table>
Do Carryover Effects Influence Attentional Bias to Threat in the Dot-Probe Task?

Joshua W. Maxwell, Lin Fang, Joshua Carlson

Threatening stimuli are often thought to have sufficient potency to bias attention, relative to neutral stimuli. Researchers and clinicians opt for frequently used paradigms to measure such bias, such as the dot-probe task. Bias to threat in the dot-probe task is indicated by a congruency effect i.e., faster responses on congruent trials than incongruent trials (also referred to as attention capture). However, recent studies have found that such congruency effects are small and suffer from poor internal reliability. One explanation to low effect sizes and poor reliability is carryover effects of threat – greater congruency effects on trials following a congruent trial relative to trials following an incongruent trial. In the current study, we investigated carryover effects of threat with two large samples of healthy undergraduate students who completed a typical dot-probe task. Although we found a small congruency effect for fearful faces (Experiment 1, \(n = 241, d = 0.15\)) and a reverse congruency effect for threatening images, (Experiment 2, \(n = 82, d = 0.11\)) whereas no carryover effects for threat were observed in either case. Bayesian analyses revealed moderate to strong evidence in favor of the null hypothesis. We conclude that carryover effects for threat do not influence attention bias for threat.

Keywords emotional cues, carryover effects, attention bias, threat, dot-probe

Emotional stimuli are thought to receive prioritized processing. Threatening stimuli like angry and fearful facial expressions of emotion often “beat-out” neutral or innocuous objects (Becker et al., 2017; Schubo et al., 2006) in the competition over spatial attentional resources (Desimone & Duncan, 1995). Attentional bias to threat is a topic of research that intersects many domains and is an essential function for a variety of organisms (Anderson & Britton, 2019).

One of the most commonly used tasks to assess attentional bias to threat is the dot-probe task (MacLeod et al., 1986). In a typical dot-probe task, participants search for the location of a target dot and indicate its position (left or right side of the screen) with a corresponding keypress. Immediately prior to the presentation of the target dot, two adjacent cues (one neutral and one threatening cue) are simultaneously presented for a very brief period (e.g., 100 millisecond (ms)). On congruent trials, the dot appears at the location of the threatening cue, whereas on incongruent trials, the dot appears at the location of the neutral cue (see Figure 1A). Each cue is completely irrelevant to the current task demands. And yet, the threatening cue biases attention to a greater extent than the neutral cue, as shown by faster response times (RTs) for congruent trials relative to incongruent trials (for reviews, see (Carretié, 2014; Imhoff et al., 2019). This effect is known as a congruency effect, or attentional capture by threat or attentional bias to threat.

While congruency effects are often assessed by the dot-probe task (Krujit et al., 2018; Mogg et al., 2017), they have come under scrutiny for poor internal and test-retest reliability (Schmukle, 2005; Staugaard, 2009). One root cause of these issues is that difference scores are inherently unreliable (for an explanation, see Hedge et al., 2018). However, some studies have found that attention (and congruency effects) may fluctuate towards and away from threat during the course of an experiment (Zvielli et al., 2015). This suggests attention bias variability could be a meaningful element of attention bias. However, these trial-to-trial variability measures appear to be strongly related to general RT variability (Carlson & Fang, 2020; Krujit et al., 2016). Identifying new sources of potential variation — like carryover effects — in the dot-probe task is beneficial because they could be accounted for in future studies.

Carryover Effects

Suppose that an individual’s attention was biased towards a threatening stimulus—like a fearful face—during a congruent trial of a dot-probe task. Such bias could impact the individual’s internal state (Panskepp & Watt, 2011), priming, or increase their association between the target and the threatening cue, which could subsequently cause them to be more biased towards threat in later trials. A persistent bias for threatening stimuli is quite plausible given that threatening cues can also bring about
Take-home Message

How is attentional bias to threat influenced by recent events? In the current study, we tested whether bias to threat (in the dot-probe task) is greater when threat cues recently appeared in a target location, relative to when they appeared in a non-target location (otherwise known as a carryover effect for threat). While we did not find carryover effects for threat, we suspect they are possible in other paradigms and that they exist in real-life scenarios. Carryover effects for threat in the dot-probe task are not likely to influence congruency effects.

expected that current trial congruency effects would be larger when the previous trial was congruent, relative to when it was incongruent (see Figure 1B). If our hypothesis was confirmed, we would have identified an important source of variability in the dot-probe task. On the other hand, if our hypothesis was not supported, we would conclude that the influence of carryover effects on congruency effects is negligible.

Experiment 1

Methods
Participants. Participants with an overall accuracy below 90% percent (eight in total) were removed from all analyses, leaving a final sample size of 241 ($M_{age} = 21.4, SD_{age} = 4.3; 178$ females). The data reported here were acquired during a screening session of a larger study on attention bias modification and brain structure, funded by the National Institute of Mental Health (NCT03092609). A portion of the current sample ($n = 127$) were included in Experiment 1 of (Carlson & Fang, 2020). The study was approved by the UNIVERSITY Institutional Review Board. Participants received monetary compensation for their participation.

Stimuli and Apparatus. The procedure was administered with a PC and a 16" LCD computer monitor. Stimuli consisted of 20 fearful and neutral grayscale faces of 10 different actors (half female; (Gur et al., 2002; Lundqvist et al., 1998). Fearful faces were rated (1-9, unpleasant to pleasant) as more negative ($M = 3.83, SD = 0.30$) than neutral pictures ($M = 4.45, SD = 0.52$), $t(18) = 3.23, p = .005, d = 1.48)$. Faces subtended a visual width and height of $5^\circ\times 7^\circ$ and the distance between the center of each face subtended $14^\circ$.

Design. The experiment consisted of five blocks of trials and each block consisted of 90 trials for a total of 450 trials. 30 congruent, 30 incongruent, and 30 baseline trials were randomly presented within each block. During baseline trials, two neutral faces were shown in the cue display and the dot was randomly presented in the same position as one of the two neutral faces (one third of all trials). Baseline trials were not included in our analyses. See Figure 1A for examples of incongruent and congruent trials.

Procedure. Each trial started with a white fixation cross (+) in the center of a black screen for 1000 ms, which was immediately followed by the cue display. The cue display was presented for 100 ms. In the cue display, there were two bilaterally presented faces. Immediately following the cue display, a single dot (the target object) was randomly presented in the central position of one of the previously shown faces in the cue display until participants responded. Participants indicated the location of the dot (left or right side of the display) by pressing a corresponding button with the pointer or middle finger of their
Examples of congruent and incongruent trials in Experiment 1. A carryover effect occurs when there is a larger congruency effect when the current congruent trial is preceded by a congruent trial, relative to when it is preceded by an incongruent trial. On the right is an example of two sequential trials where the target appears in the same location (repeating target location).

Figure 1

(A) Examples of congruent and incongruent trials in Experiment 1. (B) On the left is an example of two congruent trials—note the target (dot) appears in the location of the fearful face on the previous trial (trial n-1) and the current trial (trial n).

Results

All trials with an incorrect response (2.2% of trials) or a response time outside of 3 standard deviations from the individual-wise and condition means were removed from all RT analyses (0.6% of trials). Attentional bias was measured by subtracting mean RTs for congruent trials from incongruent trials. Positive values indicate the level of attentional bias for fearful faces. All analyses were not pre-registered but were pre-planned. As shown in Figure 2A, RTs for congruent trials ($M = 331$ ms, $SD = 72$ ms) were significantly shorter for incongruent trials ($M = 342$ ms, $SD = 72$ ms), $t(240) = 20.56, p < .001, d = 0.16$. A congruency effect was also found for percent correct (1.5%), $t(240) = 14.31, p < .001, d = 0.11$.

As shown in Figure 2B, carryover effects are measured as the interaction between previous trial congruency (pre-congruent or pre-incongruent) and current trial congruency (congruent or incongruent). The location of the fearful face and the target could repeat from trial to trial, so we further separated carryover effects by target location repetition. We examined target location repetition for two reasons: (a) it is an exploratory approach to analyzing carryover effects and (b) previous studies on carryover effects (Gladwin & Figner, 2019) have deliberately prevented cue and target locations repetitions trial-to-trial. We tested for the presence of carryover effects by conducting a $2 \times 2 \times 2$ Generalized Linear Mixed Model (GLMM) with the random effect of subject and the fixed effects of current trial congruency (congruent or incongruent), previous trial congruency (pre-congruent or pre-incongruent) and target location repetition (repeated or non-repeated) on single trial RTs using a Gamma probability distribution with a log link, which accounts for the slightly positive skew of the data. Target location repetition was included in the model because previous studies of carryover effects deliberately chose to prevent target locations repetition between trials (Gladwin & Figner, 2019), which leaves open the possibility that target repetition may moderate carryover effects. Similar to the t-test reported above, there was a main effect of current trial type, $F(1, 46115) = 348.91$, $p < .001$: RTs were faster on congruent ($M = 329$, $SD = 68$) relative to incongruent ($M = 339$, $SD = 69$) trials. There was also a main effect of target location repetition ($F(1, 46115) = 222.06, p < .001$) wherein RTs were faster for non-repeated target locations ($M = 330$, $SE = 2.32$) relative to repeated locations ($M = 338$, $SE = 2.38$). There was also an interaction between target location repetition and previous trial congruency ($F(1, 46115) = 6.08, p = .01$) such that RTs for non-repeated target locations were faster if the previous trial was congruent ($M = 329$, $SE = 2.34$) relative to incongruent ($M = 331$, $SE = 2.36$). Meanwhile, for repeated target location there was no statistically significant difference in RTs between trials where the previous trial was congruent ($M = 339$, $SE = 2.42$) relative to incongruent ($M = 338$, $SE = 2.42$). Critically, neither the 2-way interaction between current trial congruency and previous trial congruency ($F(1, 46115) = 0.59, p = .44$) nor the 3-way interaction ($F(1, 46115) = 0.84, p = .36$) were significant, indicating that no carryover effects were observed. No other effects were significant.

