Mapping the Crises Literature in psychology:

What is the crisis in psychological science?

Sven Ulpts

S3300854

Master Thesis

Theory and History of Psychology

University of Groningen

First supervisor: Jeremy Burman

Second supervisor: Maarten Derksen

Date: 26.08.2022

Mapping the Crises Literature in psychology:

What is the crisis in psychological science?

Abstract

Since 2011 numerous crisis declarations have been circulating in the psychological literature. Claims are ranging from inadequate measurement and statistical practices to overgeneralized conclusions and improper theories. Therefore, the question emerges, what does the currently claimed crisis state in psychological science actually mean? Or in other words what is wrong with psychological research, according to psychological researchers? A second question is whether researchers who see theory, replication or measurement respectively as problematic agree or disagree regarding what else is problematic in the field. To answer these questions in study one themes that are related to the crisis discussion in the literature and are linked to problems in the discipline were extracted from a systematically compiled literature sample. Subsequently, star graphs were constructed that show the occurrence and co-occurrence of identified problems in psychological science to visualize how the crisis state is represented in the literature. The results indicate that there seems to be somewhat of an agreement regarding what is wrong with psychological science. The theme graphs provide us a proxy to see what the crisis state and the replication crisis mean, regarding problems in the discipline. However, study 1 does not allow inferences about which problems are more important or dangerous to the discipline, according to psychological scientists, nor does it provide an accurate representation of what the theory and measurement crisis mean according to the literature. In the second study a crisis state network was created using the APA Thesaurus of Psychological Index Terms. Study two illustrates shortcomings in the index and recommends caution when using the index for literature searches and representing concept understandings (meanings).

Keywords: Crisis, replication, networks, theory, measurement, crises

Mapping the Crises Literature in psychology:

What is the crisis in psychological science?

Psychological science seems to be in a crisis state since at least 2011. Unfortunately, there seems to be no agreement about what this crisis actually is or means. There are numerous different crisis declaration circulating in the literature. There is for instance the theory crisis (Muthukrishna & Henrich, 2019; Oberauer & Lewandwoski, 2019), the validation crisis (Schimmack, 2021), the inference crisis (Starns et al., 2019), the statistical crisis (Gelman & Loekn, 2014), the generalizability crisis (Yarkoni, 2022), the practicality crisis (Berkman & Wilson, 2021), the replication crisis (Earp & Trafimow, 2015) and the measurement crisis (see e.g. Lilienfield & Strother, 2020).[1] Moreover, crisis discussions in psychology are nothing new, but rather a reoccurring theme in the literature throughout the history of psychological science (Sturm & Mülberger, 2012). The more recent crisis discussions since 2011 seem to have been jumpstarted by, among other things, Diederik Stapel's highly publicized fraud case that created major doubts in the self-correction mechanisms in psychological science and Daryl Bem's (2011) publication of a manuscript that claimed to have found evidence for precognition, with at the time widely accepted statistical practices, in a leading social and personality psychology journal (Stroebe et al., 2012; Wiggins & Christopherson, 2019). A lot of discussion was also fueled in 2015 by the publication of the reproducibility project by the Open Science Collaboration, because it found that less findings replicated than widely expected and that the practice of replication is harder than anticipated (Open Science Collaboration, 2015). Furthermore the rate of publication in science makes it practically impossible for an individual researcher to keep up with the state of the academic literature in psychology (Phaf, 2020). Similarly, with the amount of attention directed at as well as recently also funding allocated to crisis related issues in psychological science and alternative publishing platforms either gaining in usage as a reaction to the crisis, such as PsyArXiv, or new journals being formed, the state of the crisis literature might also be inextricable for the individual scholar. Likewise, the relevance of and the increasing amount of work concerning these issues is at least somewhat evidenced by the uprising of new areas and forms of science that aim to investigate the state of science itself as well as steer it towards a seemingly better state, with quite some attention paid to and work done in psychology. There is, for example, meta-research (see e.g. the Meta-Research Innovation

[1] Sean Devine called the last four crises (replication, measurement, practicality and generalizability) the four horsemen of the crisis in psychological science in an online post on the homepage of the Journal of Trial and Error. However, that post does not exist anymore.

Center at Stanford[2]), science of science (Wang & Barabasi, 2021), research on research (see for example the Research on Research Institute[3]) and open science (see the Center for Open Science[4]).

Importantly, all these different crisis declarations that focus on varying specific issues in psychological science might indicate that, as Jill Morawski put it, psychological science, or at the least its literature, could just be a "polygenic mess" (Morawski, 2019, p. 219). Conversely, it could also be that the literature hides connections and similarities between claimed crises in the abundance of published material. We need to find review methods that allow us to capture and visualize the content (meaning) of the vast literature. A possible way trying to capture the meaning of the crisis state or the representation of issues that are related to those declarations could be to map the occurrence and co-occurrence of themes in the literature that are linked to those crises. In other words, when someone talks about replication being a problem for the discipline such methods could tell us what else that person views as problematic for the discipline? Nelson and colleagues (2021; 2022) performed something similar. However, they did not exclusively concentrate on psychology and focused on the replication crisis. I only focus on what is claimed to go wrong in psychological science while Nelson and colleagues (2021, 2022) focus is more broad with also including the subject of investigation, potential solutions to the replication crisis and the stake of the replication crisis. Similarly Tabea Cornel and Brandon Heil conducted an occurrence and co-occurrence analysis (network) with focus on the replication crisis and open science (CEFISES at UCLouvain, 2021). They also constructed co-author and co-citation networks of scholars who write about the replication crisis and open science. Additionally, there have been other attempts at understanding the crisis and reform literature, for instance with a philosophical and epistemological reading (Derksen, 2019; Flis, 2019; Marowski, 2020). These contributions are highly important for providing a reference and lens that allows to understand the discussions from a certain perspective (e.g. Popper's philosophy of science or indigenous epistemology). However, we also need methods that allow for the representation of meanings that are intended by the original authors, instead of providing a new frame of reference that aids comprehension. Hence, we need methods that are able to tell us what these crisis declarations actually mean. What does replication crisis, theory crisis and measurement crisis actually mean? Are they different kinds of crises with different problems just because they

[2] https://metrics.stanford.edu/
[3] https://researchonresearch.org/
[4] https://www.cos.io/?hsLang=en

have different names? Or are we giving the same thing different names over and over again? To answer these questions I will use methods from the digital humanities to create an illustration for each crisis that represents the themes that are mentioned in the literature in relation to these crises. This methodology is somewhat similar to Burman and colleagues' (2015) study about what self-regulation actually means according to the PsycINFO APA Thesaurus of Psychological Index Terms.

Study 1

Methods

Literature was searched using PsycINFO with the search terms valid* AND crisis, measure* AND crisis, theor* AND crisis, confidence AND crisis, credib* AND crisis, statistic* AND crisis, method* AND crisis, generaliz* AND crisis, replica* AND crisis, reproducib* AND crisis, irreproducib* AND crisis, rigor* AND crisis and explan* AND crisis.[5] The search terms were inspired by Nelson and colleagues (2022). Furthermore google scholar was used to search for citers of the following crisis declarations: *The generalizability crisis* (Yarkoni, 2022), *the Validation crisis in Psychology* (Schimmack, 2021), *A problem in Theory* (Muthukrishna & Henrich, 2019), *Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them* (Flake & Fried, 2020), *Replication, falsification, and the confidence crisis in Social Psychology* (Earp & Trafimow, 2015), *Addressing the Theory crisis in Psychology* (Oberauer & Lewandowsky, 2019) and *So useful as a good theory? The practicality crisis in (social) psychological theory* (Berkman & Wilson, 2021). Subsequently the references of selected articles were scanned for relevant sources. The literature was limited to publications between 2011 and 2020. Sources had to fulfill at least one of the following criteria: authors had to be psychological researchers, the manuscript had to be published in a psychological outlet or the publication had to mention a crisis or focus on problems in psychological science (behavioral sciences). Preprints of articles that were only published after 2020, but were available before, were included (an example would be *The generalizability crisis* by Yarkoni (2022)). This lead to the pre-selection of 784 scientific publications, of which 121 were selected for the analysis (See Appendix 1 for a list of the selected literature). The selected literature is mostly more recent literature, as can be seen in Figure 1. All sources were read at least twice to ensure that the coding did not just represent co-occurrences, but also captured the content and meaning of the

[5] Due to the COVID crisis the search term crisis caused quite the overflow of results which made the literature search more taxing than anticipated.

texts. For instance, just because someone mentioned publication bias does not imply that the author thinks that publication bias is related to a crisis in psychological science or a problem for the discipline. Subsequently, I created themes based on grouping of codes (the coding and theme creation procedure was done similar to a thematic-analysis)[6]. I constructed the crisis graphs similar to Burman and colleagues' (2015) meaning networks. I created excel files for each graph with two columns, a *Source* column and a *Target* column. The Source column always just contains one term for each graph. For the first network it only contains crisis all the way down, for the replication graph it only contains replication, for the theory graph only theory and for the measurement graph only measurement. The Target column for each graph contains the themes that are seen as related to the crisis or as problems in the discipline. For the first crisis state graph the Target column contains the themes that occur in each of the publications. For the other three networks the Target columns respectively contain the themes that co-occur in the publications with replication, theory and measurement. Next I calculated the number of occurrences for the first graph and the number of co-occurrences in the other three graphs (how often does each theme co-occur with replication, theory and measurement respectively). Lastly, I created an excel file containing the citation counts for the publications (according to google scholar on August 17th 2022) that contain the specific themes to get the sum of citations for each theme. The visualization of the themes was done with Gephi 0.9.2.

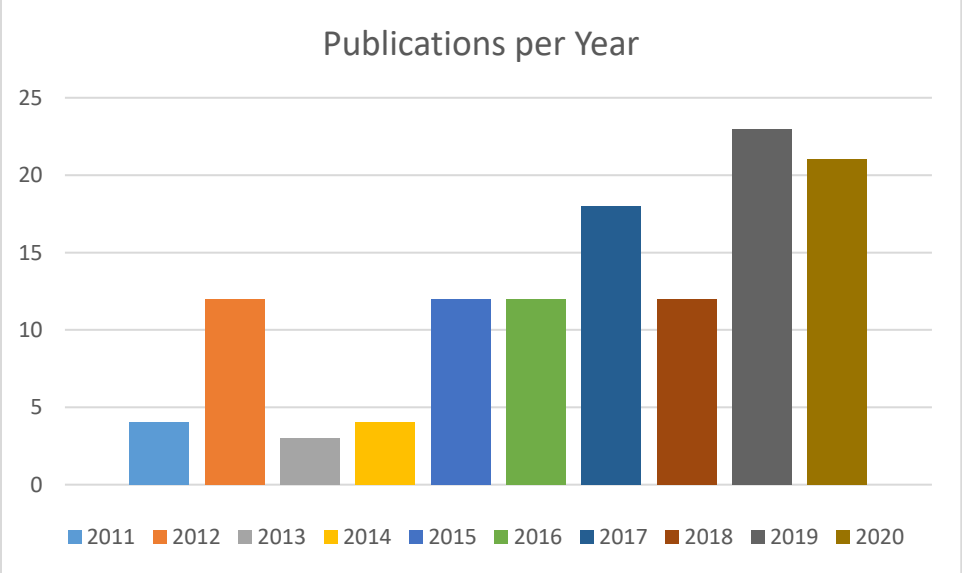


Figure 1. Distribution of Publications over the years between 2011 and 2020 in the literature sample. The distribution is clearly left skewed, meaning that more recent literature is over

[6] See, for instance, Braun & Clarke, 2006.

represented in the literature sample. The drastic rise in 2012 could be interpreted as a reaction to Bem's publication and Stapel's fraud case. The rise since 2015 could be related to the Open Science Collaborations Reproducibility Project.

Results

I identified 33 themes in the literature that are related to a crisis state in psychological science and what potentially goes wrong in psychological research. The number of themes per article varied from one to 18. The ten themes that mostly seem to capture what constitutes the general crisis state in psychological science measured by the number of occurrences are replication (two thirds of the sources mentioned replication), publication bias, questionable research practices (QRP), false positives (Type I errors), theory, incentives, power/ sample size, NHST (null hypothesis significance testing), transparency/ openness (the lack of transparency and openness) and biased researchers.

