Acta Psychologica 148 (2014) 188–194

Contents lists available at ScienceDirect

Acta Psychologica

journal homepage: www.elsevier.com/ locate/actpsy

The meaning of "significance" for different types of research [translated and annotated by Eric-Jan Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]☆

Dr. A.D. de Groot

From the Psychological Laboratory of the University of Amsterdam

ARTICLE INFO

Article history:
Received 19 January 2014
Accepted 4 February 2014
Available online 3 March 2014

Keywords:
De Groot
Exploratory research
Confirmatory research
Inference and evidence

ABSTRACT

Adrianus Dingeman de Groot (1914–2006) was one of the most influential Dutch psychologists. He became famous for his work "Thought and Choice in Chess", but his main contribution was methodological — De Groot co-founded the Department of Psychological Methods at the University of Amsterdam (together with R. F. van Naerssen), founded one of the leading testing and assessment companies (CITO), and wrote the monograph "Methodology" that centers on the empirical-scientific cycle: observation–induction–deduction–testing–evaluation. Here we translate one of De Groot's early articles, published in 1956 in the Dutch journal *Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden*. This article is more topical now than it was almost 60 years ago. De Groot stresses the difference between exploratory and confirmatory ("hypothesis testing") research and argues that statistical inference is only sensible for the latter: "One 'is allowed' to apply statistical tests in exploratory research, just as long as one realizes that they do not have evidential impact". De Groot may have also been one of the first psychologists to argue explicitly for preregistration of experiments and the associated plan of statistical analysis. The appendix provides annotations that connect De Groot's arguments to the current-day debate on transparency and reproducibility in psychological science.

The meaning of the outcomes of statistical tests — applied to psychological experiments — is subject to constant confusion. The following remarks are meant to clarify the issues at hand.

These remarks only pertain to the well-known argument, where "a hypothesis is tested", or: "the significance of certain empirical findings is assessed" by means of a null hypothesis (H_0) and an assumed significance level α. Usually H_0 is rejected whenever the calculated *P*-value is lower than the assumed threshold value α. This is considered a "positive result" — and we will use the same terminology throughout this article.

The question of interest, however, is what such a "positive result" is worth, in terms of argument, in terms of support for the hypothesis at hand. This depends on a number of factors. In this respect we wish to make a distinction, first of all, as to *the "type" of research* that provides the framework in which the relevant test is conducted.

1. Hypothesis testing research versus material-exploration

Scientific research and reasoning continually pass through the phases of the well-known empirical-scientific cycle of thought: observation – induction – deduction – testing (observe – guess – predict – check). The use of statistical tests is of course first and foremost suited for "testing", i.e., the fourth phase. In this phase one assesses whether certain consequences (predictions), derived from one or more precisely postulated hypotheses, come to pass. It is essential that these hypotheses have been precisely formulated and that the details of the testing procedure (which should be as objective as possible) have been registered in advance. This style of research, characteristic for the (third and) fourth phase of the cycle, we call *hypothesis testing research*.

This should be distinguished from a different type of research, which is common especially in (Dutch) psychology and which sometimes also uses statistical tests, namely *material-exploration*. Although assumptions and hypotheses, or at least expectations about the associations that may be present in the data, play a role here as well, the material has not been obtained specifically and has not been processed specifically as concerns the testing of one or more hypotheses that have

☆ We thank Dorothy Bishop for comments on an earlier draft, and we thank publishers Bohn Stafleu van Loghum for their permission to translate the original De Groot article and to submit the translation for publication. This work was supported in part by an ERC grant from the European Research Council. Correspondence concerning this article may be addressed to Eric-Jan Wagenmakers, University of Amsterdam, Department of Psychology, Weesperplein 4, 1018 XA Amsterdam, The Netherlands.
E-mail address: EJ.Wagenmakers@gmail.com.

http://dx.doi.org/10.1016/j.actpsy.2014.02.001
0001-6918/

been precisely postulated in advance. Instead, the attitude of the researcher is: "This is interesting material; let us see what we can find." With this attitude one tries to trace associations (e.g., validities); possible differences between subgroups, and the like. The general intention, i.e. the research topic, was probably determined beforehand, but applicable processing steps are in many respects subject to ad-hoc decisions. Perhaps qualitative data are judged, categorized, coded, and perhaps scaled; differences between classes are decided upon "as suitable as possible"; perhaps different scoring methods are tried along-side each other; and also the selection of the associations that are researched and tested for significance happens partly ad-hoc, depending on whether "something appears to be there", connected to the interpretation or extension of data that have already been processed.

When we pit the two types so sharply against each other it is not difficult to see that the second type has a character completely different from the first: it does not so much serve the testing of hypotheses as it serves *hypothesis-generation*, perhaps theory-generation — or perhaps only the interpretation of the available material itself.

In practice it is rarely possible to retain the distinction for research as sharply as has been stated here. Some research focuses partly on testing prespecified hypotheses, and party on generating new hypotheses. Even in reports of rigorous-objective research one often finds, either in the discussion of the results or intermixed with the objective text, a section with interpretation, where the writer transcends the results, and therefore *generates* new hypotheses (phase 2).

When, however, research has such a mixed character, it is still possible to *discriminate hypothesis testing parts from exploratory parts*; it is also possible, in the text, *to separate the discussion of the one type and the other*. This is not only possible, this is also highly desirable. Testing and exploration have a different scientific value, they are grounded in different modes of thought, they lead to different certainties, they labor under different uncertainties. When their results are treated in the same breath, these differences are somewhat obscured: the impression is given that the positive results of the hypothesis tests have also "proven" the results from exploration (interpretations) — or, that the meaning of hypothesis test outcomes is no different from that of other elements in the interpretative whole in which they are processed.

