Publish less, read more

R. Hans Phaf
University of Amsterdam

Abstract
A publication deluge has impeded rather than advanced theory in experimental psychology. Many researchers rely more on null-hypothesis significance testing than literature studies to determine whether results are worthwhile. Four problematic publication practices are symptomatic for the theoretical deficit: (a) reinventing the wheel, (b) the Proteus phenomenon, (c) mechanical (non) replications, and (d) the survival of discredited hypotheses. Remedies include the development of AI tools recommending semantically related references, mandatory hypothesizing before and after results are known, and theoretical syntheses guided by meta-analyses and process models. The nonlinear theoretical development shows parallels to the optimization procedure of biological evolution. Theoretical hypotheses rather than experimental results are the elementary units of science. The fittest theories may survive alongside the least fit because they are not made to compete in research publications. Even if publication practices improve, winning hypotheses will often represent local optima and still cannot be taken with absolute certainty.

Keywords
evolutionary dynamics, experimental psychology, problematic publication practices, reference recommender tool, theoretical deficit

It is simply a sad fact that in soft psychology theories rise and decline, come and go, more as a function of baffled boredom than anything else; and the enterprise shows a disturbing absence of that cumulative character that is so impressive in disciplines like astronomy, molecular biology, and genetics. (Meehl, 1978, p. 807)

The sheer, and ever-rising, volume of publications has impeded rather than accelerated theoretical progress in experimental psychology. The larger the field, the slower the citation rates of top-cited articles change over the years (Chu & Evans, 2018). New hypotheses apparently do not get noticed and researchers stick to the trusted ones. “A deluge of papers in a scientific field does not lead to a quick turnover of central ideas, but rather to
the ossification of canon” (p. 1). The number of research publications is so staggeringly high that it is humanly impossible to read them all, even if one has a very limited field of interest. Ioannidis et al. (2017) counted about 17 million articles that involve research on humans within the search engine PubMed, and each year about 1 million articles are added. All authors of research publications have to settle with the realization that they do not know the previous literature sufficiently, and therefore cannot cite the relevant studies adequately. Symptomatic for this impasse is a number of problematic publication practices (PPPs) that occur frequently in experimental psychology. Easy measures to rectify this situation do not exist, but a few improvements are suggested.

Any research effort requires attempts to retrieve what has been done before in the field. Without such attempts, researchers run the risk of rediscovering what was already known, accepting flukes as robust findings, and advancing personal intuitions and folk theories while ignoring well-established hypotheses (i.e., the PPPs, to be discussed below). Sparse referencing is certainly not specific to experimental psychology but has, for instance, also been observed in medical randomized controlled trials (RCTs). Robinson and Goodman (2011) investigated prior citations in 1523 RCTs, of which 257 systematic reviews (e.g., meta-analyses) revealed at least three prior citable trials. Only 21% of the relevant prior reports were cited in these RCTs. Systematic reviews were also scarcely cited, and indeed a considerable number of trials claimed to be the first. In addition, older studies were cited less frequently than were more recent trials. Surprisingly, the more evidence existed the more likely subsequent RCTs were to overlook it. Sparse referencing not only represents an unjustifiable waste of research funds but also threatens the theoretical development of science.

The problem is exacerbated by considering statistical significance “conclusive” evidence indicating that the finding can stand on its own. Significance tends to serve as a stop criterion for theoretical analysis. Findings often need to be labeled “novel” to get accepted for publication. The presumed novelty of the significant effect not only relieves the authors from performing comprehensive literature searches but also lowers the need for embedding the findings into a theoretical context. Many calls have been made to abandon “fatally flawed” significance testing (e.g., Amrhein et al., 2019; Cumming, 2014; Trafimow & Marks, 2015), but that it also promotes the neglect of theory is not well recognized. I do not want to reiterate extensively the classical objections against null-hypothesis significance testing (NHST) here, but will focus on the theoretical deficit in experimental psychology, and what might potentially be done about it.

Research efforts in psychology should move from a test-centered to a theory-centered approach. If significance and the apparent success of the results cannot serve as gatekeepers for publication (i.e., in preregistered research), what can be done to stem the flood of publications that, due to their large number alone, already hinder scientific progress? In the end, no editor could reject any paper unless it had explicit methodological flaws. As a main remedy, I propose to limit the number of publications by requiring the formulation of competing substantive hypotheses in every research publication. This limitation should not apply to the execution of research or making the results available in an internet database or repository, for which I favor an “anything goes” approach (cf. Feyerabend, 1975). Fewer, but more meaningful, publications may help relieve the theoretical stagnation in experimental psychology and allow for its theoretical advancement, however nonlinear and erratic that still may prove to be.
The statistical illusion

The overwhelming number of publications has led many researchers to adopt the strategy of focusing on NHST, rather than theoretical analyses. Statistical testing applies to data but not to theories. Significance serves as a decision criterion whether an “effect” is there or not. The alarmingly high number of false positives (Ioannidis, 2005; Wilson & Wixted, 2018) and false negatives (Hartgerink et al., 2017; see also Hauer, 2004; Vadillo et al., 2016) clearly defeats this strategy. The fine-tuning of statistical analyses after the data have been obtained certainly makes things worse (see Wagenmakers et al., 2012), but even in the absence of questionable research practices, a majority of significant findings may not represent “true effects” (Ioannidis, 2005). The notion that one can decide about an “effect” based on its (non-)significance, therefore, should be called illusory.

The lack of theoretical analysis is masked by the use of the term hypothesis in statistical testing. Phaf (2016) argued that the statistical H₀ and H₁ hypotheses are indifferent to theoretical hypotheses. The H₁ entails an infinity of theoretical hypotheses (i.e., all non-zero differences) and could better be renamed H∞. An even larger infinity of potential non-zero differences (i.e., “effects”), of course, can be tested in any research field. If the proportion of “true effects” to potentially “true effects” (i.e., the prior odds) equals one, it makes no sense to do research in that field. Conversely, low prior-odds findings are generally considered surprising and novel. The role of prior odds in the false-positive rate is generally underestimated, however. Novel findings come at the cost of many false positives. Wilson and Wixted (2018) noted that if the prior odds in a field are effectively zero, as may have been the case in the controversial Bem (2011) study on precognition and premonition, all significant “effects” may come to represent false positives. Paradoxically, the more surprising a significant result is, the more likely it is false.

NHST is much more common in psychology and the social sciences than in the natural sciences. Carver (1993; see also Trafimow, 2018) provided an illustrative example of how the statistical illusion could have hampered theoretical developments in physics. In a famous set of experiments, Michelson and Morley (1887) set out to investigate the luminiferous ether hypothesis. If the ether served as a medium for light to travel, a difference in the measured speed of light would be expected between parallel and perpendicular directions with respect to the earth’s motion around the sun. The observed difference was, however, only one-fortieth of the expected difference and was interpreted by Michelson and Morley as a decidedly negative result providing evidence against the ether hypothesis. Carver (1993) subjected these data to a simple analysis of variance and obtained a statistically significant difference (p < .001) for this minuscule effect. According to the statistical illusion, significance would have confirmed the existence of luminiferous ether.

That conclusion would have set back Einstein’s ideas many years, because his notions about relativity are based on light traveling in every direction at the same speed. Fortunately, Michelson and Morley did not corrupt the scientific method by testing the null hypothesis before they interpreted their data with respect to their research hypothesis. (Carver, 1993, p. 288)

The statistical illusion has shifted focus in the experimental literature from the quality of hypotheses to the quantity of “effects.” There is an abundance of literature criticizing
NHST, often from prominent statisticians (e.g., Cohen, 1990, 1994; Gelman & Stern, 2006; Gigerenzer, 2004; Schmidt, 1996), but how experimental psychology should deal with the illusion is less often discussed. Perhaps the best advice is offered by Cumming: “I strongly suggest that the best plan is simply to go cold turkey, omit any mention of NHST, and focus on finding words to give a meaningful interpretation of the ES [effect size] estimates and CIs [confidence intervals] that give the best answers to your research questions” (2014, p. 26). The reverse strategy of focusing on NHST and not interpreting the findings meaningfully contributes much to the crisis experimental psychology is presently facing. Rather than experimental effects, theoretical hypotheses should be considered the elementary units of science: “Whether you can observe a thing or not depends on the theory which you use. It is theory which decides what can be observed” (said by Einstein to Heisenberg during the latter’s 1926 Berlin lecture, as cited in Fullbrook, 2012, p. 20).