Theme: Replication

It is often pointed out that reproducibility is a cornerstone of a proper science (Simons, 2014). Numerous potential issues are mentioned in relation to replication, such as, failures to replicate (especially eminent findings in the discipline or large scale failures to replicate), a general lack of replication in psychological science (direct and systematic replications) and a lack of clarity as well as agreement about what constitutes a successful replication or a failed one for that matter (see e. g. Laraway et al., 2019; LeBel & Peters, 2011; Simons, 2014).

Theme: Publication bias

Publication Bias is usually mentioned as the problem that journals, especially high impact journals, prefer to publish novel, original, beautiful, clean, positive, significant and impactful findings (see e.g. Giner-Sorolla, 2012). This of course is closely linked to the lack of replication, because why would someone attempt a replication when no prestigious journal wants to publish such incremental work. Publication bias is also seen as causing a heavily skewed scientific literature in psychology with an abundance of positive (significant) findings. With a large bulk of the literature being hidden away in the file drawer the credibility and usefulness of literature review methods, such as, meta-analyses and systematic reviews become questionable (see e.g Ferguson & Heene, 2012; Rosenthal, 1979).

Theme: QRP

Questionable research practices are no actual scientific misconduct, but practices that

are questionable due to the consequences they have on the reported results. The number one effect that these practices have that is also seen as an important problem in the crisis of psychology is and inflation of false positives (another problem QRP potentially produces are inflated effect sizes). Among those practices are optional stopping, which is the practice of successively adding participants until a significant result is obtained and HARKing (Kerr, 1998), which is hypothesizing after the results are known. HARKing in itself is not a problem, but it becomes questionable and statistically (inferentially) problematic when post hoc explorations are not treated and declared as such (see e.g. John et al., 2012). Questionable research practices are highly linked to the problem of degrees of freedoms (see theme: degrees of freedom). For instance, researchers can run numerous small studies, exploit in each study the degrees of freedom and then selectively report only the significant ones. This is a questionable practice called selective reporting that is linked to a lack of transparency and openness.

Theme: False Positives

False positives are instances when the test indicates that something was found when in actuality that finding is false. In the context of NHST in which most of the false positive discussions are held, it means that the null hypothesis is rejected although it is true (we are ignoring ambient noise or the crud factor here)[7]. False positives in original studies are seen as one of the major causes for the failures to replicate certain findings, because it is thought that the replications do not support certain effects, because they are actually nonexistent (see e.g. Simmons et al., 2012). The problem of false positives is usually also mentioned in combination with QRP. False positives are also seen as a major motivation behind the claim that (direct) replications are necessary, because, according to some, only with replication can potential false positives be identified.

Theme: Theory

The problem in theory is usually claimed to be a general lack of appreciation for theoretical considerations, a lack of a theoretical foundation or overarching theoretical frameworks, that the theories that do exist in psychological science lack precision, empirical content, a clear deductive link between theory and hypothesis and are vague (flexible) which makes them unfalsifiable (see e.g. Fiedler, 2017; Fried, 2020; Szollosi & Donkin, 2019). Falsifiability of theories is linked to the theme of falsification, which is an adoption of

[7] See Orben & Lakens (2020) for a description of these concepts.

Popperian philosophy of science (or a narrow reading thereof) by reformers (Derksen, 2019; Earp & Trafimow, 2015).

Theme: Incentives

The theme of incentives captures the circumstance that in psychological science (or science in general) the system and institutions do not reward scientific (research) quality, but scientists are hired, promoted and get tenure based on quantitative metrics that capture productivity instead of quality (Munafo et al., 2020). Therefore, there seems to be a conflict between what is good for the individual scientist (quantity) and what is healthy for the sciences (quality) (see e.g. Smaldino & McElreath, 2016). Something that is sometimes mentioned in relation to the reward system and its focus on metrics is Goodhart's law, which states that the moment a metric becomes the target itself it loses its intended purpose as a measure.[8]

Theme: Power/ Sample size

Statistical power is the probability of rejecting the null hypothesis in the case that the null hypothesis is actually wrong. The identified problem in psychological science is a general lack of appreciation for statistical power, the lack of proper prior power calculations and chronically low power in the literature (see e.g. Button et al., 2013). Statistical power is also mentioned, because it is linked to the problems of inflated effects sizes and false positive rates. Furthermore, it is claimed that low powered original studies make power calculations for replication studies more problematic and inaccurate (see e.g. Anderson & Maxwell, 2017). Importantly, I chose to call the theme Power/ Sample size, because in the literature when power is mentioned the focus is often only on sample size. However, sample size is just one of the components determining the power of a test. The conceptualization of power in the literature seems to neglect the importance of measurement accuracy (construct validity) and manipulation strength (research design or internal validity).

Theme: NHST

Null hypothesis significance testing is seen as a problem, because it invites binary thinking and the categorization of knowledge according to binary outcomes. Linked to the statistical technique of NHST is also a seemingly discipline wide misuse, misunderstanding and misreporting of NHST (see e.g. Bakker & Wicherts, 2011). Furthermore, it is said that NHST does not provide a strong enough test of two competing hypotheses, but rather the null

[8] Which should make people who think that badges for open practices are a good idea really (re)consider the reasoning behind that thought process.

hypothesis is set up in a way that it functions as a straw man hypothesis (the nil hypothesis), while the alternative hypothesis cannot be supported directly only the null hypothesis can be rejected and the alternative hypothesis is usually underspecified (see e.g. Phaf, 2020).

Theme: Transparency/ Openness

It is expressed that there is a general lack of transparency and openness in psychological science. This is seen as an issue in the field, because of how hard it is to get data from other scientists, that the method sections of manuscripts are usually insufficient descriptions of what was actually done and lack details that would be necessary to know if someone wanted to replicate a study (Stroebe, 2019). Therefore, the lack of transparency and openness is viewed as a potential cause for the lack of replication, falsification and self-correction in science, because it prevents psychological scientists from checking the work of their colleagues. Linked to this is the theme of culture of trust, since the lack of transparency and openness can be interpreted as a symptom of that culture. When scientists blindly trust each other, then why would they be transparent or open about their research? The lack of transparency is also argued to be problematic, because it protects questionable practices such as selective reporting. If researchers had to be transparent about everything they have done then selective reporting would be practically impossible. The perceived lack of transparency and openness in psychology is also evidenced by the open science movement, which attempts to resolve that problem.

Theme: Biased Researchers

It is claimed that scientists as human beings are as biased (have motivated reasoning) like everyone else. The biased researcher, for instance has confirmation bias, which is the tendency to quickly accept results that are in accordance with one's expectations (wishes) and the quick rejection of contradicting claims or results, and hindsight bias, which is the impression that after the fact one has the impression to have known it already beforehand (hindsight bias is also often mentioned in combination with HARKing) (Nuzzo, 2015). Bias is also used as somewhat of a synonym for human subjectivity and methods as well statistics or science as a whole is seen as a sort of antidote against this subjectivity and as a protector and producer of objectivity (Munafo et al., 2017). [9]

[9] See also mechanical objectivity in Porter, 1995 & Davidson, 2018. One might argue that the reform movement is mostly also an attempt to rid science of the human factor and cleanse the manuscripts from the subjectivity of the researcher by letting more and more standardized methods guide the process of knowledge production as well as dissemination.

Hence, six of the ten most mentioned themes are related to methodology and statistics, hinting towards the notion that according to the literature (between 2011 and 2020) the crisis state seems to be mostly understood by psychological scientists as a methodological and statistical phenomenon.

Interestingly, theory, power/ sample size and NHST are not among the top ten most cited themes. However, effect sizes, competition and degrees of freedom are among the ten most cited themes, which are not among the ten most mentioned themes.

Theme: Degrees of freedom

Degrees of freedom is the ability of the researcher to make numerous different, but justifiable decisions during the research process each with different (small) effects on the outcomes (see e.g. Simmons et al., 2011). This problem is also known as, among other terms, analytic flexibility and the garden of forking paths (Gelman & Loken, 2013). Degrees of freedom are not only limited to statistical analysis, but are prevalent in all aspects of research (see for instance Flake & Fried, 2020 for an elaboration on degrees of freedom in measurement practices and questionable measurement practices). The problem of degrees of freedom is exacerbated by a lack of transparency and openness, which together can bring about the problem of QRP. If one includes motivation in the consideration, then incentives and competition are themes that also have to be mentioned in the conversation about how degrees of freedom can lead to QRP.

Theme: Effect sizes

The issue of effect sizes comprises, among other things, of the overestimation of effect sizes in psychological science due to a lack of statistical power in studies, the generally low effect sizes that actually exist in psychological science and the decline effect, which is the circumstance that effect sizes are apparently decreasing over time (see e.g. Gong & Jiao, 2019). Effect sizes are viewed as so important, because they are necessary for proper calculations of statistical power. Furthermore, an underappreciated problem that is related to the issue of effect sizes is the widespread lack of insightful interpretations and expert judgments about effect sizes in psychological science (see e.g. Davidson, 2018).

Theme: Competition

This theme is about the social structure of science. The issue of competition is, simply put, about the apparent situation in psychological science that researchers instead of working

together and collaborating to investigate or solve a problem they work in isolation and against each other. The problem of completion is linked to the problem of incentives and the perverse incentive structure which rewards productivity. Researchers seem to race against each other for publication, jobs and tenure. This competition between scientists is also seen as a contributing factor to the widespread use of questionable research practices as a means to stay ahead of the others (see e.g. Smaldino & McElreath, 2016). Furthermore, this competitive nature of science is also linked to the identified lack of transparency and openness, because when someone is competing with other scientists why would they share their insights and data with the wider scientific community?

See Figure 2 for a map of the representation of the crisis state in (a part of) the psychological literature. The size of the nodes (themes) in the graph is based on the number of occurrences in the literature. The color of the nodes is based on the sum of citations of the publications that contain a certain theme. Hence, the color goes with increasing number of citations from red to blue. In all the star graphs of study 1 the proximity of the nodes to each other has no meaning, the nodes are positioned the way they are to prevent overlap. Interestingly, generalizability (on the left of QRP in Figure 2), which also has its own crisis declaration dedicated to itself, came up in 23 publications as a problem for the discipline and in relation to the crisis state in psychological science (with a sum citation of 2890). According to the proponents of the generalizability crisis, the generalizations we make in our conclusions are not justified by the research practices (samples, methods and statistics) we employ (Yarkoni, 2022). When we look at the crisis graph in Figure 2 and focus on the largest nodes (the themes that occur the most in the literature) we can see that the general crisis state in psychological science seems to be mostly understood by psychological scientists as or in reference to problems with replication, publication bias, QRP, false positives, theory and incentives. However, when we focus on the blue nodes, which are the most cited themes, we can see that QRP (38232 citations), publication bias (38227 citations), replication (37993 citations), false positives (37255 citations), power/ sample size (32535 citations) and incentives (34411 citations) are cited the most. Such high citation counts indicate that these topics are at least acknowledged as issues that merit discussion and conversation. Citation counts do not provide any information about whether the citers see the themes as problematic for the discipline or not. Additionally, for the interpretation of the node color it is important to consider the impact of highly cited articles and the themes that were allocated to them. The three most cited articles in the literature sample were *Estimating the reproducibility of psychological science* by the Open Science Collaboration (2015) with 6865 citations, *Power*

failure by Button and colleagues (2013) with 6702 citations and *False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant* by Simmons and colleagues (2011) with 6552 citations. All of them contain the six most cited themes. These high citation counts carry even more weight once one considers that the average citation count of a publication in the sample is 360 and the average citation count of a theme is 13463. Citation behavior in science is also too complex and potentially problematic to warrant a straight forward interpretation (Horbach et al., 2021).