We followed up the GLMM approach with a Bayesian analysis. Given that there were no clear prior probabilities based on previous research, a diffuse (or uninformative) prior was used in SPSS to quantify evidence for the null hypothesis that there is no carryover effect for threat. A Bayes Factor analysis on a related-sample t-test indicated moderate evidence for the null hypothesis (BF01 = 5.38). In other words, the data is 5.38 times as likely under the null hypothesis than under the alternative hypothesis (i.e., that there is an influence of previous trial congruency on current trial congruency).

We also tested a model with a random slope for congruency x previous congruency x target location repetition interaction. Although this model fit our data better (AIC$_{without random slope} = -26903.474$ whereas AIC$_{with random slope} = -27133.846$; models with smaller AIC values fit better), the significant levels of all the fixed effects for this model was very similar to the one reported in the main text, where there was no significant carryover effect.
Experiment 2

In Experiment 2, our goal was to replicate and extend our lack of carryover effects with another category of threatening stimuli: threatening images (Lang, 2008). We used data from Experiment 2 of (Carlson & Fang, 2020). The method for Experiment 2 is the same as Experiment 1 except for the following.

Method

Participants. Participants with an overall accuracy below 90% percent (4 in total) were removed from all analyses, leaving a final sample size of 82 (Mage = 20, SDage = 2.2; 61 females).

Stimuli and Apparatus. The visual angle between the two stimuli presented in the cue display subtended a visual angle of 12°. The images used in the cue display were taken from the International Affective Picture System (IAPS; 200 Lang, 2008). 10 threatening and 10 neutral images were used, and the cue display was shown for 500 ms.

Results

Outlier RT trials were removed from all analyses (0.5% of trials). To test for carryover effects, we conducted the same 2 × 2 × 2 Generalized Linear Mixed Model on RTs as in Experiment 1. There was a main effect of current trial type (F(1, 15869) = 16.21, p < .001) wherein RTs were faster for incongruent (M = 345, SE = 4.06) relative to congruent trials (M = 349, SE = 4.11). Note that this effect indicates attentional bias to neutral, not threatening images, (see Figure 2C) which also might indicate for threatening images. There was also a main effect of target location repetition, F(1, 15869) = 167.53, p < .001), wherein RTs were faster for trials with non-repeating target locations (M = 340, SE = 4.00) compared to repeated locations (M = 354, SE = 4.17). However, as in Experiment 1, neither the 2-way interaction between current trial congruency and previous trial congruency (F(1, 15869) = 0.53, p = .47) nor were the 3-way interactions (F(1, 15869) = 0.01, p = .92) significant (see Figure 2D). Thus, no carryover effects were observed. In addition, no other effects were significant. We again followed up the GLMM approach with the same Bayesian analysis. For Experiment 2, we found strong evidence for the null hypothesis (BF01 = 11.27) indicating that the data are 11.27 times more likely under the null hypothesis than under the hypothesized existence of a carryover effect for threat.

Discussion

In two dot-probe experiments, we assessed whether attentional bias to fearful faces (Experiment 1) or threatening images (Experiment 2) would be influenced by carryover effects. Given that threatening stimuli are especially salient, their potential trial-to-trial influence on attention is quite plausible. However, the evidence supporting carryover effects for threat has been insufficient to rule out the influence of carryover effects (Gladwin et al., 2019; Hill & Duval, 2016). In Experiment 1, while we did observe a small sized congruency effect for fearful faces, we did not find an interaction between current trial congruency and previous trial congruency. Even for non-repeating target location trials (see Figure 1B), we found no such interaction. There are multiple potential explanations for faster responses on non-repeating target location trials. One is that participants strategically attend away from the emotional cue. Another possibility is that participants avoid previous target locations, as explained by inhibition of return (Posner et al., 1985). Lastly, our non-significant carryover effects for threat were replicated in Experiment 2. The results from our Bayesian analyses showed moderate to strong evidence for this null result.

The lack of carryover effects we observed is consistent with previous statistical findings from other dot-probe studies (Gladwin, 2017; Hill & Duval, 2016). These previous studies, when combined with the current findings, strongly suggest that attentional bias for threat in dot-probe tasks (congruent-incongruent) is unlikely to be moderated by carryover effects. Our study improved upon the limitations of these previous studies: we used a very large sample in Experiment 1, and did not administer any prior-experimental manipulations (like attentional training in (Hill & Duval, 2016). The current evidence suggests that previous studies and future research involving the dot-probe task do not need to consider whether congruency effects are influenced by carryover effects.

Carryover effects are analogous to the congruency sequence effect commonly observed in the flanker, Simon, and Stroop tasks (Dutchoo et al., 2014). One difference between the dot-probe task and these other tasks is that the distractor stimulus does not directly conflict with the target. For exam-

Original Purpose

The objective of this study was to assess whether carryover effects for threat would be observed in the dot-probe task. Related studies have found support for carryover effects for threat (Gladwin, 2017; Gladwin & Figner, 2019; Gladwin et al., 2019; Gladwin et al., 2020). However, there are several unique elements to the task design and procedure used in these studies that make them difficult to compare to the commonly used dot-probe task (a task often used in applied, clinical, and basic research on emotion processing). Our purpose was to extend these previous studies of carryover effects for threat to the standard dot-probe task.
ple, in the Stroop task, the font color of the word stimulus can be incongruent with the meaning of the word, which creates conflict when the goal is to indicate the word color. Similarly, in the arrow flanker task, conflict is created on incongruent trials; the two arrows to the left and two to the right point in the direction opposite to the direction of the center target arrow, like this: < <> <<. On congruent trials, all five of the arrows point in the same direction. However, in the dot-probe task, the target does not co-occur with the conflicting cue (they are shown in different displays); they are separated in time, and they are visually distinct categories of stimuli because the cue is a face, and the target is simply a single dot. Therefore, current trial congruency effects in the dot-probe task may stem from relatively low stimulus conflict and therefore be unaffected by previous trial congruency. In sum, carryover effects, or congruency sequence effects, may only arise for specific types of stimulus conflict.

There are some notable procedural differences between the standard dot-probe task (as used here) and other studies that used a similar spatial cueing task but did observe carryover effects. For example, in the unique study design by (Gladwin & Figner, 2019; Gladwin et al., 2020), they presented their cues until the participants made a response, whereas we only presented our cues for 100 ms (Experiment 1) or 500 ms (Experiment 2). Presenting cues until a response is made increases the possibility of attentional dwelling or delayed disengagement from the cue, as opposed to the initial orienting of attention towards the cue (Fox et al., 2001). Another key difference is that (Gladwin & Figner, 2019; Gladwin et al., 2020) presented a distractor (non-target object) in their search display, whereas in the classic dot-probe task, the target appears alone (Mogg & Bradley, 1998). Inhibition of the cue that appears in the previous non-target location might explain carryover effects for the cue (Gladwin et al., 2020). The presence of the non-target distractor might facilitate such inhibition, thereby creating a carryover effect (Gillich et al., 2019; van Moorselaar & Slagter, 2019). Future studies on carryover effects should carefully consider their task design procedure.

Conclusion

Researchers should carefully select their study design when examining congruency effects. As for studies involving the dot-probe task, however, carryover effects for threat do not appear to be a concern. The highly controlled and simple design of the dot-probe task adequately controls for sources of variance such as carryover effects. While threatening cues may bias attention to a greater extent than neutral cues, their residual influence on attention is limited when assessed with the dot-probe task.

Data Availability Statement

Datasets are available here.
References


Are We Meeting Best Practice Standards?: A Longitudinal Analysis of Mental Health Practices Within the Florida Child Welfare System with Implications for Child Well-being

Daniel Dunleavy

Best practice standards are one method by which medical providers ensure effective care, thus promoting well-being. Though formal guidelines have been recently implemented to direct and standardize children’s mental healthcare in Florida, little research has evaluated the extent to which they are executed in practice. This study aims to fill this gap by analyzing Florida Medicaid data. Individual-level data will be collected from a 12-month period from a random sample of children, on Medicaid, with a mental health diagnosis; to: 1) Describe the type and frequency of mental health services provided to this sample, including to those in the child welfare system; 2) Evaluate the extent to which Florida's Psychotherapeutic Medication Treatment Guidelines are adhered to; and 3) Analyze sociodemographic characteristics, to determine if there are predictive factors which account for undertreatment/overtreatment. Data will be coded for congruence with these standards and analyzed using multinomial logistic regression.

Keywords health, best practice guidelines, social work

Research proposal

General information

Institution of employment at time of application
Florida State University
Prospective host institution
Florida State University
Chapin Hall, University of Chicago
Main field
Social Work
Other fields
Child Welfare
Required length
8/7 pages

Description of the proposed research

Primary research questions

Question #1 What are the sociodemographic characteristics of children on Medicaid, diagnosed with a mental health condition? Sociodemographic characteristics will further our understanding of how frequently different groups of children are diagnosed with a mental health condition. Diagnoses will be cross-tabulated with age-group, gender, race, placement setting, etc. to answer related secondary questions regarding diagnostic prevalence by group.