**Problematic publication practices**

The neglect of prior findings and theories has led to four PPPs in experimental psychology. I will present a few examples, primarily from my fields of expertise, but further examples will readily come to mind with readers in their fields. It must be added that these PPPs are due to the overwhelming number of publications and probably were committed in good faith.

**Reinventing the wheel**

Supposedly novel findings do not necessarily entail novel hypotheses. Research methods in psychology have changed considerably over recent decades, primarily towards cognitive neuroscience. The research subject (i.e., human mental functioning) however, has not changed. New research methods, such as fMRI scanning and “noninvasive” brain stimulation techniques, are seen by many to bring novel findings. Hypotheses derived from older behavioral research can easily be ignored in the new, seemingly more dynamic, research domain, or even dismissed as no longer relevant due to their different methods and hypothesis wording. In my department, students are advised to refer in their theses to studies not older than 10 years, probably in order not to get bogged down in the preparatory literature search. Inevitably, good ideas are sometimes expressed there without referring to their original sources. Such advice may shorten further the habitual life cycle of a theoretical hypothesis, which is often said to become extinct when the generation of scientists responsible for it retires. Reinventing the wheel occurs when old hypotheses are presented as novel (see the example below). Whatever the, largely unintentional, reasons for ignoring related hypotheses, doing so should be considered a PPP.

The more emotionally charged an event or experience is, the more strongly it is remembered in the long term, as noted by William James: “An impression may be so exciting emotionally as almost to leave a scar upon the cerebral tissues. . . . The primitive impression has been accompanied by an extraordinary degree of attention, either as being horrible or delightful” (as cited in Hamann, 2001, p. 394). Both positive and negative arousal may strengthen explicit declarative memory, according to Hamann. Several research
waves into this phenomenon can be distinguished, but the latest one was subsumed under the term “enhancement effect” (e.g., Cahill & McGaugh, 1998). Astonishingly, the mainly cognitive-neuroscience researchers from the last wave did not seem aware of the flourishing behavioral research wave 40 years ago (for an overview, see Eysenck, 1977), and thus “reinvented the wheel.” The hypotheses in the previous wave even seemed specified better than in the last, with the former distinguishing short-term attentional restriction, at least with negatively valenced arousal, and long-term enhanced memory consolidation.

The Proteus phenomenon

Incompatible findings have appeared occasionally in the literature, mostly within short intervals but sometimes also over longer periods. Extreme results remain very attractive to editors and can be published more quickly than more moderate findings. They have the capacity to surpass the publication of less extreme results from experiments that may have been run simultaneously. The respective researchers may simply not be aware of each other’s work. Ioannidis and Trikalinos (2005) observed a large number of extreme contradictory estimates in the exploratory field of genetic association. Linking millions of polymorphisms in the human genome to thousands of diseases is hypothesis-generating research par excellence. In 22 of the 44 meta-analyses included in their study, the most extreme estimate was observed in the first year. Only 10 of the 44 meta-analyses showed the least favorable-ever study in the first year. These numbers were reversed in the second year, with 5 most favorable-ever and 15 least favorable-ever studies being published. Crucially, in 36 of these 44 meta-analyses, the extremes diagonally contradicted each other, with one extreme suggesting a protective effect of the genetic variant and the other extreme a higher susceptibility to the disease. Ioannidis and Trikalinos named this rapid, early succession of extreme findings the Proteus phenomenon, after the mythological sea-god who, like the sea, could assume many forms.

The Proteus phenomenon may be aggravated, and extended over longer timescales, in experimental psychology, due to the statistical illusion. Arguably more than in other fields, the neglect of previous literature has led to the publication of squarely contradicting, but both significant, research claims without referral to the other study. Significant findings not immediately familiar to the researcher almost instinctively evoke the reflex of calling the finding new, thus reducing the felt need for extensive literature study. In all fairness, it should be acknowledged that some authors argue that the selective publication of significant effects and the ensuing Proteus phenomenon of alternating findings infers benefits, which may be due to a kind of enhanced triangulation process. De Winter and Happee (2013) concluded from their simulation study that “it is worthwhile to be selective and publish only research findings that are statistically significant. After a number of publications, selective publishing yields a more accurate meta-analytic estimation of the true effect than publishing everything” (p. 4). The following description of the occurrence of the Proteus phenomenon in emotion psychology could thus accelerate rather than impede effect-size estimation.

Positive emotions have been postulated to boost flexibility and creativity (cf. Isen et al., 1987), whereas negative emotions are associated with a difficulty in disengaging
attention and even freezing (e.g., Belopolsky et al., 2011). With respect to the switching of attentional focus between global (i.e., large scale) and local (i.e., small scale) spatial features, however, diagonally opposing influences of positive affect have been reported. On the one hand, Tan et al. (2009; see also Baumann & Kuhl, 2005) reported greater switching flexibility with positive affect. On the other hand, Huntsinger et al. (2010) found an opposite empowerment of current focus, and greater rigidity, with positive affect. This study did not refer to Tan and colleagues, or Baumann and Kuhl. Although appearing in the same journal (Emotion) one year apart, their relatedness apparently was not noticed by reviewers or the editor. Details in the experimental setups, such as individual differences in dominant focus vs. induction of a global–local focus, may have caused the Proteus phenomenon here, but these should have been discussed extensively, given their apparent theoretical importance.

The Proteus phenomenon may represent one of the most extreme forms of unintentional “citation amnesia” (Garfield, 1982; see also Robinson & Goodman, 2011), or even of intentional citation bias. According to Duyx et al. (2017), positive findings, supporting the hypothesis under investigation, are cited about twice as often as null or negative findings. At first sight, it seems only natural that clear results should be cited more often than less distinct results. If the decision to select citations is based consciously on support for the investigated hypothesis however, the omission of opposing results should be viewed as a questionable research practice. What counts as a related hypothesis, of course, hinges on a deeper theoretical analysis. Without adequate a priori theoretical specification, the non-citation can be blamed retrospectively on an insufficient overlap with the hypothesis under investigation.

**Mechanical (non)replications**

The current replication movement considers experimental “effects” rather than theoretical hypotheses to be the core of experimental psychology (cf. Zwaan et al., 2018). A mechanical replication attempt usually takes statistical significance as the main criterion for successful replication but should be defined rather as being devoid of theory. The statistical illusion however, surely aggravates the theoretical neglect in replication attempts. I will present two examples that were considered direct (non)replications but that overlooked potentially crucial factors which could have been discovered through theoretical elaboration. The replicators did not notice that subtle changes they introduced to the experimental procedures were not trivial but could have suppressed the “effects.” To decide whether two studies really investigate the same phenomenon, a thorough theoretical analysis is required. The mechanical replication attempts, therefore, qualify neither as direct nor even as conceptual (see Zwaan et al., 2018).