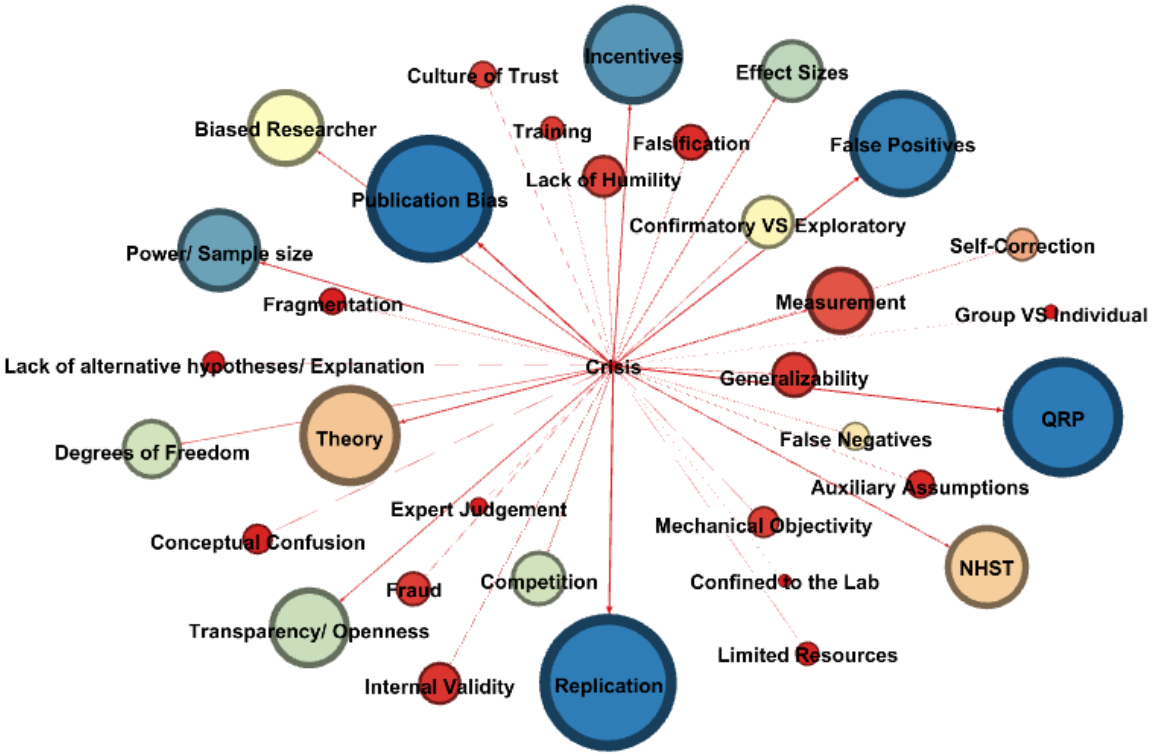


Figure 2. Star graph of what constitutes the crisis in psychological science according to the literature. The node size is determined by the number of occurrences of the themes and the node color goes from red to blue with increasing citation count.

Followingly, I constructed a graph that shows which themes in the literature are most connected to replication (Figure 3). For that purpose I operationalized connection as being mentioned in the same publication with replication. The color of the nodes goes with increasing number of co-occurrences from pink to green. The only theme that co-occurs with replication that is not among the ten most mentioned themes (, but among the most cited) in the crisis literature sample is degrees of freedom. This might be explained by the circumstance that degrees of freedoms are often mentioned in combination with QRPs and

seen as providing a playground for psychological scientists that allows them to make numerable differing choices during the analysis process. This star graph can also be interpreted as a representation of how the replication crisis is understood in the psychological literature, because replication became the focus of discussion pretty early on. When we look at the replication graph in Figure 3 and focus on the large green (green-grey) nodes we can see that the problems with replication seem to be largely understood in reference to publication bias, QRP, false positives, incentives, transparency/ openness (the lack thereof), biased researcher, power/ sample size, theory and NHST.

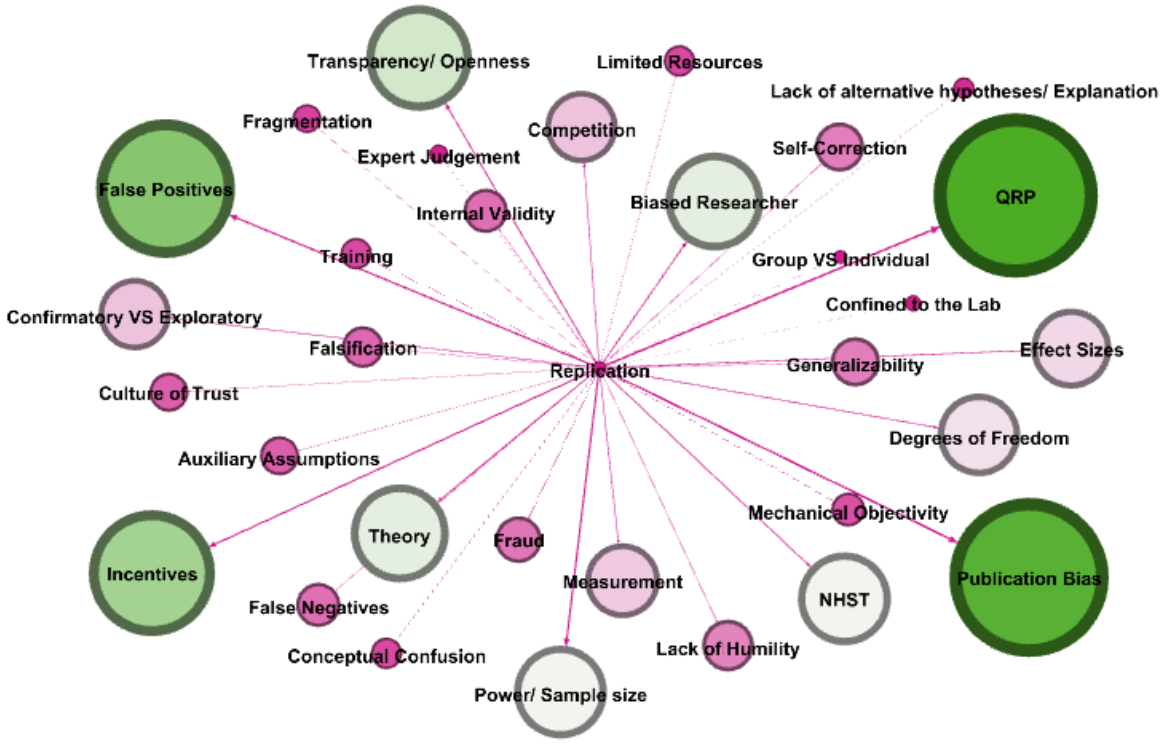


Figure 3. Replication graph showing which themes co-occur in the literature with replication. Node size and color determined by the co-occurrence count.

Afterwards, I made a star graph that represents the themes that are connected to theory in the crisis literature sample. This could be used as a proxy to see where the problem in theory is positioned n psychological science. This problem is usually mentioned under the heading of the theory crisis. Therefore, the purpose of constructing the theory graph was also to see whether it could be used to represent the meaning of theory crisis in (the sample of) the psychological literature and how the problem in theory is understood by psychological scientists (the authors). The theory graph in Figure 4 shows which themes are most connected

to theory according to the same operationalization as above. The color of the nodes goes with increasing number of occurrences from orange to purple. It can be seen in the theory graph (Figure 4) by focusing on the large purple nodes that authors who identified a problem in theory also see publication bias, QRP, replication, incentives, power/ sample size NHST and false positives as problematic for the disciplines. For this star graph an interpretation as representing understanding about the problem in theory or the theory crisis is questionable (for an explanation see the discussion section).

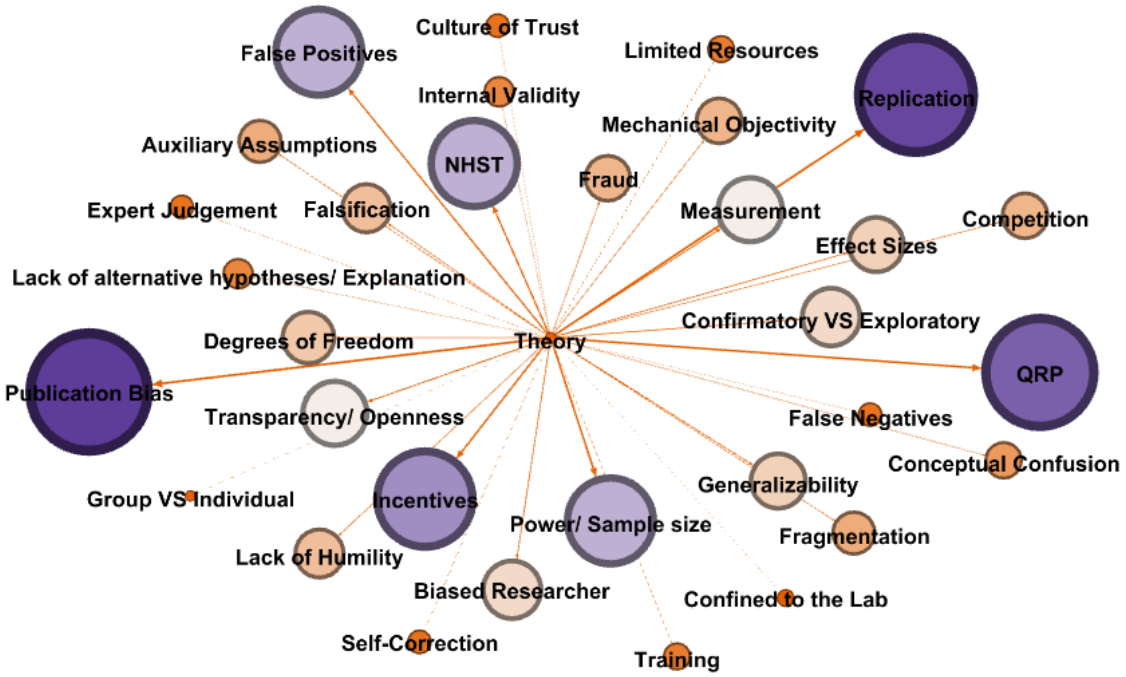


Figure 4. Theory graph showing which themes co-occur in the crisis literature with theory. Node size and color determined by the co-occurrence count.

Subsequently, I created a Measurement star graph (Figure 5) that represents the themes that are connected to measurement in the crisis literature sample. As for the theory graph the measurement graph could be used as a proxy to see where the problem in measurement is positioned in psychological science. Like for the theory graph, another motivation for constructing the measurement graph was to see whether it could be used to visualize what the problem in measurement and the (upcoming) measurement crisis actually means. The only theme that is in the top ten of the most mentioned themes that is different than in the other graphs is internal validity. The theme of internal validity represents the identified inadequacies in the research design in psychological studies, problems with random assignment, attrition rates, the lack of proper manipulation checks and a lack of control (see

e.g. Fabrigar et al., 2020; Plant, 2015). The node color goes with increasing number of co-occurrences from orange to green. By focusing on the larger green nodes in the measurement graph (Figure 5) we can see that psychological scientists who see measurement as a problem also view replication, publication bias, QRP, false positives, theory, effect sizes and internal validity (grey-green) as problematic for the discipline. Similar to the theory graph (Figure 4) an interpretation of the measurement graph as representing the understanding of the authors about the problem in measurement or the measurement crisis is questionable and would be misleading (see discussion section for an elaboration).