Question #2 What types of treatments are being provided to children diagnosed with a mental health condition? Descriptive statistics about type of mental health treatment (i.e. drug and/or psychosocial interventions) will further our understanding about how frequently each type of treatment is provided and how often specific interventions (e.g. cognitive behavioral therapy [CBT], fluoxetine [Prozac]) are utilized. This includes identifying trends in concerning pharmacological practices (e.g. polypharmacy, use of multiple antipsychotics, off-label antipsychotic use; as outlined in Government Accountability Office, 2012, p. 11). This information can be mapped across sociodemographic variables to describe what treatments each subgroup of children (e.g. Medicaid-only vs. child welfare involved children) is receiving. Knowledge gained from this research question can identify if particular groups of children are being differentially treated with respect to their peers in the population; and to determine how these rates compare with national prescription trends (Ofinson, Marcus, Weissman, & Jensen, 2002) and those within the wider child welfare system (Raghaven et al., 2005).

Question #3 To what extent are the Florida Psychotherapeutic Medication Treatment Guidelines be-
ing met? The “Florida Guidelines” (Florida Medicaid Drug Therapy Management Program for Behavioral Health, 2017) represent best practice standards. The guidelines are congruent with the most recent research (American Academy of Child and Adolescent Psychiatry, 2015, p. 29; The TADS Team, 2004), which specifies that all children with complex mental health needs should be provided psychosocial therapy concurrently with medication, where possible. Understanding the frequency of guideline adherence will elucidate how often quality, evidence-based mental health treatments are provided. While deviation from guidelines may be occasionally justified (Colbrook, 2005), their consistent implementation is essential to ensure that effective treatments are provided, which help protect and promote child well-being.

Question #4 What sociodemographic factors are predictive of treatment discrepancies? A multinomial logistic regression will be conducted to determine if particular sociodemographic characteristics are predictive of treatment quality. This analysis will determine whether a child’s age-group, gender, race, placement setting, abuse history, etc. increases the odds that they will receive: 1) Recommended care (Guideline adherence); 2) Undertreatment (i.e. no treatment, or treatment lacking medication/psychosocial treatment); or 3) Overtreatment (i.e. polypharmacy, atypical/contraindicated drug use). This information will allow for greater scrutiny within the field and ensure that particular groups of children aren’t harmed by extreme deviations in care.

Research approach

Data Source This dissertation will utilize Florida Medicaid data, which contains records of all children enrolled in Florida’s Medicaid program. This dissertation will also utilize a statewide administrative database from the Florida Department of Children and Families (DCF). The Florida Safe Families Network (FSFN) database contains records of each individual child in Florida’s child welfare system, and will be used to identify a subgroup of children on Medicaid, in child welfare placement settings, for comparison.

Population My population of interest is children, on Medicaid, in Florida, with a mental health diagnosis. A random sample will be taken from the Florida Medicaid database, from a 12-month period. The sample drawn will be random, in order to facilitate sound statistical inferences to the larger population of children (see undefined Berk, 2004, pp. 39-52). A subgroup of this population (i.e. children in the child welfare system) will be identified within the sample, using their unique Medicaid ID numbers. Merging the information from these two datasets will provide a novel data source, thereby increasing the depth of analyses and the importance of the study’s findings.

Data access Access to the Medicaid and FSFN databases will be requested in the Spring of 2018, with estimated approval by August of 2018. Permission to use the FSFN database is not an obstacle, as Florida State University’s Institute of Child Welfare (FICW) is housed within the College of Social Work. FICW is a key partner with the Florida Department of Children and Families. I’ve spoken with Dr. Jessica Pryce, director of FICW and a data analyst at the Florida DCF. Both view access to FSFN as feasible and straightforward process. In an effort to investigate the feasibility of using Florida Medicaid data, I am currently working with an AHCA, Florida-Medicaid data analyst to coordinate the data request process. Data received from Medicaid and FSFN will be anonymized, individual-level data. It will be coded into an SPSS dataset and include, among other variables, information about age-group, race, gender, placement setting, and treatment types. Other variables may be collected, pending review of data.

Study design This longitudinal study aims to assess the extent to which the Florida Guidelines are adhered to in practice. Data will be collected from a 12-month time period. All cases meeting inclusion/exclusion criteria will be assigned a unique ID number. The sample will be drawn, at random, from this group. Data will be entered into SPSS. Sociodemographic variables (independent variables) will be collected and analyzed for each case. Multinomial logistic regression analyses will be conducted at the end of data collection, to determine what characteristics are predictive of treatment variables/guideline adherence (dependent variables).

Limitations Given the size of Florida Medicaid enrollment, sample size is not perceived as a limitation. Missing data will be analyzed, in order to determine if it is Missing Completely at Random (MCAR). Missing data will be replaced using Multiple Imputation (MI), if methodologically appropriate. A second, more philosophical limitation regards psychiatric diagnosis. Given the current state of psychiatric knowledge, the act of making a diagnosis is a fallible process, relying on descriptive-level, behavioral criteria (Regier et al., 2013). Consequently, some children may be incorrectly diagnosed (false positive diagnosis), while others not in the population under study are incorrectly left undiagnosed (false negative diagnosis). As in all studies of this type, diagnoses are presumed to be correct.

Analytic methods

Descriptive statistics Descriptive statistics (e.g. frequencies, measures of central tendency, and dispersion) will be reported for all variables in order...
to describe the characteristics of the sample. A cross-tabulation will be conducted to examine the frequency of different types of treatment services and adherence across key sociodemographic variables.

Inferential statistics A multinomial logistic regression was chosen as the method for calculating the odds that a child will receive treatment that adheres to the guidelines, or that can be considered a case of overtreatment or undertreatment. While differences can be noted at the descriptive level, this analysis will help uncover which groups are at greatest risk for deviations in care, by use of frequentist inference methods. The reported odds ratio (OR) serves as the effect size (Tabachnick & Fidell, 2012, pp. 464-465) of the relationship between the predictor (i.e. IVs) and outcome variables (i.e. DVs), and is reported with the associated confidence interval (CI, 95%; Cumming & Finch, 2005). Confidence intervals provide an estimation of the precision of the OR (Szumilas, 2010, p. 227). P-values will be reported, as a tool for controlling error rates (i.e. the false positive discovery rate) and will be judged, at a minimum, according to the standard .05 threshold; though recent debates (Benjamin et al., 2018; Colquhoun, 2014; Lakens, 2017) call for a more nuanced approach. Bayesian methods, which include the use of credible intervals (Hoekstra, Morey, Rouder, & Wagenmakers, 2014), will be considered given the nature of the dataset (see generally Berk, Western, & Wiess, 1995).

Work plan

Timeline I have completed all coursework required for my degree and have successfully passed my preliminary examinations. I will begin my prospectus writing in the beginning of January 2018. I plan on defending my prospectus in April of 2018 (04/18). Upon completion I can begin formally securing the data needed for my dissertation. I expect to have approval and access to the Florida Medicaid database by August of 2018 (08/18). I expect to have approval and access to the Florida Department of Children and Families’ (DCF) Florida Safe Families Network (FSFN) database by August of 2018. Access to the FSFN database is not viewed as an obstacle. FSU is home to the Florida Institute for Child Welfare (FICW), a key partner to Florida’s DCF. Data from across a 12-month period, from these two databases, will be synthesized by September of 2019 (10/19). I expect to have analyzed the data and written my full dissertation by March of 2020 and defend my dissertation by the end of April (04/20). I plan on graduating at the end of the Spring semester of 2020 (05/20).

Relationship with your policy practice mentors My policy mentor is Mr. Robert “Bob” Whitaker. I chose Bob as my policy mentor because he has strongly influenced my professional career. My desire to get my PhD and subsequent research has been greatly informed by his published work and advocacy through the Mad in America Foundation. I value his depth and breadth of knowledge on psychiatric medications, his enthusiasm for the issues surrounding mental healthcare (i.e. diagnosis/etiology, treatment efficacy, long-term treatment outcomes), and passionate efforts to reform the systems of care.

The deepest areas of my knowledge of mental healthcare are related to the issues of diagnostic accuracy, treatment effects, and practice ethics. Bob’s work and knowledge complements my own knowledge and work, in that he has extensively covered issues surrounding long-term treatment outcomes and effects on well-being, as well as issues surrounding polypharmacy, off-label prescribing, and best-practice standards. These latter areas of focus will greatly inform my dissertation work. Bob’s ability to interpret discrepancies in treatment delivery, will help me to identify and operationalize instances of undertreatment and overtreatment. Similarly, Bob’s knowledge of best practice standards will help me advocate for practice- and legislative-level changes in children’s mental healthcare.

During the fellowship, I will be regularly in touch with Bob via phone, video chat, and email. It is also likely that we will meet during the fellowship timeframe at academic conferences to discuss issues in the field and progress of my research. I hope to utilize the mentor relationship to help disseminate my research findings, to reach a diverse array of mental health stakeholders (e.g. researchers, clinicians, advocates/policy-makers, consumers/families), and to become better versed in the nuances of policy-change and policy-advocacy. I am ecstatic to have the opportunity to learn and grow as a professional and scholar from Bob.

Knowledge utilisation

Policy practice implications

Whatever their inherent limitations, when properly implemented, practice guidelines help curb extreme deviations in care (i.e. undertreatment, overtreatment). The implications of this study are twofold. It will: 1) Improve social and emotional well-being by ensuring that evidence-based treatments are provided; and 2) Identify and reduce the harm caused to children by extreme deviation in care. Identification of groups that might be at higher risk of receiving undertreatment or overtreatment will lead to greater institutional oversight and attention by staff in practice settings. These goals are congruent with the national efforts to standardize healthcare practice (Institute of Medicine, 1990; 2011), with federal and institutional calls for greater oversight over childhood medication prescription (Government Accountability Office, 2011, 2012, 2017; Leslie, 2010).
and with the concerns of child welfare researchers (Mackie et al., 2016) and legal scholars (Strawbridge, 2011).