Darwin (1872) famously noted “The free expression by outward signs of an emotion intensifies it. On the other hand, the repression, as far as this is possible, of all outward signs softens our emotions. . . . Even the simulation of an emotion tends to arouse it in our minds” (p. 366). Strack et al. (1988) indeed found that unobtrusive manipulations of facial muscles could influence the subjective ratings of the funniness of cartoons (i.e., through facial feedback). Wagenmakers et al. (2016), however, could not replicate this finding but had introduced subtle procedural changes. In the original study, the
experimenter was present, whereas facial expressions were only filmed in the replication attempt. Both the original authors (Strack et al., 1988) and the nonreplicators (Wagenmakers et al., 2016) seemed to have overlooked audience effects (Fridlund, 1994; see also Kraut & Johnston, 1979; Ruiz-Belda et al., 2003). For instance, if participants are asked to smell pleasant, neutral, and unpleasant odors in the presence of a human observer, the corresponding facial expressions are easily recognized. If, however, those odors are smelt without someone watching, but faces are secretly filmed, few expressions can be seen, despite large differences in odor evaluation remaining (Gilbert et al., 1987). The nonreplication and original study differed exactly in this respect, whether an observer was present or the expressions were only filmed. Facial movements were induced in the nonreplication but the video recording may not have led the participants to interpret them emotionally, due to its lack of social context. If there are only facial grimaces but no emotional expressions, there also can be no evidence for facial feedback.

Something similar has happened in another nonreplication by some of the same authors. Matzke et al. (2015), in their nonreplication of saccade-induced retrieval enhancement (SIRE; e.g., Lyle et al., 2008), conformed to standard memory research practices by suppressing the recency effect, but consequently deviated from the original procedure. The importance of the recency effect, and the studied items being active during saccades, was also not identified by Lyle and colleagues but may have been one of the crucial factors in SIRE. Matzke and collaborators could have discovered the relevance of this deviation had they discussed all four or five extant eye-movement accounts (see Phaf, 2017). They chose to discuss only the one that previously had been dismissed by the original authors and had been firmly refuted by co-authors to the nonreplication. In addition, other unmotivated differences with the original study, such as the strict selection of affectively neutral words, and a defective distinction between consistent and mixed-handers, may also have contributed to the nonreplication.

**Survival of discredited hypotheses**

Hypotheses often remain highly popular after they have been refuted rather firmly (cf. Begley & Ioannidis, 2015). Even after retraction, which constitutes a severe, and publicly acknowledged, form of discrediting, a paper may be cited frequently. Cor and Sood (2018) investigated a database of 3000 retracted papers from 2001 to 2015 and their 74,000 citations. Major errors or fraud caused 51% of these retractions; 31.2% of citations occurred within a year after retraction and 91.4% of these citations were approving. The citation rate declined in later years but remained appreciable. Evidently, many of the citing researchers only superficially explored the literature they referred to. Otherwise, they surely would have come across the retraction notices.

Many more studies, and the hypotheses therein, may become discredited without being formally retracted. This mostly results from an honest scientific debate, rather than from errors or fraud. Even after a hypothesis has been replaced by an alternative better fitting the data, the former often continues to be cited as if nothing has happened. Begley and Ioannidis (2015) provided an example from cancer biology (a psychological example follows):
This is illustrated by the MDA-MB-435 cell line that has hundreds of citations falsely identifying it as a breast cell. Despite definitive reports identifying this as a melanoma cell, at the time of writing, since 2010, there have been over 170 reports continuing to falsely identify it as a breast cell and only 47 reports correctly identifying it as a melanoma cell. (p. 120)

Conservation laws were the favorite theoretical tool of physicists in the 18th and 19th centuries. The development of the steam engine drove early thermodynamics research and the first formulations of the law of energy conservation, which states that within a closed system, energy may change form but cannot be created or destroyed. In general, a conservation law holds that the total of a particular quantity in an isolated physical system remains constant over time. Similar limited-capacity (LC) models, applying the “steam-engine metaphor” to attention, also have proven attractive to psychologists for explaining interference in dual-task performance. Norman (1968), for instance, proposed that the system was limited not due to a fixed structural point, through which both tasks had to pass, but due to having only a fixed amount of general-purpose processing resources. They are viewed as an energetic input required for all types of processing, which have to be shared by concurrent tasks. Attention is paid to one task by drawing resources from a limited reservoir, at the expense of the attention to the other task, which results in dual-task interference.

The human system is very open (and far removed from thermodynamic equilibrium). It exchanges not only large amounts of energy but also large amounts of information with its environment. Not meeting the condition of a closed system precludes the formulation of any conservation law in psychology. The capacity of this mysterious resource in the human reservoir, moreover, seems to vary and escapes quantification when measured experimentally. Attempts to evade this quandary by postulating multiple resource reservoirs have also failed. In his forceful critique on LC accounts, Allport (1980, 1989) noted that the apparent amount of resources one particular task draws from the presumed reservoir varies and depends on overlap with the concurrent task. Not much interference, for instance, is observed when visually controlled copy-typing is combined with auditory-vocal shadowing of continuous speech, but recombining input and output modes by reading aloud and audio-typing proves nearly impossible to perform concurrently. Allport (1980) suggests “that any limitations in the performance of concurrent tasks are the result of quite specific sources of interference, function-specific or data-specific limitations” (p. 143). His arguments effectively do away with the concept of a general-purpose resource.

LC accounts can only be refuted if a more viable new account emerges. The more neurobiologically plausible competition hypotheses (e.g., Desimone & Duncan, 1995; Duncan, 1996; see also Phaf et al., 1990) do a much better job of explaining data and function specificity in interference. Essentially, multiple stimuli compete for a neural representation in a distributed fashion. Where the features of the stimuli and tasks meet, they try to suppress one another through mutual inhibition. In the biased-competition (BC) account, top-down influences enhance neural responses to attended stimuli, which raises their chances of winning. LC and BC accounts are mutually exclusive. Dual-task interference is explained either by drawing from a limited reservoir of resources or by mutual inhibition in concurrent tasks. Only the latter explains why interference should
depend on the level of overlap of task features. The incompatibility of LC and BC accounts, and the discrediting of the former, is not recognized by many attention researchers. To investigate whether the publication of the BC account had any influence on the citing of LC accounts, I performed a literature search in Google Scholar and Semantic Scholar covering 1995 to 2018 (see Table 1).

All counts increase with the rising number of publications covered by the search engines over subsequent years. Semantic Scholar retrieves fewer publications than does Google Scholar, presumably due to the smaller coverage (i.e., only of open access articles) and the more efficient search procedure. These numbers moreover should only be taken as a global indication of concept acceptance. There may be considerable overlap, due to contrasting of the accounts in many papers. Twenty-three years after BC has been formulated, however, one would expect that such contrasts are no longer necessary and that references to BC increase, and references to LC decrease. After 2011, the counts

| Year | Resources and Limited Capacity | Biased Competition |
|------|-------------------------------|--------------------|
|      | Google Scholar | Semantic Scholar | Google Scholar | Semantic Scholar |
| 2018 | 775             | 134               | 139            | 27              |
| 2017 | 861             | 128               | 133            | 30              |
| 2016 | 893             | 151               | 146            | 46              |
| 2015 | 905             | 158               | 153            | 42              |
| 2014 | 899             | 173               | 163            | 50              |
| 2013 | 894             | 166               | 191            | 54              |
| 2012 | 914             | 144               | 149            | 54              |
| 2011 | 1040            | 151               | 121            | 64              |
| 2010 | 652             | 142               | 129            | 56              |
| 2009 | 629             | 166               | 94             | 46              |
| 2008 | 560             | 144               | 113            | 32              |
| 2007 | 531             | 132               | 74             | 46              |
| 2006 | 532             | 127               | 71             | 48              |
| 2005 | 366             | 125               | 57             | 34              |
| 2004 | 333             | 97                | 54             | 23              |
| 2003 | 283             | 115               | 34             | 28              |
| 2002 | 279             | 81                | 20             | 24              |
| 2001 | 236             | 96                | 15             | 19              |
| 2000 | 256             | 68                | 13             | 22              |
| 1999 | 198             | 30                | 6              | 17              |
| 1998 | 180             | 49                | 4              | 12              |
| 1997 | 140             | 41                | 4              | 11              |
| 1996 | 167             | 32                | 3              | 7               |
| 1995 | 129             | 20                | 1              | 6               |

Note. Both searches also contained the search terms “Attention” and “Interference” and “Psychology.”
remain approximately constant (only the BC count in Semantic Scholar seems to halve) with the number of references to LC remaining at substantially higher levels than those to BC. The discredited LC account seems to have ossified in the attention literature. If experiments do not contribute to an incremental development of theory, it makes no sense to do further attention research.