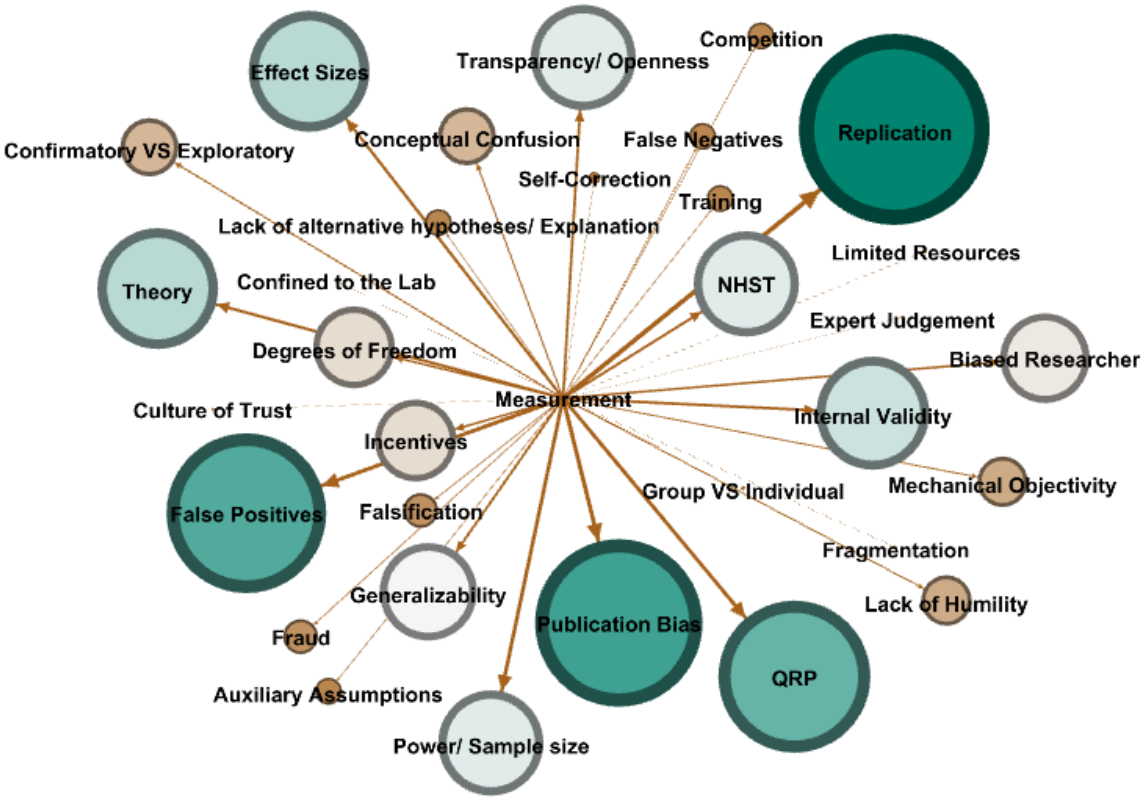


Figure 5. Measurement graph showing which themes co-occur in the crisis literature with measurement. Node size and color determined by the co-occurrence count.

The four star graphs also visibly illustrate that although the focus in the crisis literature as well as the discussions might be quite narrow as can be seen in such declarations for a crisis with foci on theory, replication, measurement, and so forth that the problems are multifaceted and too complex to capture with a crisis title that isolates one problem, when there is no such thing as an isolated problem (or crisis for that matter). One just has to think

about the inferential link from theory to statistics and the conclusion to realize that the actual question, as Szollosi and colleagues put it, might be: “How do we improve the link between psychological theory, measurement, methodology, and statistics?” (Szollosi et al., 2020), instead of just isolating one at a time. Just focusing on one of them obfuscates the existing or missing connections that are necessary for valid inferences. Focus and specialization on a specific problem are important and necessary, but the bigger picture and the interconnections between issues should not be neglected.

Another peculiarity that becomes quite visible in all the graphs is the disproportionate attention that has been paid to the two main types of error in statistical analysis (with some exceptions see for example Fiedler et al., 2012). False positives (type 1 error) are in the top five of each of the crisis graphs, while the highest position false negatives reach in all the graph is the 21st in the replication graph. In the literature for that graph false positives co-occurs 49 times with replication while false negatives only co-occurs 12 times with replication. In the complete literature sample (for the crisis graph) false positives are mentioned in 62 publications and false negatives in only 12. Oddly, this means that all publications in the sample that mention replication as problematic also mention false negatives as an issue for psychology, while 13 articles that mention replication as a problem do not mention false positive as a problem. The danger of neglecting the impact of false negatives should not be underestimated. The problem of false negatives, as Fiedler and colleagues (2012) point out are not just statistical, but also theoretical and imply the risk of neglecting potentially fruitful alternative hypotheses and explanations. This is in the graphs also represented by the theme of lack of alternative hypotheses/ Explanations.

When looking at the star graphs it is visible that there seem to be some themes that exist in the crisis literature, but receive little to no attention. Among those are *confined to the lab, group VS individual, fragmentation* and *lack of alternative hypotheses/ explanations.* Confined to the lab captures the criticism that most of psychological research and its conclusions are based on laboratory studies and that there is a general lack of naturalistic research (observation) (McGann & Speelman, 2020). This is seen as important, because the lack of naturalistic research leads to lack of a rich empirical foundation. Hence, laboratory research only provides a narrow empirical foundation. The theme group VS individual is about the circumstance that most of psychological science is based on aggregate statistics (group statistics) while neglecting the individual (see e.g. Normand, 2016). This becomes an important issue when one considers the problem of ergodicity and the fact that conclusions

based on group statistics might not apply to the individual (Speelman & McGann, 2020). The theme of fragmentation captures the concern that psychological science as a discipline lacks unification, for instance, in the form of a unifying theoretical framework (Mandler, 2011). It also contains the claimed issue that psychological scientists tend to stay within the confines of their (sub-)discipline and that the discipline shows a general lack of interdisciplinary research (Borghi, 2019). Lack of alternative hypotheses and explanations means the lack of coming up with and stating alternative explanations for a specific finding (see e.g. Holtz, 2020). This theme is related to the Duhem-Quine thesis and the underdetermination of theory by data, because it implies that just because your theory provides an explanation for the data does not imply that it is the only or even correct one. Another important node when it comes to identifying problems at their roots as well as solving the problems in the discipline that has received some attention, but not a lot, is *Training* (mentioned a total of 9 times). There seems to be a lack in proper and sophisticated statistical, theoretical and research design training in the discipline. Interesting are also the themes lack of humility and conceptual confusion, because both have quite recently gained some attention (see e.g. Bringmann et al., 2022; Hoekstra & Vazire, 2021).

According to the star graphss it seems like there is somewhat of an agreement regarding what goes wrong in psychological science and what constitutes the current crisis state. However, the disagreement might start when it comes to deciding what to focus on and which issues in the discipline are more problematic to the integrity of the scientific conduct and the validity of the inferences. Unfortunately, I did not code for strength or urgency of a certain theme, I just coded for whether it is seen as a problem for the discipline or not. After reading coding and visualizing the problems in psychological science as indicated in the literature one gets the impression that the problems that are mentioned are varied and diverse, while the focus and in depth treatment is (currently) limited to the top ten or in the graphs to the large nodes.

Limitations of Study 1

I analyzed and coded the literature alone, which could negatively influence the quality of the analysis. The literature sample for this study neglects alternative forms of publication and communication between scientists such as social media and blogs. For example, Nelson and colleagues (2021, 2022) did include blog posts in their analysis. Especially blogs have actually become quite the important medium for scientists to share their opinions on certain topics. Numerous researchers who fall into the category of reformers have

started their own blog with crisis relevant content (see for instance https://www.bayesianspectacles.org/, https://replicationindex.com/ and https://retractionwatch.com/). Furthermore the literature sample is actually quite small with only 121 publications, so any conclusions and generalizations have to be considered with caution and this should be interpreted as an inspiration for further investigation not as providing any kind of ultimate answer. Similarly, the literature sample is more representative of the second half of the chosen time interval which could affect the captured themes and makes conducting an insightful year per year occurrence and co-occurrence analysis practically uninformative. Due to the coding method in this study with a focus on what generally goes wrong in psychological research the themes remain somewhat superficial. However, those themes can be used for further deeper more targeted coding into the nature of those themes in follow-up studies. For example, with a coding targeted on what constitutes publication bias and why it is a problem for the discipline.

Discussion of Study 1

Initially the intention was to construct crisis graphss more in accordance to what Burman and colleagues (2015) did with the PsycINFO index terms, by using the index terms that are linked to each article in PsycINFO as *Subjects* as part of the articles' meta-data. However, the subjects that are attached to each article are not appropriately representative of the content in the respective articles. For instance, the article *Addressing the theory crisis in psychology* by Oberauer and Lewandowsky (2019), which is mostly about the (missing) deductive link between theory, hypothesis and the test of the hypothesis, does not have an index term as subject attached to it that would directly indicate a content related to theory. The one that is mentioned that could be interpreted as somewhat linked to theory is computational modelling, but the terms in the index for psychological terms that seem most appropriate would be theory formulation or theory verification. Therefore, themes were created for each article that represent the content in relation to what is wrong with psychological science.

The literature search in PsycINFO demonstrated certain limitations with using databases and search terms, because I had to add a good amount of literature ad hoc. I somewhat know the crisis literature in psychological science and I noticed that there were a lot of relevant publications missing in the literature that was found with the search terms in PsycINFO. Hence, for literature reviews it is not sufficient to solely and blindly rely on databases like PsycINFO or Google Scholar and it is of advantage to, at least, have some

notion of the literature one is attempting to review, because it enables one to have a critical eye on the results of the search process. As in all literature review practices, when entering literature based on personal knowledge and background it is necessary to be transparent about all the literature that is used and make a list available for others to check the work (see Appendix 1).

Since the selected literature is from between 2011 and 2020 and the focus on theory and measurement discussions in the psychological (crisis) literature came later in time (around 2019) themes and terms that are more closely related to theory or measurement (in a crisis context) are sparse and usually theory and measurement are just shortly mentioned as relevant to the problem of replication. But, since they were still mentioned as problems I coded them as such.

Importantly, the exclusive focus on crisis literature in the literature search process implies the risk of neglecting insightful publications on problems in the discipline, such as, measurement and theory that do not engage with the crisis discussions directly. Especially for the problems concerning the measurement practices in psychological science there seems to be an independent discussion going on in parallel to the crises discussions (see for instance, Bringmann & Eronen, 2016; Michell, 2013). This also demonstrates that often scientists do not have to reinvent the wheel over and over again. By paying a little bit more attention to the already existing literature they could learn a lot and maybe even save a lot of time and resources. This also shows that scholars should once in a while deviate from their usual sources and publication outlets and read literature that they usually would ignore. The literature you use should not only be informed by what you already know, some citation count or what you are used to, but it should mostly be informed by what is relevant for the specific subject that you are investigating.

To visualize the understanding of the theory and measurement crisis a more specific mapping of the measurement and theory crisis focused on the literature that concentrates on problems in theory and measurement would be more appropriate. Furthermore, the time frame of the literature would have to be expanded to ensure that enough literature can be used. Focused and elaborate discussions about measurement and theory problems or crises in psychological science started around 2019. Therefore a more appropriate time range would be 2019 until the present. Coding and qualitative analysis of the manuscripts should also not be based on what is generally going wrong in psychological research, but instead on what is wrong with theory and measurement respectively in psychological science. The focus on what

is wrong with psychological research is also problematic for later crisis declarations, because scholars who focus on theory or measurement problems do often state that things such as false positives, replication, publication bias, incentives and QRP are problematic and that they welcome such criticism of the discipline (see e.g. Fiedler, 2017; Oberauer & Lewandowsky, 2019; Schimmack, 2021). Likewise, there are researchers who focus on a problems in theory or measurement in the discipline and see these problems as an important cause for the large amount of failed replications in psychological science (Lilienfield & Strother, 2020). Hence, even though those scholars identify theory and measurement as important issues in psychological science they still view the inability of the discipline to successfully replicate its findings as a central problem in the field. The theory graph might also look so similar to the crisis state and replication crisis graph, because a lot of the replication literature mentions theory as a problem, but does not elaborate on the nature of the problem with theory. Theory is often mentioned in passing in the replication literature, which somewhat obfuscates what is actually seen as related to the problem in theory according to the constructed graph (Figure 4) in this study. Consequently, with a coding and qualitative analysis that is guided by what is wrong in psychological science earlier identified problems tend to be over represented. Additionally, the coding remains too broad. It, as mentioned before, needs to be guided by the specific problem. In the case of replication and the replication graph this is not such a big or even only a negligible problem, because replication entered the discussion and even became the focus of the discussion pretty early on (replication crisis is nearly synonymous with the general crisis state in psychological science). Hence, an interpretation of the replication graph (Figure 3) as representing (being a proxy for) how the replication crisis is understood by psychological scientists (the authors) might be justified.