Ensuring the provision of evidence-based treatment Childhood physical and sexual abuse and neglect are strongly associated with the development of childhood mental health problems (Administration for Children & Families, 2012) and the development of later adult psychopathology (Guadiano & Zimmerman, 2010; Matheson, 2013; Read & Bentall, 2012; Read, van Os, Morrison, & Pam, 2005). Therefore, addressing mental health problems of children on Medicaid and in the child welfare system is essential for promoting current and future well-being. An analysis of Florida Guideline implementation, will ensure that children diagnosed with mental health conditions are and will receive appropriate care.

Preventing the neglect and abuse of children's mental healthcare. In Florida, failure to provide proper mental healthcare can legally be considered a form of child neglect (Child Welfare Information Gateway, 2015). As the Florida Guidelines represent best practice standards, failure to consistently implement these guidelines in child welfare systems can be regarded as a form of institutional child neglect. The consequences of neglecting mental health problems are severe. Untreated depression can lead to a worsening of symptoms, resulting in suicide, self-harm, decreased academic performance (Kovacs & Goldston, 1991), and social deficits (Geller, Zimmerman, Williams, Bolhofner, & Craney, 2001). In other words, untreated mental health problems can be detrimental, if not seriously harmful, to a child's well-being. Likewise, overtreatment (overmedication, polypharmacy, and off-label prescription) can lead to serious physical and mental health problems; a trend some call "pharmacological abuse" (Healy, 2009, pp. 272-273). Overtreatment is not uncommon in child welfare systems, where children are 2-3 times more likely to be medicated than their peers in the community (Crimson & Argo, 2009; Raghavan, 2005). Foster care children are estimated to be medicated at a rate 3-13 times that of their peers (Leslie et al., 2010). Children on Medicaid are prescribed antipsychotics at twice the rate of those privately insured (Government Accountability Office, 2012). This is concerning since overtreatment can increase the rate of adverse effects (Whitaker, 2010), which include: akathisia (Looen & Stahl, 2011), suicidality (Bielefeldt, Danborg, & Gotzsche, 2016; Healy, 2005), and physical dependence syndromes (Fava et al., 2015). Long-term antipsychotic use, is associated with progressive reduction in brain volume (Fusar-Poli et al., 2013) and Parkinson's-like movement disorders (American Psychiatric Association, 1992). Thus, identifying cases of overtreatment (and undertreatment) will promote children's physical and mental well-being, prevent harmful trends in treatment practice, and lead to policy change.

Cost estimates
Total budget requested
30,000 $ US

Intended starting date
August 2018

Application for additional grants
None

Data management plan
N.A.

Ethics
N.A.

Society
Public summary
N.A.

Reviews
N.A.

Rebuttal
N.A.

Decision
Doris duke fellowships application
Doris Duke Fellowships <DDFellowships@chapinhall.org>
Tue 3/13/2018 9:18 PM
To: Daniel Dunleavy <djd09e@fsu.edu>

Dear Dan,

Thank you for applying to the Doris Duke Fellowships for the Promotion of Child Well-Being and participating in atelephone interview for finalists. Unfortunately, after careful consideration, the selection committee did not select your application for the fellowship. It was a very difficult decision that does not necessarily reflect on the quality or relevance of your proposed research. In selecting the cohort, every attempt is made to achieve a group of scholars that will complement each other's skills and represent a diversity of disciplinary perspectives, research interests, and professional experiences. Our final selection is difficult, particularly when the finalists are as highly qualified as they were this year.

We greatly appreciate your interest in the fellowship and regret that we are not able to offer you an
award. We wish you the best of luck in completing your dissertation research and in your future endeavors, and hope that you will continue to focus on the prevention of child abuse and neglect.

Sincerely,
Deborah Daro, PhD.
Chair of the Selection Committee

Doris Duke Fellowships for the Promotion of Child Well-Being
Senior Research Fellow
Chapin Hall at the University of Chicago

References


Treatment for Adolescents With Depression Study (TADS) Team. (2004). Fluoxetine, Cognitive-Behavioral Therapy, and Their Combination for Adolescents With Depression: Treatment for Adolescents With Depression Study (TADS) Randomized Controlled Trial. JAMA, 292(7), 807. https://doi.org/10.1001/jama.292.7.807

Appendix

Personal statement

Your expected benefits from the fellowship

Interdisciplinary perspective Interdisciplinary work has been an extremely valuable part of my professional and academic development. In my professional experience, I have worked as a social worker among physicians, nurses, psychologists, support staff, and other social workers. Each has enriched my experience by offering a unique lens from which to understand my clients and the challenges faced as part of a professional team. At the academic level, my education has been unquestionably enriched by taking courses across fields. It has provided me the opportunity to learn from researchers in the fields of Sociology, Psychology, Philosophy, Medicine, and Public Health, and has taught me how to analyze social problems from a variety of perspectives.

One of the things which I am looking forward to most, as a fellow, is collaborating with and becoming enriched by the Doris Duke Fellowship and its multidisciplinary peer learning network. Undoubtedly, the fellowship's interdisciplinary perspective and emphasis will help shape my approach and research regarding child well-being, as well as the scope of the issues surrounding providing mental healthcare to children, both within and outside the child welfare system. Considering that previous cohorts have been comprised of fellows from the fields of Sociology, Psychology, Education, Public Policy, Criminal Justice, Social Work, and Public Health (among others), I believe that my work and how I understand these social problems, will only become further enriched and refined.

Peer learning network The feedback that I receive most frequently from my academic peers is that I have an insatiable appetite for knowledge. I continuously strive to learn, to understand the world around me, and to solve important social problems. As a classmate and colleague, I am attentive, offer constructive feedback, and engage thoughtfully with new ideas, even where they diverge from my own particular research interests. These are personal strengths which I can contribute to the work of other fellows. While I do not pretend to know everything about every topic, I always try to sincerely and thoughtfully engage with an issue. I try to use my critical thinking skills to help others refine their ideas and to provide constructive feedback. I attribute these assets to my supplementary coursework and interest in philosophy, a field which has helped me to greatly clarify and explicate my own ideas and arguments. As described in the previous section*, one of the things I benefit from most as a social worker and academic is different perspectives. I believe the Peer Learning Network will allow me to engage with fellows from disciplines both similar to and quite different from my own, exposing me to new ideas, perspectives, arguments, and resources. It is clear to me that the peer network has great potential to improve and refine my research and ideas; as well as for me to contribute to the intellectual development and growth of other fellows. This aspect of the fellowship is among the most exciting opportunities available to me in my academic and professional development.

Looking forward

Immediate plans following graduation

Immediate professional goals My immediate plans following graduation are to secure a tenure-track position in a school of Social Work at a research university in the United States or related policy center. This will be a place where I can continue researching on child welfare, mental health, and help foster “good research practices” both within my profession and within child welfare research. Additionally, I aim to teach classes related to child welfare, psychopathology, and social policy. While research is my primary motivation for coming into academia, this latter role of teacher is no less important to me. I believe that helping develop the next generation of social workers, many of whom will work with children in some respect, is of the utmost importance in ensuring that our policies and practices are successfully implemented.

Immediate research plans I hope to extend my dissertation work, by seeking funding from federal, state, and non-governmental organizations (NGOs). I aim to analyze mental health practice guidelines in child welfare systems across the nation, to analyze how children’s mental health services are provided nationally, and to engage with policy reform to improve child well-being and mental functioning. This research agenda would allow for the attempted replication of the findings from my dissertation, a hallmark of rigorous scientific research.

Long term career aspirations My long-term aspirations can be divided into three non-exclusive parts. My goals are to: 1) Refine our understanding about mental health treatment effects, particularly with children; to 2) Contribute to the improvement of children’s mental health policy and practice locally and nationally; and to 3) Improve the quality of child welfare research.

Goal #1 My most important long-term goal is to use information gathered about state guidelines (described in Immediate Plans above), along with knowledge of long-term treatment efficacy/safety (see point 2 below), to ensure that the mental health guidelines utilized in each state are truly evidence-based and reflect ‘best-practice’ standards for children. I plan on using information gathered about

* Appendix
discrepancies in guideline adherence to ensure that these improved standards are effectively implemented with all vulnerable children, once they are in place.

Goal #2. Despite enormous research funding, few studies have been undertaken describing the long-term benefits, harms, and efficacy of many mental health treatments frequently used with children. Long-term evaluation will help us to develop more effective policies and practices for vulnerable children with mental health needs. While effective guideline implementation will minimize extreme deviations in care, this goal will evaluate the long-term effects of services.

Goal #3. “Good research” is founded on rigorous research methodology, sound statistical inferences, attempted replication and transparent/open research practices. This is true of child welfare research and the long-term areas of research outlined above. I aim to advocate for, research, teach, and practice “good research practices”, as exemplified by the Center for Open Science, Society for Improving Psychological Science, and others.