**Potential remedies**

*Reference recommender tools*

Instrumental in the PPPs is the unfeasibility of finding all references relevant to a study. Many scientific search tools, such as ISI Web of Science, PsycINFO, PubMed, and Google Scholar, are available, but often deliver tens or even hundreds of thousands of hits. Many of the hits concern articles that are not even slightly relevant to the research in question, because the tools search for the occurrence of loose words but have no inkling of their connectedness or meaning. To extract semantic knowledge from these papers, a researcher would have to read at least all abstracts. This requires months or years of continuous reading. Modern AI techniques may come to the aid, however. A string of new AI tools is being developed to chart the literature in order to validate existing theoretical hypotheses, reveal hidden connections between separate findings, and eventually even suggest new hypotheses for guiding research (cf. Extance, 2018). I propose here to further develop these tools so that they can recommend an ordered list of the most relevant references in response to an article’s full text. Authors may correct and improve their papers based on these recommendations, and editors and reviewers can check whether the article pays sufficient attention to foregoing theory.

The best-known of the emerging tools is the free scientific search tool, Semantic Scholar, which combines machine learning, natural language processing (i.e., for the discovery of key concepts), and machine vision (i.e., for recognizing graphs and figures) to add semantical citation analysis to traditional citation analysis (see Fricke, 2018). The search is currently restricted to about 40 million open access papers, mainly from computer science and neuroscience. Although capable of removing the long tail of irrelevant studies, Semantic Scholar may still leave less relevant studies among top-ranked articles (see Arum, 2016). Searches in Semantic Scholar are done with individual search terms rather than with full texts, which is needed for the proposed tool.

In experimental psychology, theoretical hypotheses are mainly formulated in terms of relations between concepts. The popular Word2Vec technique is able to derive such relations from word embeddings in large text corpora (Mikolov et al., 2013). The likelihood of co-occurrence of words in a text is learned by a back-propagation neural network. Through this classical supervised-learning procedure, Word2Vec trains words (i.e., as network input) against words (i.e., as output constituting the supervision) that neighbor them in the corpus. The network learns to transform each word into a node-activation vector of high dimensionality (i.e., axes representing all words in the corpus). In a well-trained network, semantically related words are located close to each other in vector space. The cosine of the angle between two vectors can act as a similarity measure between two concepts. Other simple arithmetical vector operations, such as addition or subtraction, also
provide meaningful results. The amazing ability by Word2Vec to represent semantic relations is illustrated by the famous “King – Man + Woman” equals “Queen” vector example. By employing basic vector-offset methods it is also very good at answering analogy questions, such as “Man” is to “Woman” as “Uncle” is to.. ? (“Aunt”).

Many wonderful applications of Word2Vec, in such diverse fields as machine translation, music classification and recommendation, and automatic extraction of gene disease associations, have already been developed. The most pertinent to the present purposes may be the semantic space analysis of election debates by US presidential candidates from 1999 to 2016 (Li et al., 2017), revealing distinct semantic spaces of Democratic and Republican parties and even of individual candidates. Similarity was operationalized here by the mean Euclidean distance between key concepts in the two semantic spaces. This could also serve as a correspondence measure between a new text and published articles in the proposed tool. Using Word2Vec, it should be capable of producing an ordered list of the most semantically related references by calculating the mean Euclidean distance between key concepts in the target article and potential reference articles.

The reference recommender tool may be useful in both the writing and review phases of a paper. The first run would be on a first draft, which should not include the references list yet. If it works well, it should endorse already-cited references. The recommended, noncited, references provide an opportunity for correction and improvement. Subsequently, the tool should be re-run on the final version. A list of the, say 100, most appropriate references could also be submitted for review purposes. The authors can then elucidate their reasons for not including particular references. The tool surely helps to combat each of the PPPs. Particularly mechanical replication attempts, which are usually preregistered, may turn into more direct, and conceptual, studies if the tool is run at preregistration. Audience effects, for instance, would have come to light in the preregistration of the replication attempt by Wagenmakers et al. (2016) of the facial-feedback experiment of Strack et al. (1988).

**Mandatory hypothesizing**

To redirect the focus from making binary decisions on the occurrence of “effects” to logical and creative thinking, hypothesizing should be made mandatory for publication, but not necessarily for performing the research. All too strict statistical rules and normative approaches (e.g., Wagenmakers et al., 2012) do not fit in the “anarchist” development of science (cf. Feyerabend, 1975), because they could block unforeseen, but very adaptive, solutions to scientific problems. Reviewers should be highly critical of papers suffering from the statistical illusion and while doing so evade a proper theoretical analysis. A clear indicator is when authors confuse the results of their ANOVAs with the results of their experiment, and only present inferential statistics but no, or very few, descriptive statistics (e.g., means and standard deviations) in their results sections. This unacceptable, but common, manner of reporting potentially stifles meaningful interpretation of results. Reporting only test statistics can also jeopardize the outcomes of meta-analyses. Effect sizes can be derived from test statistics, but this may lead to overestimation (Dunlap et al., 1996). Undetected errors in statistical testing, which are quite prevalent (Bakker & Wicherts, 2011), may further inflate effect sizes when calculated from test
statistics. Indeed, in one of our meta-analyses, we found that in one study, the effect size calculated from the test statistic was 10 times larger than the effect size calculated from means and standard deviations.

Peer review does not always encourage theoretical elaboration. A shallow treatment seems a wise strategy for getting a paper accepted. Perry et al. (2004) noted that:

Reviews for grants or manuscripts generally involve two or three peer researchers. When you submit a hypothesis, the likelihood of an opposing reviewer reading your proposal is high. If 80% of people disagree with you, the chance of two reviewers agreeing with you is only 4%. In contrast, if your view is accepted by 80%, the likelihood of getting two positive reviews is 64% or 16 times as likely. The odds are even more skewed by grant reviews made by panels where a consensus of 20 or more peers is required. (p. 6)

The likelihood of having opposing reviewers increases with the level of detail to which a hypothesis is specified. To promote deeper theoretical analyses, reviewers should, in my opinion, also make their own theoretical position on the subject explicit in some conflict-of-interest statement, and both reviewers and editors should strive to discount differences of opinion in their reviews and editorial decisions.

Following Francis Bacon, Platt (1964) argued that the most rapid intellectual advance in science could be made by “strong inference.” The risk of becoming “method-oriented” rather than “problem-oriented,” which was recognized and already blamed on researcher overload by Platt, seems to have materialized in experimental psychology. Strong inference can be compared to climbing a tree, involving three steps that are consecutively repeated. The first step involves generating as many alternative hypotheses as possible, the second devising a crucial experiment in which they compete to enable the exclusion of one or more of these hypotheses, and the third performing the experiment to get a clean result. Subsequently, according to Platt, the procedure should be recycled, inventing sub-hypotheses and sequential hypotheses to refine the remaining possibilities. Crucial in this procedure is the exploration of all possible substantive hypotheses and the elaboration on previously successful hypotheses. The inclusion of posterior hypothesizing in the discussion section thus seems essential for extending the research beyond the study presented in a paper.