However, a positive side effect of the coding and analysis approach in this study is that it allowed me to analyze and visualize what psychological scientists generally view as problematic in their discipline and whether scientists who identified a problem in theory, a problem in replication or a problem in measurement agree or disagree when it comes to what are the problems in the discipline. The star graphs and the occurrence and co-occurrence analysis show that there is quite the agreement regarding what are the problems in the discipline, while the disagreement seems to start when it comes to what are the most important, influential and urgent issues in psychological research. Although to analyze that the analysis and coding approach has to be, as mentioned earlier, adapted for that purpose.

Future research should use larger literature samples and look at the time trajectory of occurrence and co-occurrence of crisis themes to see whether the focus shifts over time. A larger literature sample would also decrease the influence of some outlier publications that have extremely high citation counts. Moreover, since crisis seems to be a reoccurring theme in the psychological literature. It could be a good idea to do occurrence and co-occurrence analyses and visualizations of the historical crisis literature during earlier crisis periods in the history of psychology. Those results could then be compared to more recent crisis discussions. It might be an insightful next step to look for differences & similarities to see whether we ignored the historical literature and could have known beforehand what was coming for us or whether something is decidedly different this time around in our understanding of the crisis. It is always important to remember that the literature analysis about crises in psychological science does not tell us what the actual problems in psychological science are, they just provide us information about how psychological scientists understand those crises and what they identify as problematic in the discipline.

Future studies could also incorporate the use of PsycINFO's psychological index terms. One possible implementation would be to add index term networks, as the ones created by Burman and colleagues (2015) for self-regulation, at points of theme graphs where understanding or research is lacking, because such meaning networks based on the psychological index terms which are two to three layers deep allow one to visualize the scope of the problems or a certain topic. This can provide insights regarding what to consider when searching the literature. Another possible first step would be to compare the theme graphs based on problems that appear in the literature to (meaning) networks of index terms to see which relevant connections or topics have not yet received enough or any attention in the literature. (See Appendix 2 for the network of the index term Experimental Replication. This could for instance be compared to the replication graph in this study.)

Such review methods as meta-analysis or systematic reviews and the one presented here are by no means a substitute for reading the literature and a critical attitude. However, since it is practically impossible for the individual psychological scientist to actually read all the literature on a certain topic, such methods as the one here could aid the scientist in selecting the literature in a more targeted and focused manner to ensure that the relevant literature is considered. The network approach to reviewing the literature also enables the researcher to see connection in the literature that provide insights regarding what has to be included depending on the focus of the intended work.

Study 2

Method

Lastly, I constructed a network out of the PsycINFO's psychological index terms, by first creating an excel file that contains index terms that are related to the three types of crisis that I introduced earlier, namely the replication crisis, the theory crisis and the measurement crisis. I used the index terms *experimental replication*, *theory formulation* and *psychometrics*. The index term measurement is not focused on measurement itself, but is also linked to a lot of term that are about specific kinds of measurements in the discipline. I chose the index term psychometrics instead of measurement, because in the index psychometrics is more linked to processes of measurement and methodological terms, while measurement is more linked to the specific context a measurement is applied in or to the subject and phenomena of investigation a measurement is intended to capture. Similarly, theory formulation was used instead of theories, because theories just lists names of theories in psychology. Theory, verification could also have been used, but I chose theory formulation since a lot of the theory crisis literature is about theory development. The narrower terms, broader terms, related terms and used for terms in the index are stored as targets in the excel file. For the second layer the targets (broader terms, narrower terms and related terms) of the first layer are the sources and their related, narrower, broader and used for terms are the targets. Subsequently, the same procedure was executed for the third layer. Afterwards the spreadsheet was used to plot a network in Gephi. The network was plotted using *PageRank* to determine the size of the nodes and *Modularity Class* to color and cluster the nodes. Put differently, the size of the nodes is based on how connected a node is to other nodes and the color of the nodes is based on which nodes are highly connected with each other, but less with other nodes. The layout was set with Force Atlas 2 with Dissuade Hubs enabled. The resulting network is hardly interpretable. However, even this somewhat messy map in Figure 6 already shows some central nodes. Afterwards, the filter In Degree Range was set to one and activated, to improve interpretability and highlight more important nodes. However, that was not enough to improve interpretability and highlight the most central nodes. Therefore, the In Degree Range was set to five (the procedure in this study was inspired by Burman et al., 2015). A way to analyze not only the meaning of the crisis state around theory, replication and measurement, but also to get an idea of the influence of certain problems that make up this crisis, is to look at clusters in the network and calculate how much each cluster contributes to the whole

network (filtered network). I chose to focus on the filtered network, to keep the interpretability and highlight the most important subjects. The calculations are based on the PageRank of the nodes.

Results

As can be seen in Figure 6 there are some groups of nodes that do not really seem to belong, because they are pushed to the outside. These are, for instance, a religion group, an evolution group and a military group. The two nodes that seem to connect the more relevant to the less relevant outsider nodes are measurement and theories. Hence, it might be an interesting first step to exclude those two from the dataset to see what effect that has on the network and whether it would focus the network and analysis more on actual aspects of research instead of application and practice. It can clearly be seen in Figure 7 that statistical analysis is the most central term, enacting quite the gravitational force on the other nodes in the methodological cluster. One has to be careful with interpreting this high centrality of Statistical Analysis, because it might be that it is so high, because it is so densely connected to crisis related index terms. However, it could also be that it is so highly connected, because in psychological science statistical analysis seems to bind and guide the research practice in the discipline (see e.g., Danziger, 1985, 1990). Danziger calls this the *methodological imperative* and it might be that the critics and reformers inherit that imperative from the (history of the) discipline. In the current study in Figure 7 the In Degree Range filter was set to 4, to focus more on the relevant subjects for the network.

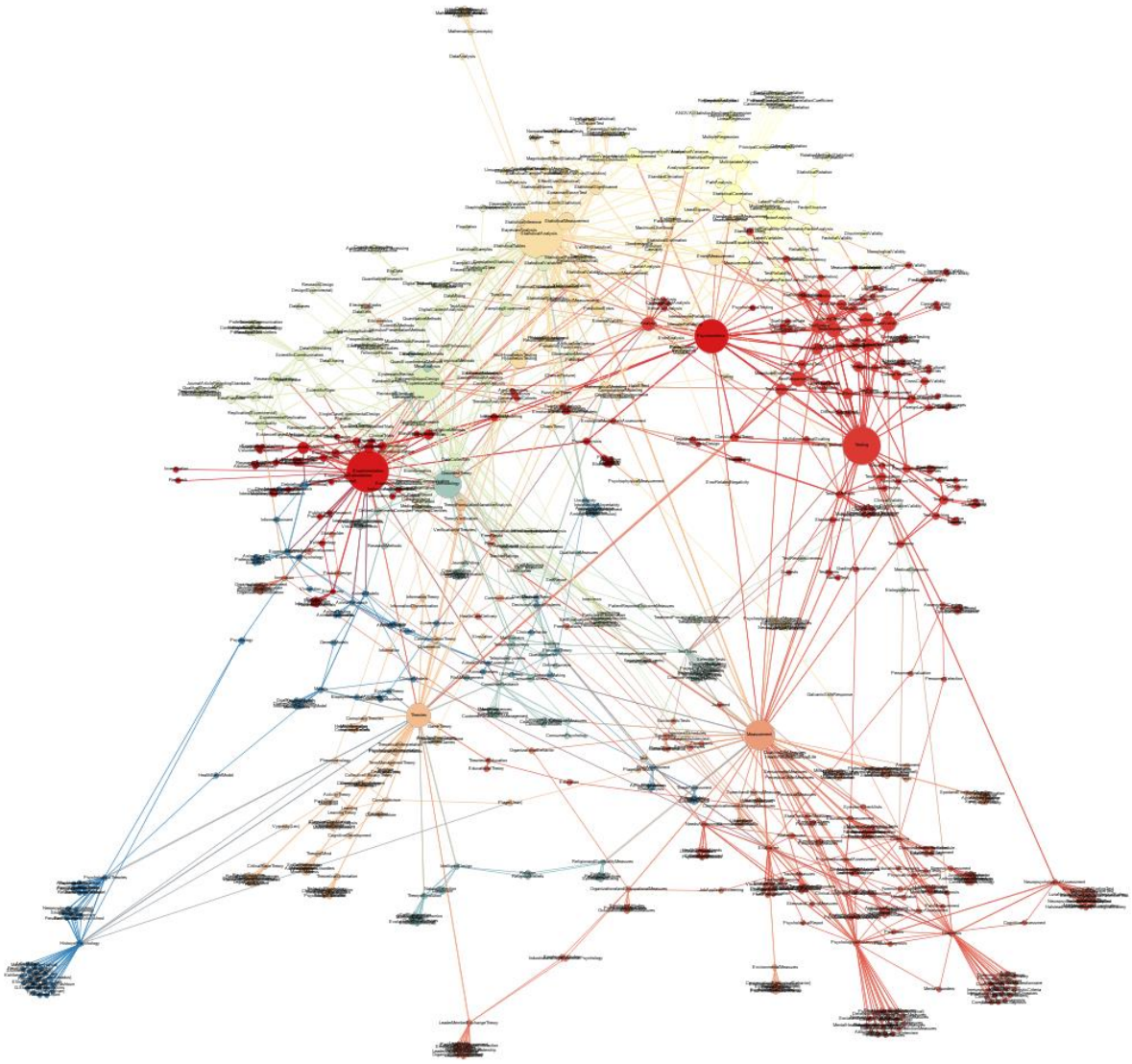


Figure 6. Network, three layers deep, of the three crisis terms experimental replication, theory formulation and psychometrics from the PsycINFO Psychological index terms together. Layout Force Atlas 2 was used with Dissuade Hubs activated and then expansion was activated three times to prevent at least some overlap.

In Figure 7 it is even somewhat clearer than in Figure 6 that to a certain degree measurement seems to be on the outside when compared to how close most of the other nodes are. Furthermore, measurement connects the three most central nodes, but so does psychometrics. Hence, Figure 7 suggests even further that it might be worth a try to exclude measurement from the network. Theories on the other hand seems to be more central than it first appeared in Figure 6. Therefore, one should compare networks with theories and measurement included

to networks with only measurement excluded and with a network where both are excluded to see the impact of those nodes on the networks.

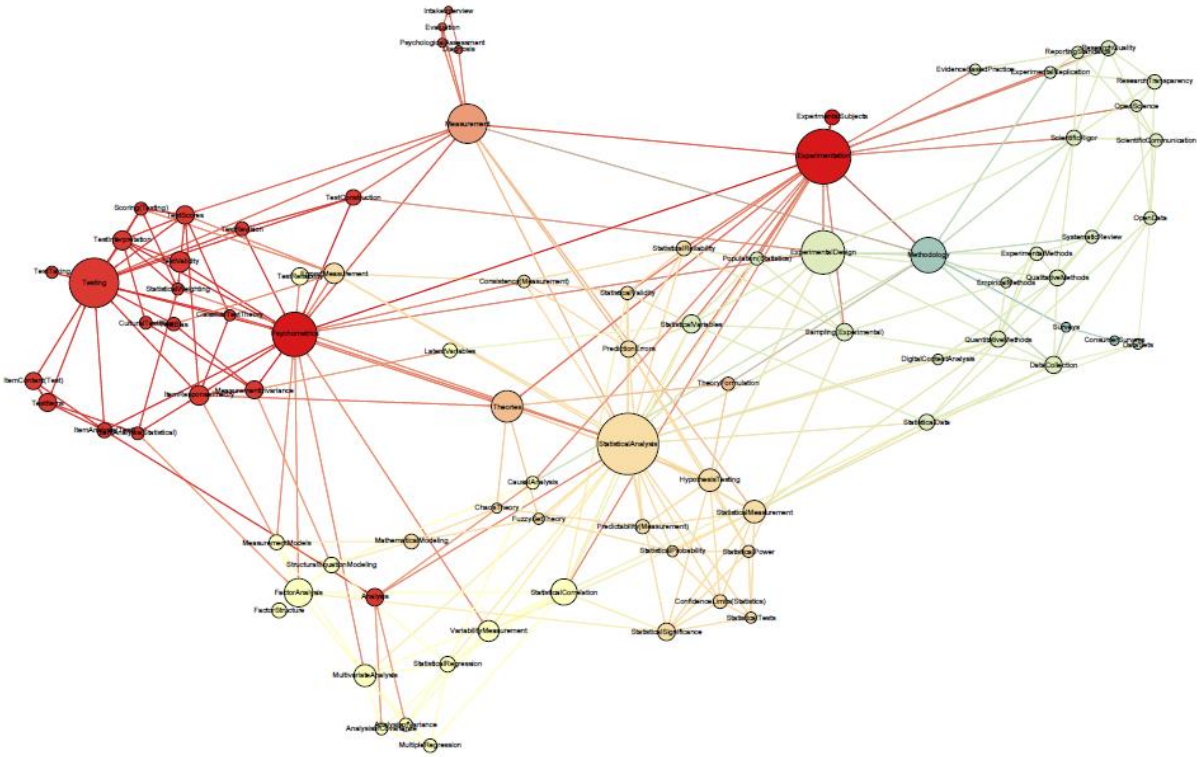


Figure 7. Same network as in Figure 6, but with the In Degree Filter Range set to five to improve interpretability and highlight the more central nodes.[10]

Experimentation Cluster

The experimentation cluster is made up of the term experimentation, psychometrics and experimental subjects. Combined these three terms contribute 8.38 percent to the meaning of the combined crisis state. This cluster somewhat function as a connection between all the nodes in the network. It is like the bridging or connecting cluster that ties everything together.