Is there anything else you want to share with the selection committee? I am thrilled to have the opportunity to be considered as a Doris Duke Fellow. The Foundation’s mission, to develop and disseminate knowledge; and the Fellowship’s goal of creating long-lasting improvement in child well-being, strongly coheres with my own research goals and ethos. This is my second time applying for the fellowship. The past year has provided me the opportunity to clarify and refine my research topics and goals. I am reapplying because I strongly value the opportunity to engage with, collaborate, and learn from the Fellowship’s members and multidisciplinary peer learning network. It would greatly improve my scholarly work and professional development to be a Fellow. I believe my intellectual drive, critical thinking skills, and strong research background will make me a valued colleague and collaborator within both the Doris Duke Foundation and Fellowship program. Thank you to the Selection Committee for taking the time to consider this application.
Discourses surrounding migration and integration often see language, and in particular the knowledge of the native language, as a crucial barrier to minorities’ access to healthcare and welfare benefits, equal healthcare treatment, social integration, and psychological wellbeing. Using methods of ethnographic and interactional sociolinguistic and conversational analysis our project investigates how healthcare and welfare professionals and Greenlandic patients define, interpret and manage communication and language inequalities in face-to-face encounters. What are the practical, cognitive, psychological and social consequences of “miscommunication” for the Danish Greenlanders? We examine four distinct aspects of communication: conversational strategies, non-verbal behavior, linguistic insecurity, and attitudes. Our aim is to understand the entire communicative circuit (i.e. channels by which information is transmitted), developing on our idea of “affective language economies of health”.

Keywords language, health, communication, equality, Greenlanders in Denmark

Research proposal

General Information

Institution of employment at time of application

Copenhagen Business School, Department of Management, Society and Communication

Prospective host institution

University of Copenhagen, Department of Cross-Cultural and Regional Studies

Main field

Intercultural communication, health studies

Other fields

Anthropology/linguistics

Length requirement

7 pages, 22,000 characters, including tables, but excluding bibliography

Description of the proposed research

Language, and in particular knowledge of the Danish language, is often seen as a crucial barrier to equal healthcare treatment and to minorities’ access to Danish healthcare and welfare benefits.

Sharing the national and international ambition to achieve more equity in healthcare and welfare system, this research project investigates the role language and non-verbal behaviour play, when Greenlanders move to Denmark and encounter the Danish healthcare and welfare professionals (SVPs), and where this poses a potential communication barrier.

Our aim is to examine the entire communicative circuit (i.e. channels by which information is transmitted), providing new empirical knowledge and theoretical insights into the emerging field of study of language and health.

Aim and objectives

The main objectives are

• Empirically, we seek to elucidate how language ideologies (i.e. sets of beliefs and ideas about language articulated by users), continuously re-produced in institutionalized practices, and non-verbal behaviour may lead to a feeling of miscommunication and injustice among the Greenlandic patients, cause cognitive blocking, emotional stress, self-exclusion and feelings of injustice among them. It will provide new original knowledge about the context-specific tendencies and potentials that occur when people with distinct social, health, psychological, and communicative “challenges” meet the Danish SVPs, and add new knowledge on values and behaviour (verbal and non-verbal) of Greenlanders in Denmark.

• Theoretically, this study seeks to expand the field of study of language and health by bridging health studies with intercultural communication theories, linguistics, and emotional theories under what we call “affective language economies of health”.

• Together with public sector partners and professional universities we seek to develop a method through which identified by participants barriers may be overcome or reduced in the everyday practice of SVPs in their encounters with patients, who...
have a different first language.

Research questions The objectives will be pursued by focusing on the following research questions:

To what extend does language (linguistic proficiencies and ideologies) and non-verbal behaviour influence communication in clinical encounters and welfare in Denmark? What are the practical, societal, health and psychological consequences of potential miscommunication for the patients?

Empirical focus: the Danish Greenlanders. In the recent years, there has been an increasing focus in the Danish media and among the Danish researchers and policy makers on social integration and inclusion of ethnic minorities (migrants, refugees and asylum seekers) from countries outside of Nordic countries, EU and North America. Greenlanders, on the other hand, have been “overlooked” in the Danish context.

As of today, there are 16.370 Greenlanders in Denmark. This number is constantly growing.

Another thousand come to Denmark each year to receive medical treatment. Yet, because Greenlanders are Danish citizens they are not offered the same integration aid as refugees and immigrants when they visit for health treatment or move to Denmark. Not only does this “lack of integration effort” contrasts Canada and the US attempts to accommodate indigenous minorities’ cultural and linguistic needs in recognition of Indigenous rights, but it is also “a form of misunderstood equality that you will treat the Greenlanders as Danish nationals and give them the same rights” (Toft, 2018).

Indeed, although about 80% of Greenlanders manage well, and the majority are well integrated into Danish society, there is still a great distance between Denmark and Greenland, geographically, culturally, and linguistically. Similar to refugees, many Greenlanders who arrive in Denmark have experienced social problems at home and/or have been traumatised (Laage-Petersen, 2015). Many lack a regular income, feel socially isolated and/or anxious living on their own in Denmark (Socialministeriet, 2003). They too experience prejudice and discrimination when encountering the Danish public authorities and healthcare. According to recent sociological research language (“poor Danish language skills”), negative representations and stigmatization of Greenlanders, and lack of knowledge of rights and appeal regulation procedures are among the major challengers (Laage-Petersen, 2015; Togeby, 2004).

According to Heidi Conradsen, project manager at Oqquumut, Activation Center for Greenlanders in Esbjerg Municipality, welfare professionals often experience “significant cultural differences” with regard to conversation style, humility, and a general lack of societal understanding (personal communication). In the White Paper on “socially disadvantaged Greenlanders", The Danish Ministry of Social Affairs points out that Danish welfare and healthcare professionals need concrete help in order to recognize Greenlanders special needs and address the challenges, because Greenlanders are often met with prejudice, negative and stigmatizing expectations, which may become “a self-fulfilling prophecy” (Socialministeriet 2003 s.21). Nonetheless, healthcare professionals and social workers are not offered any specific training.

Our project addresses some of these critical issues by investigating situated practices (what is actually going on in conversations between Greenlanders and SVPs) and experiences of the mentioned above characteristics and of clinical encounters of a selected group of Greenlanders. In the analysis, we focus on four distinct aspects of communication: language proficiencies, conversational styles, societal attitudes, and non-verbal behaviour.

Theoretical approach: Previously, researchers have focused mainly on social inequality as an effect of negative biases (social stereotypes and prejudiced attitudes) against Greenlanders in Denmark, measuring reported behaviours and attitudes (Togeby, 2004; Laage-Petersen, 2015). These studies do not look at what is actually going on during the encounters, nor do they address the role language (linguistic competences and ideologies) actually plays in the construction of these biases.

Yet, as a series of linguistic anthropological studies have shown, languages is anything but neutral. Language is not only crucial for immigrants’ political and social participation, and individual's ability to communicate with the national authorities (Togeby, 2004), but it is also a prerequisite of economic and psychological wellbeing in a foreign country (Chiswick & Miller, 1995; Ting-Toomey, 2011). Because ways of speaking often become a way through which people are identified and judged, and through which legitimacies and illegitimacy, inclusion and exclusion, and group identities are symbolized (Aikhenvald, 2002; Blackledge, 2002; Piller, 2016; Schwalbe 2015; Woolard, 1998), a lack of necessary language skills may lead to negative categorization, exclusion, and ethnic segregation (Ting-Toomey, 2011). Moreover, negative attitudes towards certain ways of speaking may cause a feeling of insecurity, unworthiness, and low self-esteem. This may block cognitive ability of the exposed, leading to lack of motivation, participation refusal and self-exclusion (Collier & Thomas, 2001).

In this research project, we shall focus on how imaginations of language (and “justice”) affect speakers' perceptions of each other and their sense of belonging. Since non-verbal communication affects individual's sense of belonging – it is through non-verbal cues that groups' identities and relations are
symbolized (studies have in fact claimed that between 65 and 80% of the message is communicated non-verbally, cf. Hall 1959; Ting-Toomey, 2011) – we also attend the non-verbal "rules" of behaviour that at play when Greenlandic patients meet the Danish SVPs.

Methods ethnographic and linguistic Our focus is on situated interaction, i.e. what people actually do and say or not say, how they say it, when and under what circumstances. We are interested in both the linguistic (language proficiencies, conversational strategies, styles of communication, turn-taking) and the non-verbal (body positioning, gestures, facial expressions, eye contact, smile, tone of voice, intonation, irony, silence and restrain) aspects of communication. We have identified 5 priority sub-topics: (1) turn-taking (2) silence (3) trust and restrain (4) linguistic insecurity (5) directness vs. indirectness.

We draw on methods and findings of interactional linguistics (Schegloff, 1990; Gumperz 1982) and ethnographic (Hymes, 1962; 1964; Basso, 1970) sociolinguistics, cultural studies (Goffman, 1974; 1970) sociolinguistics, cultural studies (Goffman, 1982; Hall, 1997; (Riggins, 1997)), and intercultural communication theory (Hall, 1959; Johnstone, 1989, Meyer 2014). The analysis builds on the work of Aagaard (2014) on patients' cultural perspective on healthcare in Greenland, Curtis (2001) and Kaufert & O'Neil (1998) work on the role of interpreters in clinical encounters. When developing the idea of 'language economies', i.e. systems of demand and supply displayed in group's dynamics, we draw on Bourdieu (1991) notion of linguistic capital and on Briggs (2005) idea of communicability, i.e. “the productive relation between discursive ideologies or practices and social relations” as a way of understanding how diversity is managed within a society.