Hypothesizing after results are known (HARKing; Kerr, 1998) is considered a scientific sin in a test-centered research approach, but may be a scientific duty in a theory-centered approach, if done as transparently as possible (cf. Hollenbeck & Wright, 2017). If significance tests are performed on posterior hypotheses as if they were planned all ahead, this certainly constitutes a violation of statistical principles. This is akin to doing 100 statistical tests and only presenting the five effects that coincidentally reached significance. Obviously, a priori hypotheses are corroborated much more strongly by supporting results than posterior hypotheses. Only the former have undergone the purifying action of selection by competition. With post-hoc theorizing, you run the risk of fitting hypotheses to noise, or of storytelling after the fact (cf. Gigerenzer, 2004). If hypothesis generation has taken place beforehand, the chances of interpreting noise are smaller. This should be no reason, however, to refrain from highly profitable posterior theorizing. In the theory-centered “strong inference” approach (Platt, 1964), the risks of storytelling...
The expansion of science into new territories depends heavily on the reinterpretation of serendipitous findings. Anomalous findings are easily ignored when they do not fit in the existing theory (cf. Lightman & Gingerich, 1992). Their innovative value is only recognized when a new conceptual framework is developed. Lightman and Gingerich presented many examples of anomalous findings leading to major shifts in theory, mostly from the physical sciences (e.g., Rutherford’s famous alpha-scattering experiment), but they can also be found in experimental psychology and cognitive neuroscience. The discovery of mirror neurons (Di Pellegrino et al., 1992), for instance, started out as informal research into the visuo-motor system, and only became anomalous when Rizzolatti realized that these neurons became active not only when the monkey grasped the object but also when the researcher made a similar grasping movement (goCognitive, 2011).

Performing only confirmatory research, in which the experimental design, statistical hypotheses, and tests are preregistered, excluding any form of HARKing (cf. Wagenmakers et al., 2012), would rob us of the most fruitful spin-offs of our investigations. Had Rutherford or Rizzolatti only performed purely confirmatory research and refrained from any post-hoc theorizing, this would have stopped the development of atomic models and of the mirror neuron concept. Estimates of the incidence of unexpected findings run astonishingly high, ranging from 40% to 60% (Dunbar, 2000). A majority of these findings are considered to be of unknown origin and thus not further interpreted (cf. Lightman & Gingerich, 1992), some 40% blamed on a methodological problem or mistake, and only about 8% were attributed to a new mechanism, according to Dunbar (2000). History of science abounds with many extremely valuable concepts that originated from serendipitous findings, such as the discovery of penicillin, plastic, X-rays, Buckyballs, gravity, and background radiation resulting from the Big Bang (cf. Lehrer, 2009). It is even hard to think of any major concept in empirical science that did not result from some unexpected finding. Given the costs associated with having performed the research and potential benefits being missed out on moreover, not engaging in post-hoc interpretation should be considered highly problematic and even intolerable. HARKing, as the origin of creative innovation, should be encouraged rather than discouraged.

Mandatory theorizing can turn nonreplications into constructive research. According to Lakatos (1970), there should be no refutation without confirmation. Ideally, the insubstantive “null hypothesis” would be banned from any replication attempt, and at least two substantive hypotheses would be pitted against each other. Nonreplications should not simply stop with “nothing is there” but should always advance theory, at least in order to be published. The replication attempt of nonconscious behavioral priming by old-age words by Doyen et al. (2012) for instance, extended theory by postulating that previously ignored experimenter expectations were crucial for the priming to occur (even when measured with infrared equipment). The theoretical contribution of a replication attempt, which often can be judged already at preregistration, should act as primary criterion for publication in a peer-reviewed journal. Mechanical replication attempts can better be deposited in an internet repository without formal review. Such repositories usually provide ample opportunity for informal comments, which may help turn them into conceptual, and even
direct, replication attempts. As explained above, there is no objection against inventing hypotheses after obtaining the data, which can serve as prior hypotheses in follow-up research. Such measures would make replication attempts more relevant to the field, could increase interest in these studies, and even enhance their publication rate.

**Theoretical synthesis**

The essence of empirical research is to collect new information, to pull together all pieces of information on a particular subject and bring them under a common roof. Meta-analyses are well suited for this purpose. Not only do they require extensive literature searches, but they also allow for the detection and arguably the correction of biases (but see Simonsohn et al., 2014). Criticisms of NHST have traditionally been accompanied by pleas for meta-analyses (e.g., Cumming, 2014; Schmidt, 1996). Cumming sees meta-analysis as an extended form of effect-size estimation and the credible limits wherein it may lie (i.e., the Confidence Interval; CI). He strongly encourages meta-analytic thinking: “Any one study is most likely contributing rather than determining; it needs to be considered alongside any comparable past studies and with the assumption that future studies will build on its contribution” (2014, p. 21). Applying CIs as a concealed significance test (i.e., determining whether zero is included in the CI) is at odds with this meta-analytic thinking. Instead of trying to establish whether an “effect” is there or not, a meta-analysis should attempt to quantify its most likely extent and narrow its CI relative to individual studies.

A meta-analysis can serve as an instrument for theoretical synthesis by distinguishing theoretically motivated moderator variables. In a meta-analysis on emotional Stroop, for instance (Phaf & Kan, 2007), it was postulated that interference by negative emotion words is larger in the subsequent trial (i.e., slow Stroop) than in the current trial (i.e., fast Stroop). This hypothesis was investigated by comparing experimental designs mixing emotional and neutral trials with designs blocking these trials. More interference was found in blocked than in mixed designs. The conclusion that emotional Stroop interference more likely arises from slow control processes than from fast automatic processes, was also supported by another moderator. In contrast to supraliminal presentation conditions of the emotion words, subliminal presentation conditions only revealed negligible interference.

In this meta-analysis, the moderators were chosen in advance but I do not see many objections to performing meta-analyses incrementally. An anomalous result from a meta-analysis may suggest new moderators for meta-analyses on the same study set. Fast emotional Stroop interference, for instance, may not be emotional at all but may depend on the semantic relatedness to the participants’ concerns (e.g., bird names to ornithologists; see Dalgleish, 1995). The meta-analysis of Phaf and Kan (2007) did not follow up on this hypothesis by including word-meaning and expertise moderators, yet this would clearly enhance our understanding of emotional Stroop. Incremental moderator specification according to the strong inference paradigm of Platt (1964) may strengthen the hypothesis-generating and synthesizing capacity of meta-analyses considerably. The best kind of meta-analysis may thus be a meta-synthesis.

Psychological theory does not appear to have attained the same standards in elucidating mechanisms as have many of the nonsocial sciences. Klein (2014) blames this partly on the tendency to investigate hypotheses via binary opposition rather than via
quantitative manipulation. Hypotheses that predict the dependent variable to wax and wane over a range of values of the independent variable should be preferred above simple present–absent hypotheses. Deciding between alternative hypotheses is facilitated considerably if they predict contrasting quantitative dependencies. Schmidt et al. (2011) pleaded for engagement in more parametric experiments, not only in their own domain but also in the whole of cognitive science. They argue that this advances theorizing by revealing functional relationships that may otherwise have been overlooked.

Parametric studies help to optimize conditions so that the largest effects are obtained. Particularly, research on new and unconventional theories may be hindered by a lack of knowledge about the best stimuli, task settings, and context conditions (cf. Fiedler et al., 2012). These theories may consequently, but incorrectly, be rejected due to their small effect sizes in suboptimally calibrated conditions. Dose-response relationship studies, for instance, are crucial in finding the optimal dose of a medicine because not all doses normally yield beneficial effects and some may even be harmful. Sign changes are particularly important, according to Schmidt and colleagues (2011), if the field of interest is not well known yet. Quantitative parametric variation, rather than binary opposition, boosts the formation of process models. Discontinuities and nonlinearities enable the decomposition into different subprocesses. A prime example in experimental psychology is the double dissociation, in which two dependent variables move in contrasting directions as a function of the parametric variation of a single independent variable. In the context of consciousness research, double dissociations due to prime-visibility variation imply that response priming and awareness cannot be driven by a single source of conscious information.