Testing Cluster

The cluster around testing is the group of terms that contributes the most to the meaning of the crisis state, with 22.15 percent. The testing cluster contains such terms (other than testing) as, Item analysis (test), classical test theory, Item Analysis (statistical), Test bias, test revision, test validity, item response theory, cultural test bias and test construction.

[10] For a bigger version of Figure 7 see Appendix 3

Theory Cluster

The theory cluster contains the terms theory formulation and theories, which together contribute 3.12 percent to the meaning of the combined crisis state. Which is a quite small contribution considering that theory has its own crisis declaration dedicated to itself. I might be that the lack of coverage of theoretical content or terms in the index is a representation of a lack of appreciation and understanding of theoretical issues among psychological scientists. In other words, this meager contribution of theory could be a manifestation of what Borsboom (2013) called theoretical amnesia. Theoretical amnesia describes psychological scientists' incompetence regarding theoretical issues (Borsboom, 2013)

Measurement Cluster

The measurement cluster is just the term measurement which contributes 2.88 percent to the meaning of the (filtered) combined crisis state.

Assessment Cluster

The Assessment cluster comprises of the terms psychological assessment, evaluation, intake interview and diagnosis. Together, these more practical terms contribute 1.78 percent to the meaning of the combined crisis state. This cluster is the least relevant for capturing the identified problems in the discipline or the understanding of psychological researchers about the aspects that have been identified as problematic. However, this cluster does provide insights about the potential scope of the problems in the field.

Statistical Analysis Cluster

Statistical analysis is the third most influential cluster in the crisis state with a contribution to the meaning of 20.14 percent. This cluster contains terms such as Hypothesis testing, statistical power, statistical probability, mathematical modelling and prediction error (e.g. Type I and Type II errors). Interestingly, this cluster also contains the term chaos theory.

Factor Analysis Cluster

The factor analysis cluster contains terms, such as, factor analysis, factor structure, latent variables, statistical correlation, analysis of covariance, analysis of variance, statistical regression and measurement models. Together the terms in this cluster contribute 16.17 percent to the meaning of the combined crisis.

Experimental Design Cluster

The cluster surrounding experimental design is the second most important cluster when it comes to the meaning of the crisis state in psychological science. It contributes 21.7 percent to the meaning of the combined crisis (theory, measurement and replication). The experimental design cluster contains terms, such as, Sampling (Experimental), experimental methods, qualitative methods, quantitative methods, scientific rigor, research quality and experimental replication. Significantly, it also contains terms that might be more interpreted as solutions to the problems in methodology, such as, Research transparency, Open science and open data.

Methodology Cluster

The methodology cluster contributes 3.69 percent to the meaning of the combined crisis state. It contains, next to methodology, the terms consumer surveys and surveys.

One could argue that seven out of the nine clusters are mostly methodological. These nine clusters together contribute about 95 percent to the meaning of the crisis state in psychological science. While the remaining five percent are either irrelevant to the purpose of this study or theoretical in nature. Even if the representation of theory related aspects in this network is only to a certain degree indicative of the appreciation for theoretical issues in psychological science, then it is no surprise that there could be something like a theory crisis lurking out in the open unrecognizable to psychological scientists due to their methodological lenses.

Limitations of Study 2

Some of the themes in study one do not have a representative index term in the APA Thesaurus of Psychological Index Terms. For instance, *culture of trust*, fraud and *publication bias* do not have a fitting index term. Moreover, some themes, such as *incentives*, do have an index term with the same name, but they focus on different things. The index term incentives does not capture the reward structure in science, but is about rewards and motivation in general. Furthermore, the APA's Thesaurus of Psychological Index Terms is regularly updated, which makes any index term network obsolete within months[11]. Resultantly, any term network remains a work in progress and has to be consciously updated.

Discussion of Study 2

[11] To keep up to date the new topic mapper could also be used: https://www.youtube.com/watch?v=n0326mYxpbw

According to the PageRank centrality indication in the index term networks in Figure 6 and Figure 7 the two terms that mostly relate to the crisis in psychological science are statistical analysis and experimentation. Creating such networks based on the index terms and comparing those to the themes in the literature might lead to the identification and visualization of terms that are currently underrepresented in the literature. The index term for measurement in this study makes also quite clear that one cannot just blindly trust the thesaurus for psychological index terms and take the terms at face value. In contrast, decisions have to be made regarding what to include and, maybe even more importantly, what to exclude, based on the focus of the research, literature search and considerations about the terms that are related to the targeted topic. In this study, for example, although measurement superficially seems to represent the theme measurement that was identified in study 1 it turned out that it was related and linked to too many less relevant terms. The better term for the current purpose was psychometrics. The APA claims that PsycINFO is the leading bibliography of psychology, but one has to ask how relevant such a statement and leading position of a databank is, when so many themes that are related to a subject like the crisis state in the field that garners such an interest among psychologists and is the focus of so many publications is not properly represented in their index of psychological terms. Among the themes that seem to be missing are: degrees of freedom, culture of trust, lack of humility, falsification, self-correction and fragmentation. In other words, although it might be the leading bibliography in psychology, one should not blindly trust PsycINFO. But, instead when using search engines, such as, PsycINFO researchers should take the responsibility and be critical as well as considerate of other sources of literature. Consequently, the crisis network that I constructed with the index terms is hardly interpretative as providing the meaning of the crisis state or the understanding of the crisis state among psychological scientists. However, single terms, such as experimental replication can be used, following prior inspection of the index entries, to provide more targeted insights about the meaning and understanding of crisis related terms that are actually represented in the index (see Appendix 2). Similar to the literature review and the theme graphs the term network analysis indicates that theory is a drastically underrepresented issue within the discipline. Comparing the themes to the clusters that were identified in term network it becomes quite visible that cultural and systematic issues within the discipline, such as culture of trust, publication bias or perverse incentives are not captured in the APA's index of psychological terms. Which should make searches with index terms harder (or even impossible) and narrower than for instances searches concerning replication or measurement. More focused term networks of one term

(e.g. experimental replication see Appendix 2) should be combined with theme networks based on theme identifications with the proposed kind of coding and qualitative analysis in study 1. This would allow one to see whether we even have words in our dictionary that can capture the more specific issues of each crisis in the field.

Conclusion

The theme graphs suggest that there is some consensus about what issues are relevant to the crisis state in psychological science. Scholars who identified replication, theory or measurement as problematic show quite the overlap regarding other issues in the discipline. According to the graphs the crisis state is mostly defined by statistical and methodological problems, with QRP, publication bias, false positives, replication and theory as the most widely mentioned problems.

However, the methodology in study 1 does not allow conclusions about the importance or dangerousness of certain issues for the discipline, nor does the coding approach allow for interpretations regarding the meaning and understanding of the theory crisis and measurement crisis in the literature sample.

Study 2 demonstrated that the use of databases for searches and the use of PsycINFO's APA Thesaurus of Psychological Index Terms requires a critical attitude and the researcher has to be careful and ideally be somewhat informed about the subject of interest. Moreover, the index might not yet be properly equipped to aid in literature searches that are related to the crisis state in psychology.

Future research is needed to figure out how network methods can complement already existing review methods and how, for instance, PsycINFO's psychological index terms might be used in a more practical and informative way to aid literature searches and reviews.

References

Anderson, S. F., & Maxwell, S. E. (2017). Addressing the "Replication Crisis": Using Original Studies to Design Replication Studies with Appropriate Statistical Power. *Multivariate Behavioral Research*, *52*(3), 305–324. https://doi.org/10.1080/00273171.2017.1289361

Bakker, M., & Wicherts, J. M. (2011). The (mis)reporting of statistical results in psychology journals. *Behavior Research Methods*, *43*(3), 666–678. https://doi.org/10.3758/s13428-011-0089-5

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, *100*(3), 407–425. https://doi.org/10.1037/a0021524

Berkman, E. T., & Wilson, S. M. (2021). So Useful as a Good Theory? The Practicality Crisis in (Social) Psychological Theory. *Perspectives on Psychological Science*, *16*(4), 864–874. https://doi.org/10.1177/1745691620969650

Borghi, A. M., & Fini, C. (2019). Theories and Explanations in Psychology. *Frontiers in Psychology*, *10*, 958. https://doi.org/10.3389/fpsyg.2019.00958

Borsboom, D. (2013). *Open Science Collaboration Blog · Theoretical Amnesia*. Retrieved January 25, 2020, from http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia/

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Bringmann, L. F., Elmer, T., & Eronen, M. I. (2022). Back to Basics: The Importance of Conceptual Clarification in Psychological Science. *Current Directions in Psychological Science*, *31*(4), 340–346. https://doi.org/10.1177/09637214221096485

Bringmann, L. F., & Eronen, M. I. (2016). Heating up the measurement debate: What psychologists can learn from the history of physics. *Theory & Psychology*, *26*(1), 27–43. https://doi.org/10.1177/0959354315617253

Burman, J. T., Green, C. D., & Shanker, S. (2015). On the Meanings of Self-Regulation: Digital Humanities in Service of Conceptual Clarity. *Child Development*, *86*(5), 1507–1521. https://doi.org/10.1111/cdev.12395

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience*, *14*(5), 365–376. https://doi.org/10.1038/nrn3475

CEFISES at UCLouvain, (2021, March 18). *Cornel & Heil – Crises, causes, and cures – DS² 2021* [Video]. YouTube. https://www.youtube.com/watch?v=SU6O7_CyzOo

Danziger, K. (1985). The Methodological Imperative in Psychology. *Philosophy of the Social Sciences*, *15*(1), 1–13. https://doi.org/10.1177/004839318501500101

Danziger, K. (1990). *Constructing the subject: Historical origins of psychological research.* Cambridge University Press. https://doi.org/10.1017/CBO9780511524059

Davidson, I. J. (2018). The Ouroboros of Psychological Methodology: The Case of Effect Sizes (Mechanical Objectivity vs. Expertise). *Review of General Psychology*, *22*(4), 469–476. https://doi.org/10.1037/gpr0000154

Derksen, M. (2019). Putting Popper to work. *Theory & Psychology*, *29*(4), 449–465. https://doi.org/10.1177/0959354319838343

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00621

Fabrigar, L. R., Wegener, D. T., & Petty, R. E. (2020). A Validity-Based Framework for Understanding Replication in Psychology. *Personality and Social Psychology Review*, *24*(4), 316–344. https://doi.org/10.1177/1088868320931366

Ferguson, C. J., & Heene, M. (2012). A Vast Graveyard of Undead Theories: Publication Bias and Psychological Science's Aversion to the Null. *Perspectives on Psychological Science*, *7*(6), 555–561. https://doi.org/10.1177/1745691612459059

Fiedler, K. (2017). What Constitutes Strong Psychological Science? The (Neglected) Role of Diagnosticity and A Priori Theorizing. *Perspectives on Psychological Science*, *12*(1), 46–61. https://doi.org/10.1177/1745691616654458