Research plan
In order to achieve our academic, societal, and practical goals, the research will be organised in three phases: (P1) ethical approvals, research planning, and data-collection and analysis; (P2) comparative data-analysis and data-collection (P3) a practice-development and dissemination phase. P1 will be divided into four work-packages:

- WP1: Greenlanders' encounters with the Danish healthcare professionals
- WP2: Greenlanders encounters with welfare professionals
- WP3: Greenlandic way of talking a comparative perspective
- WP4: Practice development

WP 1: Greenlanders' encounters with the Danish Healthcare professionals (University of Copenhagen (UCPH) & Copenhagen University Hospital) This research will be conducted by principal investigator Daria Schwalbe (DS), in collaboration with Unit for Greenlandic Coordination, Copenhagen University Hospital. Following methods of interactional linguistics (Gumperz, 1964/1972) and conversational analysis (Schegloff, 1990; 1964/1972) and conversational analysis (Schegloff, 1991; Gumperz, 1964/1972) and conversational analysis (Schegloff, 1991), we will use a three-fold ethnographic data-collection method: 1) Method of observation will be used when collecting data on clinical encounters (scheduled appointments and patient-conversations) between the Danish healthcare professionals at the Copenhagen University Hospital and the Danish-speaking Greenlanders. To access different types of clinical care pathways and patient-conversations, data-collection will be carried out at three departments at the Copenhagen University Hospital: cardiology, thoracic surgery, and haematology at Copenhagen University Hospital (Rigshospitalet), and spinal cord surgery at Glostrup Hospital. The predominant analytical focus in this type of data collection is on communicative strategies and styles (said and unsaid in a conversation, directness, indirectness, silence and restrain), and non-verbal behaviour as communicative barriers. To investigate both the verbal and non-verbal aspects of communication, clinical encounters will be video recorded. 2) Method of participant-observations will be used to observe patient rounds, nursing encounters, educational group sessions (at Glostrup Hospital) and daily routines of the patients; 3) Methods of semi-structured interviews with patients and healthcare professionals will be conducted to understand how the participants experience and interpret clinical encounters, language proficiencies, and non-verbal behaviour (Spradley, 1997).
fare professionals in a Greenlandic context is often linked to negative emotions, such as shame and distrust. We are thus particularly interested in exploring how emotional economies (“trust”, “shame” and “restrain”), “silence” and “directness” may affect the encounter, and how these dynamics are different from clinical encounters. The research will use similar to WP1 methodology (video recordings, participant-observation and semi-structured interviews). To validate our data, we will conduct 6 focus group interviews with our Greenlandic participants (6-8 people each) with the help of a research assistant. These will be conducted at the House of Greenlanders in Copenhagen, recorded and transcribed.

WP 3: Greenlandic ways of talking (UCHP & The Greenlanders Patient Home in Copenhagen) To understand cultural and conversational differences that are at play, this part will focus on communication in Greenlandic, and more specifically on the turn-taking principle (Tanen, 2012; Steensig, 2011) in Greenlandic. The data will be collected primarily at The Greenlanders’ Patient Home in Copenhagen, as well as include data from earlier investigations (Trondhjem, 2008). The data collected and investigated in this part of the project will be used as a comparative point to the main investigation of the intercultural communication between the Danish SVP and Greenlanders (WP1 and WP2). This part of the project will also shed light on turn-taking in Greenlandic, which has not been investigated before.

WP4: Practice development (UCHP & Copenhagen Business School) Together with CBS Teaching & Learning Development and Support Unit and with public sector partners, DS and NT will develop strategies and educational materials (case studies and educational videos) through which identified barriers may be overcome or reduced in the everyday practice of SVPs.

Organisation of the research The project will be hosted at the Department of Cross-Cultural and Regional Studies, UCHP, and directed by Daria Schwalbe. Another key participant is Naja Trondhjem, associate professor, Greenlandic language specialist, UCHP. The core collaboration partners of the project are The Danish House of Greenlanders in Copenhagen and Copenhagen University Hospital. We use collaborative methodology (Phillips, 2018), engaging our partners in the discussion of the production of knowledge. Project will be part of the Diversity and Difference Platform at CBS, which will allow us to bring organisational perspective into the research.

The project makes use of an Advisory Board (AB) and a Stakeholder Panel (SP). AB consists of Associate Prof. Tenna Jensen, Center for Health Research in the Humanities, UCHP; Dr. Christina Larsen, senior advisor, Center for Peoples Health in Greenland, University of Southern Denmark; Prof. Alex Klinge, CBS; Prof. Louise Jane Phillips, Roskilde University; Prof. Josee Lavoie, Department of Community Health Sciences, University of Manitoba, Canada, and Associate Prof. Katie Cueva, Institute of Social and Economic Research University of Alaska, Anchorage, United States. The board will meet twice a year, during workshops or via videoconferences to discuss and identify productive theoretical questions and empirical findings. The Canadian and American researchers will be used as a testing ground for comparative interpretations and analyses.

SP includes Leise Johnsen, lead manager at The House of Greenlanders in Copenhagen; Pauline Albeck Thomsen, Greenlanders’ coordinator, Copenhagen University College; Aaja Chemnitz Larsen, The Danish Parliament, Inuit Ataqatigiit (IA); and Heidi Conradsen, Project manager at Qoqumut, Esbjerg Municipality. We are in a dialogue with Metropolitan University College and Ilisimatusarfik. The Panel will meet twice a year to secure practical implementation and public outreach.

A full-time research assistant (RA) will be employed for a fixed term of two year to support data integration, international outreach, and practice development activities, including management of international symposium, revision and publication of symposium proceedings into an anthology.

Knowledge Utilisation

Potential

Research Understanding how language and culture work in clinical encounters (and culture is communication, cf. Hall 1959), and how language “difficulties” and non-verbal behaviour may contribute to cultural biases, feeling of social injustice and exclusion is of extreme importance if we want to achieve a higher degree of participation and social inclusion, necessary to secure a more inclusive integration agenda and social cohesion in Denmark.

Implementation

Research We will communicate our results at conferences and in Danish and international peer-reviewed publication channels, including Inuit/Ethudes/Studies, Linguistic Anthropology, and Journal of Communication in Health Care. The articles topics include: 1) “Yes, no, I don’t know”: cultural values in healthcare. 2) Imprisoned in Danish. 3) Trust, shame and restrain in a conversation 4) Silence and turn-taking in Greenlandic. 5) “Keeping quite”: On the value of ‘quiet language’. 6) The negative effect of Danish directness. Proceedings of an international symposium on language and health in Copenhagen will be published in an anthology in English under the title “Affective language economies of health”.
The overall project plan is depicted in the following Gantt chart:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>P 2: data analysis</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P 3: practice development</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP 1 (DS)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP 2 (DS, NT &amp; RA)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP 3 (NT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WP 4 (DS &amp; NT)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research visits*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conferences &amp; symp.*</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dissemination</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>


Figure 1  Overall project plan

with contribution from international researchers (7). Educational materials (videos and case studies) for a broad pedagogical environment from university colleges to universities will be developed in the final stage of the project (WP4). We intend to establish Communication, Equality and Health Platform, and to offer interactive workshops for SVPs at the University Hospital and Municipalities across Denmark, and contribute with a feature article in a Danish and international newspaper.

Cost estimates
Total budget requested
4,284,188, excl. overheads

Intended starting date
1.08.2019 (The project shall run for 3 1/2 years. Estimated end date is 31.01.2023)

Application for additional grants
None

Data Management Plan
We plan to record 12-14 clinical encounters at Copenhagen University Hospital and 10 conversations with welfare professionals, divided between Copenhagen and Esbjerg; 6 additional encounters will be collected by NT at The Greenlandic Patient Home in Copenhagen. All encounters will be video recorded. Ethnographic interviews with the participants and SVPs (approx. 30) and (8-10) focus group interviews will be recorded on digital device. All data will be transcribed, classified, and stored in a ‘closed-access’ database with the help of student assistants.

Ethics
The research deals with sensitive patient data. Prior to the research, we will seek ethical approval by sending application to The Danish Council of Ethics and The Committee of Research Ethics in Greenland. Management of the Data till follow Anthropological Ethical guidelines and include written consent agreement and anonymity of data. All data will be transcribed, classified and stored in a ‘closed-access’ secure database.

Society
Public summary
Danish Med dette projekt undersøger vi de kommunikative processer, samtalestrategier og sproglige udfordringer, som gør sig gældende, når grønlandsk-dansk talende tilflyttere fra Grønland møder det danske sundheds- og velfærdspersonale. Skont de indfødte grønlændere er danske statsborgere, føler mange af dem der flytter til Danmark, at det er vanskeligt at tilpasse sig det danske samfund og samtalekultur. Ifølge grønlændere, snakker danskerne meget og de snakker højt, de er meget direkte samtidig med at deres ordvalg og tone til tider kan være grov. Derfor kan det som grønlændere være svært at komme til orde, og man får ikke danske venner, før man er megasocial, hvorfor mange føler sig isoleret og blive ensomme. Det er især i mødet med det offentlig system kommuner og sundhedsvæsen at udfordringerne er størst. Her bliver grønlændere ofte opfattet som vanskelige, netop på grund af de
The project presents an integrative approach to a societal problem, the inequality of health care for Greenlanders in Denmark. The combination of linguistic performance and non-verbal behaviour is an important step towards a more holistic understanding of a multi-faceted problem. There is an extensive integration of stakeholders outside the PI’s own research environment, both within health-care and the Greenlandic community. The research questions are clear and well connected to the focus on situated practices. The dissemination includes non-academic instances Weaknesses: There is a risk of essentializing culture, positioning it as a given “natural” precondition, and of creating a negative presupposition that there (always) are problems in the encounter between Greenlanders and Danish health staff, and that the problems are situated on one side of the encounter. There is no multimodal expertise though the project includes body language.