Molding process models into a more concrete format remains a challenge for experimental psychology. Mathematical models do not seem to work as well there as they do in physics. Computational models, particularly those inspired by brain processes, have more often been employed successfully to simulate psychological functions. The neural-network program AlphaZero, for instance, has been trained solely via playing against itself, not only playing the board game Go, but also chess and shogi, with only the basic rules being preprogrammed (Silver et al., 2018). After a few days of self-training, it was able to outplay any human player in these games. Human chess grandmasters were so impressed by AlphaZero, that one of them even described it as play from a superior alien species. Primarily for the associative part of human functioning, we have seen this remarkable progress, but it comes at the cost of the behavior of these models becoming intractable. There is no insight into how exactly AlphaZero accomplishes this astounding play.

The hallmark of good theory is the ability to make surprising predictions that are later corroborated by research. Such predictions abound in physics. The existence of Neptune and the return of Halley’s Comet were predicted by Newtonian classical mechanics. Eddington’s observation of light bending by the sun’s gravitational field confirmed one of the astonishing predictions from general relativity. The endorsement of surprising predictions strengthens confidence in the theory’s generalizability: “Because Einstein’s general theory of relativity has successfully predicted many things we can observe, we also take seriously its predictions for things we cannot, such as the internal structure of black holes” (Tegmark, 2007, p. 24). As another example, Dirac’s prediction of
anti-matter was at first met with disbelief, with these particles (i.e., positrons) only later being discovered in cosmic-ray traces in a cloud chamber ("Paul Dirac," n.d.; Vidmar, 2011). Many scientists may have actually observed them before, but probably disregarded them as a fluke. This again supports Einstein’s remark that it is theory which decides what can be observed. Experimental results in the absence of suitable theory only rarely seem to drive scientific developments (cf. Lightman & Gingerich, 1992).

Surprising predictions should be distinguished, of course, from surprising findings. The former arise from a posteriori theorizing, whereas the latter follow from a priori theorizing. Obviously, surprising predictions that are supported by research provide a much stronger corroboration of the theory than do surprising findings. Although surprising findings abound, not many examples of surprising predictions come to mind in experimental psychology. A nonexhaustive internet search yielded a finding that was heralded as a surprising prediction.1 Few would expect that people like things better when they are located at their dominant side, and that right- and left-handers show opposite preferences. From his body-specificity hypothesis, Casasanto (2009) indeed predicted and observed this phenomenon. Although his theoretical efforts should be commended, it may be another example of reinventing the wheel, however (see Levy, 1976).

The limitations of constructing process models and deriving concrete predictions in experimental psychology should be acknowledged. It is a cliché to say that the level of complexity is much higher in psychology than in physics. Only in human research, awareness of expectations matters for the research outcomes. Participants in a psychological experiment are routinely not told what results are expected, because this may change their behavior. Contrary to psychology, in meteorology, weather forecasts will not affect the weather. The recursive, or reflexive, nature (i.e., referring or acting upon itself) of psychological processes increases complexity manifold. Excluding recursivity, however, may still leave a complexity of unmanageable proportions in psychological theory.

The apparent success of physical theories may be deceptive because they can be applied mainly to relatively simple problems. The number of phenomena that are amenable to precise analytic description is relatively small. The famous French mathematician Henri Poincaré even proved that the motions of as few as three bodies in mutual gravitational attraction (e.g., sun, moon, and Earth), although governed by exact Newtonian laws, defy an exact solution (e.g., see Stewart, 2011). They can be predicted thousands, but not millions, of years ahead. The differential equations of such humble phenomena as a dripping faucet or a double pendulum can be written down but cannot be solved analytically, and consequently, their behavior cannot be predicted in great detail. In psychology however, the number and complexity of interacting entities, be it persons or neurons, is generally much higher than in physics, which effectively precludes the formulation of precise mathematical models.

**Evolutionary dynamics of theory**

Even when publication practices are improved, theory development in experimental psychology will not be linear. The history of science seems to defy sharply delineated general schemes (cf. Feyerabend, 1975). Its irregular development may show striking
parallels to biological evolution (cf. Marcum, 2017), which probably represents nature’s most efficient optimization procedure (cf. Dawkins, 1986). A theory is not overturned by simple Popperian falsification (Popper, 1959), but by an accumulation of anomalies (Kuhn, 1970), and only if it can be replaced by a more adaptive theory (cf. Lakatos, 1970). Selection always takes place in the presence of competing theories and makes no sense on isolated theories. Inventing alternative hypotheses and successive selection by experimentation (i.e., strong inference; Platt, 1964), indeed very much resemble random variation and selection of the fittest in evolution. The “genes” of science are not experimental “effects” but theoretical hypotheses that can mutate and be selected most effectively through strong inference.

Transparent competition between alternative theories is essential for the evolutionary dynamics of science. The PPPs have in common that they insufficiently put alternative hypotheses in competition. In experimental psychology, the fittest theories may survive alongside the least fit because they are not made to actually compete in many of its publications. The statistical illusion has quashed the selection of the fittest theoretical hypotheses and thus led to a proliferation of research papers in experimental psychology. Many of these only make a negligible contribution to the quasi-evolutionary development of theory, and instead, through their sheer number, stifle innovation. In addition, the blanket publication of all preregistered research, regardless of theoretical elaboration, could easily multiply the number of publications manifold. Publication in a peer-reviewed scientific journal should be limited to meaningful studies. Every published empirical article should deploy at least two substantive (i.e., not the statistical null and infinite) hypotheses that are made to compete by the research.

Also, after improvement of publication practices, the research outcomes and resulting hypotheses will likely represent local optima (i.e., better solutions may be reached after passing through suboptimal regions in solution space) and should not be taken with absolute certainty. Unfortunately, you cannot tell whether the solutions represent global (i.e., overall highest fitness) or local optima. As is the case in biological evolution, the future development of science is intrinsically unpredictable. Society, however, and many of the researchers funded by it, desire from science definite knowledge and established “facts.” The paradox of the scientific optimization procedure is that its solutions are sometimes very good, and can and should be applied practically, but that they still remain liable to a healthy scientific skepticism. From the evolutionary perspective, some hypotheses are merely “fitter” than others, but they can never be seen to constitute the full truth.

“In general, there is a degree of doubt, caution, and modesty, which, in all kinds of scrutiny and decision, ought for ever to accompany a just reasoner” (Hume, 1748/2008, p. 89).

**Acknowledgements**

The author thanks Gezinus Wolters for his helpful comments on a previous draft of the manuscript, and David Trafimow and Maarten Derksen for their constructive reviews.
Declaration of conflicting interests

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

R. Hans Phaf https://orcid.org/0000-0002-7691-981X

Note

1. This information was published in a 2009 press release by the Max Planck Institute in Nijmegen, entitled “Good and bad in right- and left-handers,” which has now been removed from the Internet.

References

Allport, D. A. (1980). Attention and performance. In G. Claxton (Ed.), Cognitive psychology: New directions (pp. 112–153). Routledge & Kegan Paul.

Allport, D. A. (1989). Visual attention. In M. I. Posner (Ed.), Foundations of cognitive science (pp. 631–682). MIT Press.