Fiedler, K., Kutzner, F., & Krueger, J. I. (2012). The Long Way From α-Error Control to Validity Proper: Problems With a Short-Sighted False-Positive Debate. *Perspectives on Psychological Science*, *7*(6), 661–669. https://doi.org/10.1177/1745691612462587

Flake, J. K., & Fried, E. I. (2020). Measurement Schmeasurement: Questionable Measurement Practices and How to Avoid Them. *Advances in Methods and Practices in Psychological Science*, 456–465. https://doi.org/10.1177/2515245920952393

Flis, I. (2019). Psychologists psychologizing scientific psychology: An epistemological reading of the replication crisis. *Theory & Psychology*, *29*(2), 158–181. https://doi.org/10.1177/0959354319835322

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the

research hypothesis was posited ahead of time. Retrieved from http://www.stat.columbia.edu/~gelman/research/ unpublished/p_hacking.pdf

Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, *102*(6), 460. https://doi.org/10.1511/2014.111.460

Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science*, *7*(6), 562–571. https://doi.org/10.1177/1745691612457576

Gong, Z., & Jiao, X. (2019). Are Effect Sizes in Emotional Intelligence Field Declining? A Meta-Meta Analysis. *Frontiers in psychology*, *10*, 1655. https://doi.org/10.3389/fpsyg.2019.01655

Hoekstra, R., & Vazire, S. (2021). Aspiring to greater intellectual humility in science. *Nature Human Behaviour*, *5*(12), 1602–1607. https://doi.org/10.1038/s41562-021-01203-8

Holtz, P. (2020). Two Questions to Foster Critical Thinking in the Field of Psychology. *Meta-Psychology*, *4*. https://doi.org/10.15626/MP.2018.984

Horbach, S. P. J. M., Aagaard, K., & Schneider, J. W. (2021, February 22). Meta-Research: How problematic citing practices distort science. https://doi.org/10.31222/osf.io/aqyhg

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review*, *2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Laraway, S., Snycerski, S., Pradhan, S., & Huitema, B. E. (2019). An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis. *Perspectives on Behavior Science*, *42*(1), 33–57. https://doi.org/10.1007/s40614-019-00193-3

LeBel, E. P., & Peters, K. R. (2011). Fearing the Future of Empirical Psychology: Bem's (2011) Evidence of Psi as a Case Study of Deficiencies in Modal Research Practice. *Review of General Psychology*, *15*(4), 371–379. https://doi.org/10.1037/a0025172

Lilienfeld, S. O., & Strother, A. N. (2020). Psychological measurement and the replication crisis: Four sacred cows. *Canadian Psychology / Psychologie Canadienne*, *61*(4), 281–288. https://doi.org/10.1037/cap0000236

McGann, M., & Speelman, C. P. (2020). Two kinds of theory: What psychology can learn from Einstein. *Theory & Psychology*, *30*(5), 674–689.

https://doi.org/10.1177/0959354320937804

Michell, J. (2013). Constructs, inferences, and mental measurement. *New Ideas in Psychology*, *31*(1), 13–21. https://doi.org/10.1016/j.newideapsych.2011.02.004

Morawski, J. (2019). The replication crisis: How might philosophy and theory of psychology be of use? *Journal of Theoretical and Philosophical Psychology*, *39*(4), 218–238. https://doi.org/10.1037/teo0000129

Morawski, J. (2020). Psychologists' psychologies of psychologists in a time of crisis. *History of Psychology*, *23*(2), 176–198. https://doi.org/10.1037/hop0000140

Munafò, M. R., Chambers, C. D., Collins, A. M., Fortunato, L., & Macleod, M. R. (2020). Research Culture and Reproducibility. *Trends in Cognitive Sciences*, *24*(2), 91–93. https://doi.org/10.1016/j.tics.2019.12.002

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, *1*(1), 0021. https://doi.org/10.1038/s41562-016-0021

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229. https://doi.org/10.1038/s41562-018-0522-1

Nelson, N. C., Chung, J., Ichikawa, K., & Malik, M. M. (2022). Psychology Exceptionalism and the Multiple Discovery of the Replication Crisis. *Review of General Psychology*, *26*(2), 184–198. https://doi.org/10.1177/10892680211046508

Nelson, N. C., Ichikawa, K., Chung, J., & Malik, M. M. (2021). Mapping the discursive dimensions of the reproducibility crisis: A mixed methods analysis. *PLOS ONE*, *16*(7), e0254090. https://doi.org/10.1371/journal.pone.0254090

Normand, M. P. (2016). Less Is More: Psychologists Can Learn More by Studying Fewer People. *Frontiers in Psychology*, *7*. https://doi.org/10.3389/fpsyg.2016.00934

Nuzzo, R. (2015). How scientists fool themselves – and how they can stop. *Nature*, *526*(7572), 182–185. https://doi.org/10.1038/526182a

Oberauer, K., & Lewandowsky, S. (2019). Addressing the theory crisis in psychology. *Psychonomic Bulletin & Review*, *26*(5), 1596–1618. https://doi.org/10.3758/s13423-019-01645-2

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Orben, A., & Lakens, D. (2020). Crud (Re)Defined. *Advances in Methods and Practices in Psychological Science*, *3*(2), 238–247. https://doi.org/10.1177/2515245920917961

Phaf, R. H. (2020). Publish less, read more. *Theory & Psychology*, *30*(2), 263–285. https://doi.org/10.1177/0959354319898250

Plant, R. R. (2016). A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter. *Behavior Research Methods*, *48*(1), 408–411. https://doi.org/10.3758/s13428-015-0577-0

Porter, T. M. (1995). *Trust in Numbers. The Pursuit of Objectivity in Science and Public Life.* Princeton, New Jersey: Princeton University Press.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, *86*(3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638

Schimmack, U. (2021). The Validation Crisis in Psychology. *Meta-Psychology*, *5*. https://doi.org/10.15626/MP.2019.1645

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simons, D. J. (2014). The Value of Direct Replication. *Perspectives on Psychological Science*, *9*(1), 76–80. https://doi.org/10.1177/1745691613514755

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384. https://doi.org/10.1098/rsos.160384

Speelman, C. P., & McGann, M. (2020). Statements About the Pervasiveness of Behavior Require Data About the Pervasiveness of Behavior. *Frontiers in Psychology*, *11*, 594675. https://doi.org/10.3389/fpsyg.2020.594675

Starns, J. J., Cataldo, A. M., Rotello, C. M., Annis, J., Aschenbrenner, A., Bröder, A., Cox, G., Criss, A., Curl, R. A., Dobbins, I. G., Dunn, J., Enam, T., Evans, N. J., Farrell, S., Fraundorf, S. H., Gronlund, S. D., Heathcote, A., Heck, D. W., Hicks, J. L., … Wilson, J. (2019). Assessing Theoretical Conclusions With Blinded Inference to Investigate a Potential Inference Crisis. *Advances in Methods and Practices in Psychological Science*, *2*(4), 335–349. https://doi.org/10.1177/2515245919869583

Stroebe, W. (2019). What Can We Learn from Many Labs Replications? *Basic and Applied Social Psychology*, *41*(2), 91–103. https://doi.org/10.1080/01973533.2019.1577736

Stroebe, W., Postmes, T., & Spears, R. (2012). Scientific Misconduct and the Myth of Self-Correction in Science. *Perspectives on Psychological Science*, *7*(6), 670–688. https://doi.org/10.1177/1745691612460687

Sturm, T., & Mülberger, A. (2012). Crisis discussions in psychology—New historical and

philosophical perspectives. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *43*(2), 425–433. https://doi.org/10.1016/j.shpsc.2011.11.001

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2020). Is Preregistration Worthwhile? *Trends in Cognitive Sciences*, *24*(2), 94–95. https://doi.org/10.1016/j.tics.2019.11.009

Wang, D., & Barabási, A.-L. (2021). *The science of science*. Cambridge University Press. https://doi.org/10.1017/9781108610834

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217. https://doi.org/10.1037/teo0000137

Yarkoni, T. (2022). The generalizability crisis. *Behavioral and Brain Sciences*, *45*, e1. https://doi.org/10.1017/S0140525X20001685

Appendix 1

Literature Sample
Statistical Rituals The Replication Delusion and How We Got There
It's Time to Broaden the Replicability Conversation Thoughts for and from clinical psychological science
Psychologists' psychologies of psychologists in a time of crisis
Neglected Sources of Flexibility in Psychological Theories: from Replicability to Good Explanations
Two kinds of theory What psychology can learn from Einstein
Measurement Schmeasurement
Theory and ontology in behavioural science
Low replicability can support robust and efficient science
Power failure button et al
Psychology's replication crisis and the grant culture righting the ship
Measuring the Prevalence of Questionable Research Practicces with incentives for truth telling
Fearing the Future of Empirical Psychology
The Rules of the Game Called Psychological Science
An Agenda for purely confirmatory research
Scientific Misconduct and the Myth of Self-Correction in Science
The (mis)reporting of statistical results in psychology journals
The natural selection of bad science
Change Starts With Journal Editors In Response to Makel (2014)
A Model of Political Bias in Social Science Research
Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science
P-value, confidence intervals, and statistical inference A new dataset of misinterpretation
Psychology, Science, and Knowledge COnstruction Broadening Perspectives from the replication crisis
Mandatory theorizing and synthesesis comment of Phaf (2020)
Addressing the Replication Crisis Using Original Studies to Design Replication Studies with Appropriate Statistical Power
Ferguson and Heene - 2012 - A Vast Graveyard of Undead Theories Publication B

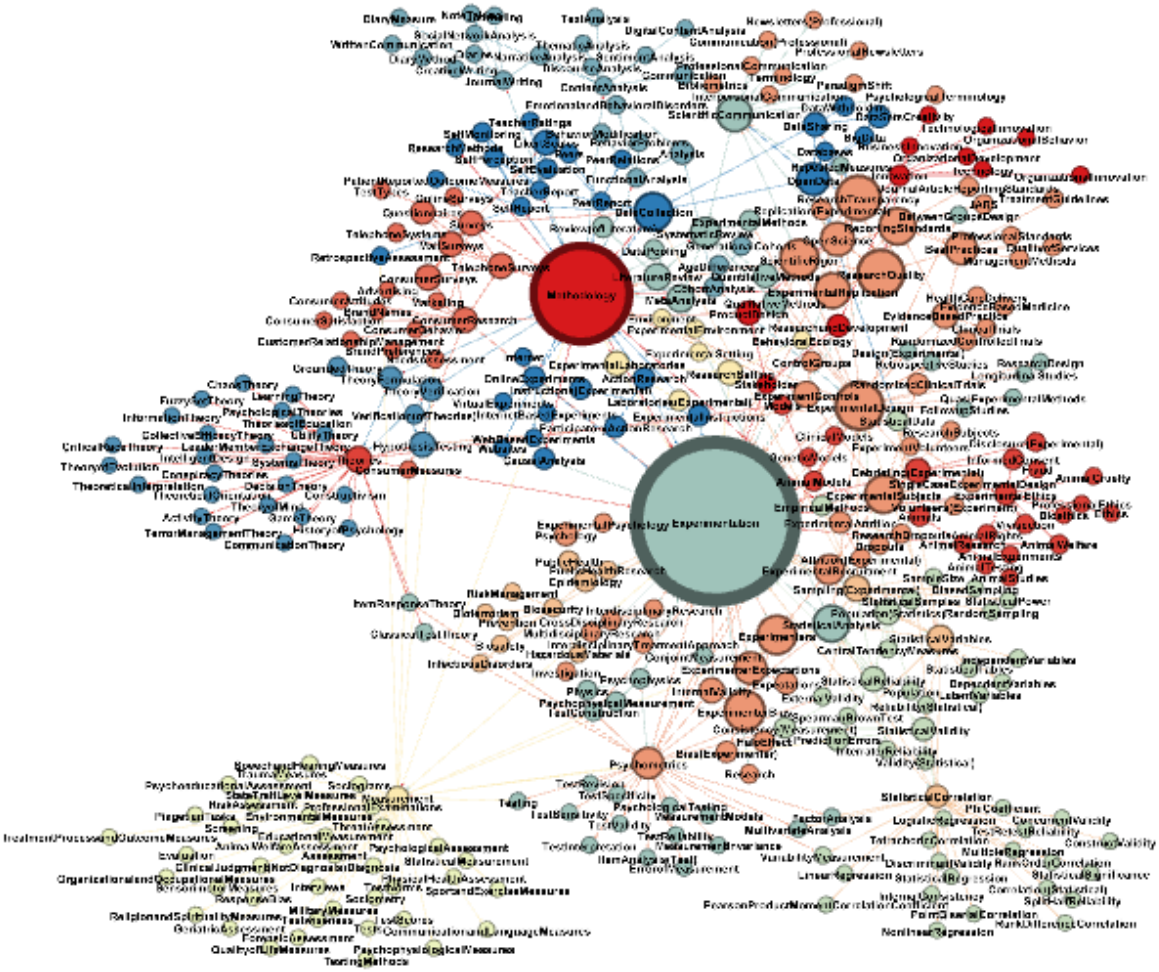
The Ouroboros of Psychological Methodology
Beyond Power Calculations
Aggressive video games research emerges from its replication crisis (sort of)
Is psychology suffering from a replication crisis What does 'failure to replicate' really mean
Addiction - 2016 - Humphreys - Grappling with the generalizability crisis in addiction treatment research
Theories and Explanations in Psychology
The statistical crisis in science how is it relevant to clinical neuropsychology
Thinking about effect sizes From the replication crisis to a cumulative psychological science
What Can We Learn from Many Labs Replications
Psychology's Crisis and the Need for Reflection. A Plea for Modesty in Psychological Theorizing
Introduction to the Special Issue on Methodological Rigor and Replicability
How scientists fool themselves – and how they can stop
Grand challenges for personality and social psychology
From discovery to justification
Estimating the reproducibility of psychological science
Caring about Power Analyses
Can't make it better nor worse
Are we at a socio-political and scientific crisis
Are Effect Sizes in Emotional Intelligence Field Declining
An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis
A Social Psychological Model of Scientific Practices
A Validity-Based Framework for Understanding Replication in Psychology
Is the call to abandon p-values the red herring of the replicability crisis
Leppink-Pérez-Fuster2017_Article_WeNeedMoreReplicationResearchA
Significance - 2015 - Peng - The reproducibility crisis in science A statistical counterattack
The Value of Direct Replication
The Epistemic Importance of Establishing the Absence of an Effect
The need for reporting negative results – a 90 year update
Putting Popper to work