**Grade 6**

Strength Both the PI and other researchers have good qualifications in relation to Greenlandic/Inuit languages and culture. Both applicants have a solid experience in managing academic research and administration. There is a large amount of hands-on experience.

Weaknesses The project staff does not include expertise on body language and multimodal communication.

**Overall impression**

**Conclusion**

A well formulated project addressing a societal problem in a nonreductive manner, combining practical expertise with theoretical knowledge. And yet, there is a risk of essentializing culture, positioning it as a given “natural” precondition, and of creating a negative presupposition.

*Grade 5: A very good proposal demonstrating good quality. It generally meets scientific standards, but some points can be improved.

**Grade 6: The proposal is internationally excellent and it meets all scientific standards and excels some of these.**

**Rebuttal**

None

**Decision**

English translation Not granted

The Council thanks you for your application and regrets to announce that you have not been granted a grant. In the assessment of your application, the Council has weighted it against the other applications it was in competition with. Your application was in the top approx. third of the field, and only minor conditions were decisive for its failure to achieve grant. The reasons for the refusal below must be seen in this light.

The Council’s consideration was based on the assessment criteria of the call, and the reasons for the refusal reflect which criteria were given heavy weight in the decision to refuse.

The council finds that the project has a high societal value. In the assessment of your application,
however, the Council placed particular emphasis on:
that the project does not seem to be very well oriented in the field medical anthropology.
that the project primarily sees the problem as communicative, but only involves modest extent further anthropological and sociological aspects of it communicative.
that only one Greenlandic-speaking researcher has been associated for 7 months, while the duration of the project is significantly longer. PI itself indicates its competencies in Greenlandic as being at a level that will hardly be enough implementation of the project. The Council's decision has been taken on the basis of your application and the assessment made by the external reviewers. In this context, it should be noted that external assessments serve as an extension of the Council's decision – and are for the guidance only. The Council makes the final decision based on its own assessment of the applications and the prioritization of the overall field of application.

References


Neurological Markers of Maladaptive Brain Activity in Fibromyalgia and their Relationship with Treatment Effectiveness

Elia Valentini

Chronic pain (CP) is estimated to affect at least one-third of the population in the United Kingdom. Fibromyalgia (FM) is one of the most disabling CP conditions. Epidemiological research suggests its global prevalence to be between 2-8%. The unknown pathogenesis, lack of biological markers to monitor its development, and lack of successful treatment make FM a crucial target of pre-clinical research.

The goal of this project is twofold. The project aims to 1) identify robust neurological markers (i.e., electrochemical brain activity) by applying a combination of advanced electroencephalography (EEG) signal processing (i.e., functional connectivity of oscillatory activity) and neuroinflammatory (NI) responses (i.e., estimation of pro-inflammatory cytokines intake), through which 2) characterizing successfully and unsuccessfully treated FM patients (compared to age-matched healthy controls). These measures, seldom combined, have been successfully applied to the study of psychiatric conditions and sleep. Crucially, the identification of neurological markers at rest and during arousing sensory stimulation will allow us to estimate the relationship between these neurological markers and treatment effectiveness. This proposal is important because it aims to generate a robust pre-clinical neurological tool to identify FM and its relationship with measures of treatment effectiveness. The successful identification of neurological markers will improve the assessment of the development of maladaptive changes in FM and will kick-start further research on treatment effectiveness.

This project is of great medical relevance as it will identify pathological signatures of FM that can then inform research on etiology and treatment of this condition.

Keywords: fibromyalgia, pain, EEG, cytokines, neuroinflammation, brain

Research proposal

The institution of employment at time of application
University of Essex

Prospective host institution
University of Essex

Main field
Neuroscience

Other fields
Pain

Overall aim and key objectives

Pain is a major burden for our society. In the United Kingdom, chronic pain is estimated to affect between one-third and one-half of the population. This impressive figure is likely to increase with the progressive ageing of the population [1]. A recent survey revealed that 19% of the European population have experienced moderate or severe pain for at least six months duration and at least twice a week. About 40% of these people report insufficient pain treatment, both pharmacological and non-pharmacological [2].

Fibromyalgia (FM) is a chronic pain syndrome affecting the central, autonomic, endocrine, and immune systems, often leading to an intense state of suffering and disability. Epidemiological research suggests its global prevalence to be between 2-8% with a predominant female/male ratio [3]. FM diagnosis is currently accomplished through clinical assessment, as no validated biological markers have been identified. Its diagnosis relies on the assessment of subjective dimensions such as pain, fatigue, and sleep quality using the Symptom Severity (SS) and Widespread Pain Index (WPI) [4]. As a result, there is large variability between diagnoses [5]. Research suggests that central nervous system sensitization is considered the hallmark of the pathological trajectory of FM, leading to an increased gain in pain and sensory processing [6].

Neuroimaging studies using functional magnetic resonance imaging show altered brain communication between cerebral regions in FM patients,
namely intrinsic resting state connectivity [7, 8]. Further evidence points to inflammatory mechanisms triggered by the activation of mast cells [9]. Current targets of investigation in FM are IL-1, IL-6, and IL-8 and the tumor necrosis factor (TNF) [10]. Despite increased investigation into the relationship between increase in cytokines synthesis and increased excitability of the brain [11], no research project has studied them in combination with EEG activity in FM.

Importantly, many FM patients report dissatisfaction with the available treatments [12]. Management strategies for FM vary and the effect size of most treatments is modest [13]. Therefore, not only is there a lack of diagnostic tools to effectively identify FM, but there is also high variability and limited effectiveness of its treatment. By “treatment effectiveness”, the project refers to how beneficial a FM treatment is in everyday life compared to no treatment, assessed by a series of outcome measures, such as pain, fatigue, mood and anxiety, sleep quality, daily functioning, and well-being [14]. An effective treatment is one that is likely to reduce/improve the aforementioned symptoms. It follows that there may be substantial differences in brain neuroinflammatory (NI) correlates between healthy and FM individuals, as well as between FM individuals who are experiencing an effective treatment compared to those who have not had a successful therapy. Neither issue has received enough investigation so far.

Research plan

Design The project addresses the following question: Can we exploit functional connectivity of oscillatory EEG activity and pro-inflammatory cytokines intake as pre-clinical neurological markers and treatment effectiveness in FM? In other words, this project aims to extend existing knowledge on abnormal brain processing and enhanced NI activity in FM. Another element of this project's originality comes from our attempt to provide a characterization of neurological patterns in FM individuals undergoing effective vs. ineffective treatment (compared to age-matched healthy controls). Thus, there will be two specific objectives: 1) to identify distinct EEG and NI markers of FM; 2) to characterize the EEG and NI profile associated with treatment effectiveness in FM. FM patients will be assigned to two different outcome groups based on scores in pain, fatigue, mood and anxiety, sleep quality, daily functioning, and well-being. The project will then measure their neurological trajectories prior to and 3 months after the start of their treatment (see tasks in Fig. 1 timeline) using a combination of advanced EEG signal processing and NI responses. These measures, seldom combined, have been successfully applied to the study of psychiatric conditions and sleep. In the meantime, EEG responses will also be recorded in a group of healthy age-matched individuals. Therefore, the research design will consist of three groups: FM patients with effective treatment (Feff), FM patients with ineffective treatment (Fineff), and age-matched healthy controls (Ctrl). Treatment effectiveness will be assessed based on reductions in current symptoms compared to the pre-treatment. By doing so, the project essentially asks whether discriminatory patterns of EEG and NI can be detected not only between Ctrl and FM patients, but also between Feff and Fineff.

It is estimated that at least 20 participants per each experimental group will enter the study (in line with a power estimate based on previous neuroimaging studies). Participants will be all female (because of the large predominance of the disease in the female population) and selected on the basis of stringent exclusion criteria [7]. The control group of healthy age-matched individuals (n=30) will outnumber the experimental groups to increase the accuracy of the findings.

Methodology

Self report measures The main criterion for participation will be a diagnosis of FM and the final score in the SS and WPI. Participants who do not meet the inclusion criterion of WPI > 7 and SS > 5 or WPI between 3-6 and SS > 9 will be not allowed to enter the study, as they would not clinically classify as FM patients. Patients who enter the study will then be submitted to a series of measures. The project will use the Fibromyalgia Impact Questionnaire (FIQ), which is a multidimensional self-administered instrument that covers the following dimensions: physical functioning (11 items), well-being (1 item), work situation (2 items), pain (1 item), fatigue/sleep (2 items), stiffness (2 items), and psychological symptoms (2 items). The Patient Health Questionnaire (PHQ-9) will be used to investigate current disposition to depressive symptoms [15] and the General Anxiety Disorder questionnaire [16] will be used to assess disposition to anxiety. In addition, participants will complete the Pain Catastrophizing Scale (PCS) [17], a questionnaire that measures the cognitive/emotional attitude towards pain, and the Warwick-Edinburgh Mental Well-being Scale (WEMWBS) [18] to quantify well-being.