Amrhein, V., Greenland, S., & McShane, B. (2019, March 20). Scientists rise up against statistical significance. Nature, 567, 305–307. https://doi.org/10.1038/d41586-019-00857-9

Arum, N. S. (2016, December 6). A look at semantic scholar and Google scholar. Academia. https://www.academia.edu/31196224/A_look_at_Semantic_Scholar_and_Google_Scholar_comparisons_and_recommendations

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

Baumann, N., & Kuhl, J. (2005). Positive affect and flexibility: Overcoming the precedence of global over local processing of visual information. Motivation and Emotion, 29(2), 123–134. https://doi.org/10.1007/s11031-005-7957-1

Begley, C. G., & Ioannidis, J. P. (2015). Reproducibility in science. Circulation Research, 116(1), 116–126. https://doi.org/10.1161/CIRCRESAHA.114.303819

Belopolsky, A. V., Devue, C., & Theeuwes, J. (2011). Angry faces hold the eyes. Visual Cognition, 19(1), 27–36. https://doi.org/10.1080/13506285.2010.536186

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

Cahill, L., & McGaugh, J. L. (1998). Mechanisms of emotional arousal and lasting declarative memory. Trends in Neurosciences, 21(7), 294–299. https://doi.org/10.1016/S0166-2236(97)01214-9

Carver, R. P. (1993). The case against statistical significance testing, revisited. The Journal of Experimental Education, 61(4), 287–292. https://doi.org/10.1080/00220973.1993.10806591

Casasanto, D. (2009). Embodiment of abstract concepts: Good and bad in right- and left-handers. Journal of Experimental Psychology: General, 138(3), 351–367. https://doi.org/10.1037/a0015854
Chu, J. S., & Evans, J. A. (2018, March 2). Too many papers? Slowed canonical progress in large fields of science. SocArXiv. https://doi.org/10.31235/osf.io/jk63c
Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45(12), 1304–1312. https://doi.org/10.1037/0003-066X.45.12.1304
Cohen, J. (1994). The earth is round (p < .05). American Psychologist, 49(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997
Cor, K., & Sood, G. (2018, December 31). Propagation of error: Approving citations to problematic research. http://kennethcor.com/wp-content/uploads/2018/08/error.pdf
Cumming, G. (2014). The new statistics: Why and how. Psychological Science, 25(1), 7–29. https://doi.org/10.1177/0956797613504966
Dalgleish, T. (1995). Performance on the emotional Stroop task in groups of anxious, expert, and control subjects: A comparison of computer and card presentation formats. Cognition & Emotion, 9(4), 341–362. https://doi.org/10.1080/02699939508408971
Darwin, C. (1872). The expression of emotions in man and animals. Murray.
Dawkins, R. (1986). The blind watchmaker: Why the evidence of evolution reveals a universe without design. W. W. Norton.
Desimone, R., & Duncan, J. (1995, March). Neural mechanisms of selective visual attention. Annual Review of Neuroscience, 18, 193–222. https://doi.org/10.1146/annurev.ne.18.030195.001205
de Winter, J., & Happee, R. (2013). Why selective publication of statistically significant results can be effective. PLOS ONE 8(6), Article e66463. https://doi.org/10.1371/journal.pone.0066463
Di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. Experimental Brain Research, 91(1), 176–180. https://doi.org/10.1007/BF00230027
Doyen, S., Klein, O., Pichon, C. L., & Cleeremans, A. (2012). Behavioral priming: It’s all in the mind, but whose mind? PLOS ONE, 7(1), Article e29081. https://doi.org/10.1371/journal.pone.0029081
Dunbar, K. (2000). How scientists think in the real world: Implications for science education. Journal of Applied Developmental Psychology, 21(1), 49–58. https://doi.org/10.1016/S0193-3973(99)00050-7
Duncan, J. (1996). Cooperating brain systems in selective perception and action. In T. Inui & J. L. McClelland (Eds.), Attention and performance XVI (pp. 549–578). MIT Press.
Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. Psychological Methods, 1(2), 170–177. https://doi.org/10.1037/1082-989X.1.2.170
Duyx, B., Urlings, M. J., Swaen, G. M. H., Bouter, L. M., & Zeegers, M. P. (2017). Scientific citations favor positive results: A systematic review and meta-analysis. Journal of Clinical Epidemiology, 88, 92–101. https://doi.org/10.1016/j.clinepi.2017.06.002
Extance, A. (2018, September 10). How AI technology can tame the scientific literature. Nature, 561(7722), 273–274.
Eysenck, M. W. (1977). International series in experimental psychology: Vol. 22. Human memory: Theory, research, and individual differences. Pergamon Press.
Feyerabend, P. K. (1975). Against method: Outline of an anarchist theory of knowledge. Humanities Press.
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
Fricke, S. (2018). Semantic scholar. Journal of the Medical Library Association, 106(1), 145–147. https://dx.doi.org/10.5195/jmla.2018.280
Fridlund, A. J. (1994). *Human facial expression: An evolutionary view*. Academic Press.

Fullbrook, E. (2012). To observe or not to observe: Complementary pluralism in physics and economics. *Real-World Economics Review*, (62), 20–28.

Garfield, E. (1982). More on the ethics of scientific publication: Abuses of authorship attribution and citation amnesia undermine the reward system of science. *Current Contents: Social and Behavioral Sciences, 30*, 5–10.

Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *The American Statistician, 60*(4), 328–331. https://doi.org/10.1198/000313006X152649

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606. https://doi.org/10.1016/j.sococ.2004.09.033

Gilbert, A. N., Fridlund, A. J., & Sabini, J. (1987). Hedonic and social determinants of facial displays to odors. *Chemical Senses, 12*(2), 355–363. https://doi.org/10.1093/chemse/12.2.355
goCognitive. (2011). Giacomo Rizzolatti—Mirror neurons. http://gocognitive.net/interviews/giacomo-rizzolatti-mirror-neurons

Hamann, S. (2001). Cognitive and neural mechanisms of emotional memory. *Trends in Cognitive Sciences*, 5(9), 394–400. https://doi.org/10.1016/S1364-6613(00)01707-1

Hartgerink, C. H. J., Wicherts, J. M., & van Assen, M. A. L. M. (2017). Too good to be false: Nonsignificant results revisited. *Collabra: Psychology, 3*(1), Article 9. http://doi.org/10.1525/colla.71

Hauer, E. (2004). The harm done by tests of significance. *Accident Analysis & Prevention, 36*, 495–500.

Hollenbeck, J. R., & Wright, P. M. (2017). Harking, sharking, and tharking: Making the case for post hoc analysis of scientific data. *Journal of Management, 43*(1), 5–18. https://doi.org/10.1177/0149206316679487

Hume, D. (2008). *An enquiry concerning human understanding*. Oxford University Press. (Original work published 1748)

Huntsinger, J. R., Clore, G. L., & Bar-Anan, Y. (2010). Mood and global–local focus: Priming a local focus reverses the link between mood and global–local processing. *Emotion, 10*(5), 722–726. https://doi.org/10.1037/a0019356

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLOS Medicine, 2*(8), Article e124. http://doi.org/10.1371/journal.pmed.0020124

Ioannidis, J. P., Stuart, M. E., Brownlee, S., & Strite, S. A. (2017). How to survive the medical misinformation mess. *European Journal of Clinical Investigation, 47*(11), 795–802. https://doi.org/10.1111/eci.12834

Ioannidis, J. P. A., & Trikalinos, T. A. (2005). Early extreme contradictory estimates may appear in published research: The Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology, 58*(6), 543–549. https://doi.org/10.1016/j.jclinepi.2004.10.019

Isen, A. M., Daubman, K. A., & Nowicki, G. P. (1987). Positive affect facilitates creative problem solving. *Journal of Personality and Social Psychology, 52*(6), 1122–1131. https://doi.org/10.1037/0022-3514.52.6.1122

Kerr, N. L. (1998). HARKing: Hypothesizing after the results are known. *Personality & Social Psychology Review, 2*(3), 196–217. https://doi.org/10.1207/s15327957pspr0203_4

Klein, S. B. (2014). What can recent replication failures tell us about the theoretical commitments of psychology? *Theory & Psychology, 24*(3), 326–338. https://doi.org/10.1177/0959354314529616
Kraut, R. E., & Johnston, R. E. (1979). Social and emotional messages of smiling: An ethological approach. *Journal of Personality and Social Psychology, 37*(9), 1539–1553. https://doi.org/10.1037/0022-3514.37.9.1539

Kuhn, T. S. (1970). *The structure of scientific revolutions* (2nd ed.). University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of scientific research programmes. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the growth of knowledge* (pp. 91–196). Cambridge University Press.