Measurement error and the replication crisis
Can the behavioral sciences self-correct A social epistemic study
A manifesto for reproducible science
Barriers and solutions for early career researchers in tackling the reproducibility crisis in cognitive neuroscience
Does the conclusion follow from the evidence Recommendations for improving research
What Constitutes Strong Psychological Science The (Neglected) Role of Diagnosticity and A priori Theorizing
Fiedler - 2018 - The Creative Cycle and the Growth of Psychological
The Long Way From α-Error Control to Validity Proper Problems with a short sighted false-pisitive debate
Scientific Utopia II
Scientific Utopia I Opening Scientific Communication
J Theory Soc Behav - 2019 - Trafimow - Why successful replications across contexts and Operationalizations might not be
When more data steer us wrong
Should We Say Goodbye to Latent Constructs to Overcome Replication Crisis or Should We Take Into Account Epistemological Considerations
Research Culture and Reproducibility
Replicator degrees of freedom
Better methods can't make up for mediocre theory
Crises and problems seen from experimental psychology
Editors' Introduction to the Special Section on Replicability in Psychological Science A Crisis of Confidence
fact-or-fiction-reducing-the-proportion-and-impact-of-false-positives
Replicability Crisis and Scientific Reforms Overlooked Issues and Unmet Challenges
The Replication Crisis and Open Science in Psychology
The Crisis of Confidence in Research Findings in Psychology Is Lack of Replication the real Problem Or is it something else
Lack of Theory Building and Testing Impedes Progress in The Factor and Network Literature
From the Role of Context to the Measurement Problem The Dutch Connection Pays Tribute to Guy Van Orden

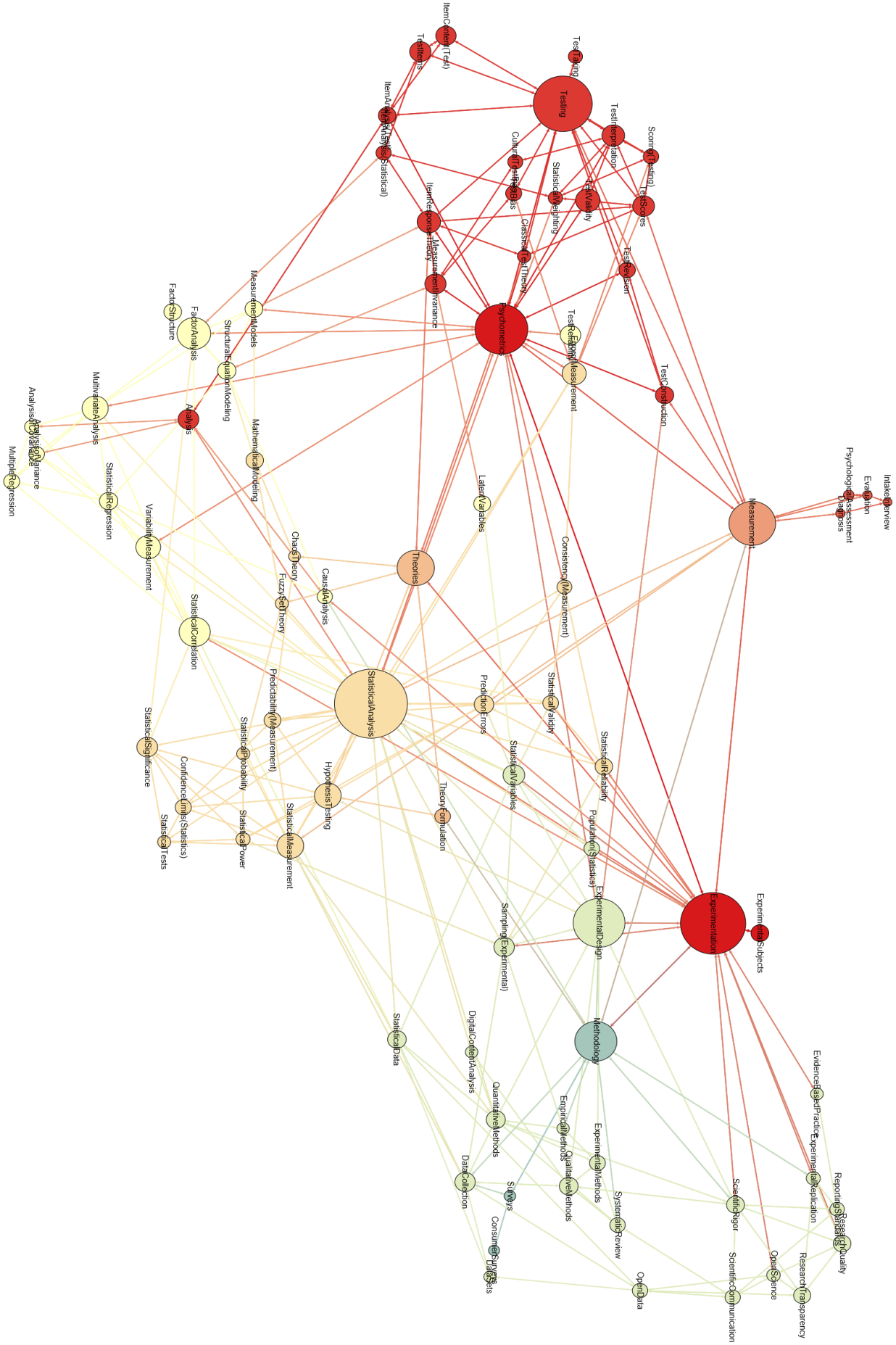
Culture in psychology Perennial problems and the contemporary methodological crisis
Implications of the credibility revolution for productivity, creativity, and progress
Mashine learning and psychological research the unexplored effect of measurement
less is more
Beyond statistics Accepting the Null Hypothesis in Mature Sciences
Muthukrishna_Henrich_2019_A problem in theory
Publish less, read more
The reproducibility crisis in psychology Attack of the clones or phantom menace
Psychology, replication & beyond
J Theory Soc Behav - 2017 - Holtz - Falsificationism is not just potential falsifiability but requires actual
Experimental power comes from powerful theories — the real problem in null hypothesis testing
False-Positive Psychology Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant
Moving Beyond 20 Questions We (Still) Need Stronger Psychological Theory
Challenges to Ego-Depletion
Badly specified theories are not responsible for the replication crisis in social psychology comment on Klein
Why Replication Is Overrated
Psychological measurement and the replication crisis Four sacred cows.
Is Preregistration worthwile
Psychologists psychologizing scientific psychology An epistemological reading of the replication crisis
The replication crisis in psychology An overview for theoretical and philosophical psychology
Replication, falsification, and the crisis of confidence in social psychology
Is the Replicability Crisis Overblown
An overview of issues in infant and developmental research for the creation of robust and replicable science
Advanced analytic and statistical methods in health psychology
The generalizability crisis
A Short Introduction to the Reproducibility Debate in Psychology

A Tutorial on Hunting Statistical Significance by Chasing N
An Open, Large-SCale, Collaborative Effort to Estimate the Reproducibility of Psychological SCience
Assessing Theoretical Conclusions With Blinded Inference to Investigate a Potential Inference Crisis
A tragedy of the (academic) commons
An a priori solution to the replication crisis
Assessing scientists for hiring, promotion, and tenure
Topics in Cognitive Science - 2017 - Gobet - Allen Newell s Program of Research The Video-Game Test
A reminder on millisecond timing accuracy and potential replication failure in computer-based psychology experiments: An open letter
Are Psychology Journals Anti-replication
A meta-psychological perspective on the decade of replication failures in social psychology
A meta-analytical answer to the crisis of confidence of psychology
Two Questions to Foster Critical Thinking in the Field of Psychology
A Short (Personal) Future History of Revolution 2.0
Everybody knows psychology is not a real science
Oberauer_Lewandowsky_2019_Addressing the theory crisis in psychology
The Validation Crisis
What can recent replication failures tell us about the theoretical commitments of psychology?

Appendix 2



Appendix 3


Testing
Psychometrics
Measurement
Experimentation
Theories
StatisticalAnalysis
Methodology
ExperimentalDesign
Analysis
FactorAnalysis
StructuralEquationModeling
MultivariateAnalysis
MeasurementModels
MathematicalModeling
StatisticalCorrelation
StatisticalRegression
VariabilityMeasurement
ChaosTheory
FuzzyLogicTheory
CausalAnalysis
LatentVariables
Consistency(Measurement)
PredictionErrors
StatisticalValidity
StatisticalReliability
StatisticalVariables
PopulationStatistics
ExperimentalSubjects
Sampling(Experimental)
HypothesisTesting
TheoryFormulation
StatisticalMeasurement
Predictability(Measurement)
StatisticalProbability
StatisticalPower
ConfidenceLimits(Statistics)
StatisticalTests
StatisticalSignificance
StatisticalData
DigitalContentAnalysis
QuantitativeMethods
QualitativeMethods
ExperimentalMethods
SystematicReview
DataCollection
Surveys
ConsumerSurveys
DataSets
OpenData
ScientificRigor
ScientificCommunication
OpenScience
ResearchTransparency
ReportingStandards
ResearchQuality
ExperimentalReplication
EvidenceBasedPractice
ItemResponseTheory
ClassicalTestTheory
TestReliability
TestValidity
TestScores
TestScoring
TestInterpretation
TestConstruction
TestDevelopment(Measurement)
ItemAnalysis(Statistical)
CulturalTestBiases
ItemContent(Test)
TestItems
TestTaking
Scoring(Testing)
MeasurementInvariance
FactorStructure
MultivariateAnalysisOfVariance
MultipleRegression
Intake/Interview
Evaluation
Diagnosis
PsychologicalAssessment