EEG recording and data analysis I have extensive experience in EEG techniques and pain research [19, 20]. The project will take advantage of my collaboration with Fibromyalgia Research UK to recruit FM patients and with the molecular biology lab led by Dr. Metodiev to extract cytokines from participants’ blood samples and quantify them for each participant. As FM is known to be associated with abnormal processing in both rest and stimulation states [6], EEG activity will be measured at rest and
The scientific objectives of the project will be achieved through 4 Tasks (T) and verified at 4 Milestones (M). The first month of the project will involve the PI and the PDRA in collecting relevant clinical data and self-report questionnaire scores of Fibromyalgia (FM) patients who just started therapy (T1, red). These will be collected again after 3 months to determine which of the patients are benefiting from their therapy (i.e., showing reduction in symptoms). Based on this assessment, participants will be assigned to either the effective treatment (Feff) group, or to the ineffective treatment (Fineff) group (M1). Both groups will undergo a first experimental session (T2.1, green) whereby symptoms will be recorded, blood sampling for extraction of neuroinflammatory (NI) indexes will be performed, and electroencephalography (EEG) data will be obtained at month 10 (M2). Symptoms and EEG measurements will also be collected in age-matched healthy controls (Ctrl) which will start in parallel with T2.1 (T3, orange) but will proceed at lower priority and end at month 12. A minimum lapse of 3 months will be interleaved before a FM patient goes through a second session of symptoms, NI, and EEG data collection (T2.2, light green), wherein the same staff will be involved. T2.2 will initially be paced slowly because of the overlapping T2.1 and T3. The last task (T4, blue) will start at month 16 and will consist of statistical analysis of self-report, NI, and EEG data for the three groups, and preparation of the findings for dissemination in the shape of conference presentations and scientific articles (M4). T4 will involve the PDRA, the PI and the scientific collaborator (Dr. Metodiev).

**Figure 1  Project timeline.**

Cytokines extraction and analysis. Cytokines IL-1, IL-6, and IL-8 and TNF will be obtained through blood sampling before each EEG session. Symptoms will be collected before, during, and after the experimental session. In particular, the project will analyze the relationship between self-report measures (e.g., pain), EEG magnitude, and connectivity in different frequencies and cytokines.

Commercial enzyme-linked immuno-sorbent assay (ELISA) kits will be used to quantify the cytokines according to manufacturers’ instructions. An example is the IL-1 kit from Thermo Fisher. ELISA is a sensitive and robust method for quantitative analysis of low-level biomarkers in blood and plasma samples. More specifically, the project will use the sandwich ELISA method, which uses two antigen-specific antibodies—one to capture the target molecule and the other to generate the signal for quantification. Commercial sandwich ELISA kits are available for all of the proposed cytokines. They are well-validated and will provide the necessary level of sensitivity and specificity for the study. The cytokine analyses will be performed in Dr. Metodiev’s lab, which is equipped with all necessary instruments and staffed by personnel experienced in biochemical analyses and proteomics.

**Knowledge utilization**

The successful identification of neurological markers will pave the way for experimental studies aimed at disentangling the effects of both pharmacological and non-pharmacological treatments on these neurological measures. The novel combination of EEG and NI measures in this study will improve the assessment of development of maladaptive changes in FM and inform research on treatment effectiveness. The output of this research will be published in reputable journals such as e.g., Cortex, eLife, Frontiers in Neuroscience, Journal of Neuroinflammation, NeuroImage, and Pain.
Figure 2  Functional connectivity analysis. Example of a preliminary test of connectivity analysis on a sample of 20 healthy individuals administered with the somatosensory repeated stimulation mentioned in the current project. The left plot in the upper panel shows the frequency spectrum of EEG power (1-25 Hz) across single trials in the sample. The plot on the right provides independent components (as specified by the independent component analysis) of the EEG power at 6 Hz. This activity is well-known to represent the vertex EEG response commonly elicited by salient sensory stimuli. The left plot in the centre panel shows the main result of the functional connectivity analysis (as performed by EEGLAB). The normalized direct transfer function (nDTF) reveals an increased connectivity from the first component (IC1) to the fourth component (IC4) in the target frequency (peaking at 6 Hz and ≈ 500 ms poststimulus relative to baseline). Dipole modelling of the ICs (centre right) localizes IC1 and IC4 at Brodmann areas 6 (premotor or supplementary motor area, green dipole) and 24 (medial portion of the anterior cingulate cortex). The bottom panel figure extends this analysis by indicating an inflow from IC4 to IC1 as well as reduced flow between these two ICs and the second IC located in the prefrontal region. This pattern takes place right before the arrival of the somatosensory stimulus. Altogether, these preliminary findings hint at a potential twofold pattern of pre-post stimulus connectivity between these regions that could be considered as a small world network for the identification of a purported neurological marker in fibromyalgia patients.
Figure 2 (Cont.) The left plot in the centre panel shows the main result of the functional connectivity analysis (as performed by EEGLAB): The normalized direct transfer function (nDTF) reveals an increased connectivity from the first component (IC1) to the fourth component (IC4) in the target frequency (peaking at 6 Hz and \( \approx 500 \) ms poststimulus relative to baseline). Dipole modelling of the ICs (centre right) localizes IC1 and IC4 at Brodmann areas 6 (premotor or supplementary motor area, green dipole) and 24 (medial portion of the anterior cingulate cortex). The bottom panel figure extends this analysis by indicating an inflow from IC4 to IC1 as well as reduced flow between these two ICs and the second IC located in the prefrontal region. This pattern takes place right before the arrival of the somatosensory stimulus. Altogether, these preliminary findings hint at a potential twofold pattern of pre-post stimulus connectivity between these regions that could be considered as a small world network for the identification of a purported neurological marker in fibromyalgia patients.

Cost estimates

*Total budget requested*

£99,555.00

*Intended starting date*

29/5/2019

*Application for additional grants*

This application has been revised and resubmitted to the Wellcome Trust Seed Grant Round 2019 (but no feedback is provided in that scheme).

Data management plan

The project will generate different types of data that can be measured on ordinal and interval scales, and in different numerical formats. Self-report data will be electronically collected through specific software (e.g., E-prime, Qualtrics) and exported to statistics software format (.sav in SPSS or .csv for more sophisticated analyses in R open source software). Other data formats include .cnt (Neuroscan EEG recording) and .set (EEGlab software) for processing of EEG activity. Cytokine concentration will be recorded from readings of absorbance (in nanometers) and stored in digital format. They will be of value for other researchers in the field of pain, particularly those with a background in clinical neuroscience. Data quality will be checked to ensure that the scientific community can be warned about any missing, excluded, or compromised data. The post-doctoral researcher and all technical staff will receive detailed guidance in best practices of task administration, data collection, and storage to guarantee that data is of consistent high quality. Dr. Metodiev and I will ultimately be responsible for ensuring the proper management of the data; we will supervise all computations and verify statistics to ensure accuracy. Data will be backed-up on a cloud server (e.g., OneDrive) in compliance with the European Data Protection Act. The University of Essex's network is secure and protected from viruses and intrusions by a firewall and the Sophos anti-virus program. The consent forms will be stored and locked in our offices at the University. Data will then be shared upon completion of data collection and before data analysis. As I previously mentioned, data will be made available on the Open Science Framework (OSF) as well as on the University of Essex Research Repository (http://repository.essex.ac.uk/), so they will be openly accessible to other researchers. Individual data will be anonymized and stored on password-protected computers.

Ethical aspects

I have had 3 different ethics applications being approved by the University of Essex so far and all of them implied the use of nociceptive stimulation and induction of tolerable pain (EV1701, EV1501, EV1801). The ethics approval procedure entails the submission of different forms to the departmental Ethics Officer. Applications will first be assessed by the Ethics Officer before being passed on to the University’s Ethics Committee. A copy of my research proposal and any necessary supporting documentation (e.g., consent form, recruiting materials, etc.) will be attached and will be assessed by the Research Governance and Planning Manager. Being familiar with this procedure, I expect to submit the ethics application for the present project at the end of the current year and for the approval process to take no longer than two months. Therefore, I foresee receiving full approval by the end of March 2019.

*Declarations*

Head of department’s statement of support I fully support this application. Fibromyalgia is a very poorly understood, under-researched condition with a huge impact on quality of life. Understanding the organic underpinnings of this chronic pain condition is important for both establishing the nature of the conditions and developing interventions. Dr. Valentini’s expertise in pain will provide vital insights into the biomarkers of this condition by combining EEG and cytokine analysis. His expertise, which he will continue to develop through this project, will position him excellently for this. He is unusual in combining a strong physiological background with a well-rounded understanding of the person as a whole, and the subjective experience in this condition of chronic pain. Elia has recently become a permanent member of the department, having established his research here very successfully, combined with being an excellent teacher and valuable colleague. The university has invested significantly in the infrastructure required for research in pain,
and in the appropriate EEG techniques. Elia has matched this with his own investment in getting this facility up and running safely and successfully and securing funding for a PhD student. This continued investment will diversify his skills and represent long-term commitment to establishing his excellent research program and academic career.

Society
Public summary

Lay summary Fibromyalgia (FM) is one of the most disabling chronic pain conditions. Science is still struggling to understand how it develops and how to treat it. More research on the basic mechanisms of this disease is needed. We currently know that FM patients show imbalanced brain activity and produce a range of neural chemicals associated with response to stress and inflammation, but we do not know how to use this information to discriminate between patients who positively respond to treatment and those who do not. The lack of effective treatments has led clinical practitioners to call for more research addressing the development of the disease and the effectiveness of treatment.

The goal of this project is to identify robust neurological measures of FM (i.e., electrochemical brain activity) by applying a combination of advanced analyses of electrical brain activity and inflammatory brain responses in FM patients who experience successful vs. unsuccessful treatment (compared to healthy individuals). The University of Essex is perfectly suited in terms of expertise and facilities to identify the electrical brain responses during rest and sensory stimulation, as well as measuring chemical stress responses at the same time.

This project is important because the successful identification of specific cerebral measures of FM can improve the assessment of the development of this disease and pave the way to study its relationship with measures of treatment effectiveness. This research could significantly impact patients’ prospects of better identification and treatment of this chronic pain condition.

Reviews
See supplemental file 1.

Rebuttal
Not available

Decision
See supplemental file 2.

References