Lehrer, J. (2009, December 21). Accept defeat: The neuroscience of screwing up. *Wired*. http://www.wired.com/2009/12/fail_accept_defeat/

Levy, J. (1976). Lateral dominance and aesthetic preference. *Neuropsychologia, 14*(4), 431–445. https://doi.org/10.1016/0028-3932(76)90072-5

Li, P., Schloss, B., & Follmer, D. J. (2017). Speaking two “languages” in America: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior Research Methods, 49*(5), 1668–1685. https://doi.org/10.3758/s13428-017-0931-5

Lightman, A., & Gingerich, O. (1992, February 7). When do anomalies begin? *Science, 255*(5045), 690–695. https://doi.org/10.1126/science.255.5045.690

Lyle, K. B., Logan, J. M., & Roediger, H. L. (2008). Eye movements enhance memory for individuals who are strongly right-handed and harm it for individuals who are not. *Psychonomic Bulletin & Review, 15*(3), 515–520. https://doi.org/10.3758/PBR.15.3.515

Marcum, J. A. (2017). Evolutionary philosophy of science: A new image of science and stance towards general philosophy of science. *Philosophies, 2*(Article 25). https://doi.org/10.3390/philosophies2040025

Matzke, D., Nieuwenhuis, S., van Rijn, H., Slagter, H. A., van der Molen, M. W., & Wagenmakers, E.-J. (2015). The effect of horizontal eye movements on free recall: A preregistered adversarial collaboration. *Journal of Experimental Psychology: General, 144*(1), Article e1–e15. https://doi.org/10.1037/xge0000038

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. https://doi.org/10.1037/0022-006X.46.4.806

Michelson, A. A., & Morley, E. W. (1887). On the relative motion of the earth and the luminiferous ether. *American Journal of Science, 34*, 333–345.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Cornell University. https://arxiv.org/abs/1301.3781

Norman, D. A. (1968). Towards a theory of memory and attention. *Psychological Review, 75*, 522–536.

“Paul Dirac.” (n.d.). *Famous scientists: The art of genius*. https://www.famousscients.org/paul-dirac/

Perry, G., Raina, A. K., Cohen, M. L., & Smith, M. A. (2004, December). When hypotheses dominate: Established ideas can strangle the assessment of valid alternatives. *The Scientist, 18*, 6–7.

Phaf, R. H. (2016, December 26). Replication requires psychological rather than statistical hypotheses: The case of eye movements enhancing word recollection. *Frontiers in Psychology, 7*, Article 2023. https://doi.org/10.3389/fpsyg.2016.02023

Phaf, R. H. (2017). Eye movements enhance recollection of re-imagined negative words: A link between EMDR and SIRE?. *Journal of Experimental Psychopathology, 8*(4), 364–375. https://doi.org/10.5127/jep.059916
Phaf, R. H., & Kan, K.-J. (2007). The automaticity of emotional Stroop: A meta-analysis. *Journal of Behavior Therapy and Experimental Psychiatry, 38*(2), 184–199. https://doi.org/10.1016/j.jbtep.2006.10.008

Phaf, R. H., van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology, 22*(3), 273–341. https://doi.org/10.1016/0142-1165(90)90006-P

Platt, J. R. (1964, October 16). Strong inference. *Science, 146*(3642), 347–353.

Popper, K. R. (1959). *The logic of scientific discovery*. Hutchinson.

Robinson, K. A., & Goodman, S. N. (2011). A systematic examination of the citation of prior research in reports of randomized, controlled trials. *Annals of Internal Medicine, 154*(1), 50–55. https://doi.org/10.7326/0003-4819-154-1-201101040-00007

Ruiz-Belda, M.-A., Fernández-Dols, J.-M., Carrera, P., & Barchard, K. (2003). Spontaneous facial expressions of happy bowlers and soccer fans. *Cognition & Emotion, 17*(2), 315–326. https://doi.org/10.1080/02699930302288

Schmidt, F. L. (1996). Statistical significance testing and cumulative knowledge in psychology: Implications for training of researchers. *Psychological Methods, 1*(2), 115–129. https://doi.org/10.1037/1082-989X.1.2.115

Schmidt, F., Haberkamp, A., & Schmidt, T. (2011). Do’s and don’ts in response priming research. *Advances in Cognitive Psychology, 7*(2), 120–131. https://doi.org/10.2478/v10053-008-0092-2

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018, December 7). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science, 362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2014, December 3). Trim-and-fill is full of it (bias). *Data Colada*. http://datacolada.org/30

Stewart, I. (2011). Sources of uncertainty in deterministic dynamics: An informal overview. *Philosophical Transactions of the Royal Society A, 369*(1956), 4705–4729. https://doi.org/10.1098/rsta.2011.0113

Strack, F., Martin, L. L., & Stepper, S. (1988). Inhibiting and facilitating conditions of the human smile: A nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology, 54*(5), 768–777. https://doi.org/10.1037/0022-3514.54.5.768

Tan, H. K., Jones, G. V., & Watson, D. G. (2009). Encouraging the perceptual underdog: Positive affective priming of nonpreferred local–global processes. *Emotion, 9*(2), 238–247. https://doi.org/10.1037/a0014713

Tegmark, M. (2007, July 5). Many lives in many worlds. *Nature, 448*, 23–24. https://doi.org/10.1038/448023a

Trafimow, D. (2018). An a priori solution to the replication crisis. *Philosophical Psychology, 31*(8), 1188–1214. https://doi.org/10.1080/01446110.2018.1490707

Trafimow, D., & Marks, M. (2015). Editorial: Banning null hypothesis significance testing procedures. *Basic and Applied Social Psychology, 37*(1), 1–2. https://doi.org/10.1080/01973533.2015.1012991

Vadillo, M. A., Konstantinidis, E., & Shanks, D. R. (2016). Underpowered samples, false negatives, and unconscious learning. *Psychonomic Bulletin & Review, 23*(1), 87–102. https://doi.org/10.3758/s13420-015-0892-6

Vidmar, D. (2011). *The Dirac equation and the prediction of antimatter*. http://multimidia.ufrgs.br/contenido/frontdaciencia/dirac%20antimatter%20paper.pdf

Wagenmakers, E.-J., Beek, T., Dijkhoff, L., Gronau, Q. F., Acosta, A., Adams, R. B., Jr., Albohn, D. N., Allard, E. S., Benning, S. D., Blouin-Hudson, E.-M., Bulnes, L. C., Caldwell, T. L., Calin-Jageman, R. J., Capaldi, C. A., Carfagno, N. S., Chasten, K. T., Cleeremans, A.,
Connell, L., DeCicco, J. M., Dijkstra, K., . . . Zwaan, R. A. (2016). Registered replication report: Strack, Martin, & Stepper (1988). *Perspectives on Psychological Science, 11*(6), 917–928. https://doi.org/10.1177/1745691616674458

Wagenmakers, E.-J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Wilson, B. M., & Wixted, J. T. (2018). The prior odds of testing a true effect in cognitive and social psychology. *Advances in Methods and Practices in Psychological Science, 1*(2), 186–197. https://doi.org/10.1177/2515245918767122

Zwaan, R. A., Etz, A. Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41*, Article e120. https://doi.org/10.1017/S0140525X17001972

**Author biography**

R. Hans Phaf is an associate professor at the Amsterdam Brain and Cognition Center (ABC) of the University of Amsterdam. He combines experimental research with connectionist and evolutionary modelling in the fields of attention, memory, emotion, and contrasts between conscious and non-conscious processes. His work has been published in journals such as *Emotion; Journal of Cognitive Neuroscience; Cognitive, Affective, & Behavioral Neuroscience; Attention, Perception, & Psychophysics; and Frontiers in Psychology.*