In the following we discuss, as far as the material-exploration is concerned, only the special case where it features counting and measurement and even the calculation of significances. It is possible, however, that the results of the comparison of this case with that of hypothesis testing research also illuminates the problems and dangers of exploration in general (interpretation and hineininterpretieren).

2. Hypothesis testing research for a single hypothesis

The simplest case, from the perspective of statistical reasoning, is the one where a single predetermined hypothesis is tested in a predetermined fashion.

Assuming that no errors have been made in the way in which the material has been obtained, in this case in the experimentation, (a) and that this material can indeed be considered as a random sample (b) from a population that has been defined sufficiently precisely and clearly (c) then the statistical reasoning holds precisely: a "positive result" means exactly that, *if H_0 holds in the population*, the exceedance probability for a finding such as the one at hand (e.g., the probability for a chi-square that is just as large or larger, or a difference in means that is just as large or larger) is smaller than the threshold value α.[1] In addition the selected threshold α has been determined in advance: as holds for all other processing methods, it is not allowed to "adjust" this threshold to the findings.

This ideal case happens occasionally, but often there are complications at play. Among others, these can go in two directions: there can be *multiple hypotheses* that are researched simultaneously; the research can contain *elements of the material-exploration type*.[2]

As far as the validity and the interpretation of the outcomes of significance tests are concerned, these two kinds of complications can be treated from a single perspective.

3. Hypothesis testing research for multiple hypotheses

When multiple separate hypotheses are assessed for their significance in a *strictly hypothesis testing research paradigm* and when the interpretation of the observed "positive results" occurs exclusively *under the assumption that H_0 holds in the population* — both of these preconditions we will maintain for now — then this problem is manageable. When we test N (null) hypotheses, then, if H_0 is true in all cases, the probability of falsely rejecting H_0 on the basis of the sample results for each of the hypotheses separately equals α. The situation therefore appears to be identical to the case of a single hypothesis.

Nevertheless, a complication arises: the probability, that e.g. *one or two* of the N null hypotheses, that *have not been selected in advance*, are falsely rejected, is not at all equal to α.

For instance, when $N = 10$ it is as if one participates — again: when H_0 holds in all 10 cases — in a game of chance with "probability of losing" α for each "draw" or "throw". The probability, that we do not lose *a single time* in 10 draws can be calculated in the case that the draws are independent[3]; it equals $(1 - \alpha)^{10}$. For $\alpha = 0.05$, the traditional 5% level, this becomes $0.95^{10} = 0.60$. This means, therefore, that we have a 40% chance of rejecting at least one of our 10 null hypotheses — falsely. Had we used the 1% level, the error probability under this scenario — H_0 holds in the population for all 10 — equals $1 - 0.99^{10} = 1 - 0.91 = 0.09$; still 9%.

The situation, where "n out of N studied associations proved to be significant", i.e. in our terminology yielded "positive results", is apparently rather treacherous. Especially when n is small relative to None one is well advised to keep in mind, that (when all null hypotheses are true) on average αN accidental "positive results" are expected. Hence one cannot just rely on such "positive results".

An obvious control on the value of *the research as a whole* is: assess whether the observed n is significantly larger than αN, i.e. to calculate the exceedance probability for n out of N "losses" (or "hits") when the probability of losing (or getting hit) is $p = \alpha$ on every occasion. For $N = 10$ and $\alpha = 0.05$ we find e.g. the exceedance probabilities: for 1 accidental "positive result" $P(n \geq 1) = 0.40$ (see above), for 2 accidental "positive results" $P(n \geq 2) = 0.09$, for 3 accidental "positive results" $P(n \geq 3) = 0.01$.

This means that from $n = 3$ onward there is sufficient cause to reject the joint null hypothesis, viz. that all 10 null hypotheses are true. When we do so, we reject the thought that *all three* positive results are produced by chance; this does not, however, exclude the possibility that one or two of the three are produced by chance.

The question of which results are produced by chance and which are not can only be addressed on the basis of additional findings (the size of the respective P-values; a possible substantive connection between the hypotheses; and, for a more exact answer: a replication of the experiment). We will not delve deeper into this issue. The main purpose of this exposition was to demonstrate the serious weakening of the argument from significance in case n is small relative to N. This weakening

[1] For a more detailed treatment of this way of reasoning, see the accompanying article by J. C. Spitz.

[2] Other causes of complications can lie in not fulfilling the preconditions mentioned under (a), (b), and (c) above: contaminated materials (a), the sample is not random (b), the population is ill-defined (c). These are not considered here. Even in the "ideal" case discussed here the interpretation of outcomes of significance-research can easily lead to indefensible conclusions, as discussed in the article of J. C. Spitz.

[3] The *calculation* indeed holds exactly only when the samples are independent — e.g. when the same hypothesis is tested in different nonoverlapping subgroups of the entire sample; the *weakening of the "argument from significance"*, which is at stake here, also occurs when independence does not hold strictly — e.g. for validation of different (correlated) predictors of a single criterion variable — but is more difficult to calculate.

is a consequence of the fact, that the evaluation of the outcomes of the statistical tests is *preceded by a selection, on the basis of those same outcomes*. In the case of a single hypothesis there is no selection; in the case of n positive results from N the effect is more serious the smaller n is (closer to αN); in the case of material-exploration it is impossible, as we will see, to estimate the seriousness of the selection effect even as an approximation.[4]

4. Material-exploration: *N* becomes unspecified

In exploratory processing of materials the available empirical material is explored and processed under different perspectives and in different ways that have not been prespecified, with the aim of *finding associations*, or also to seek confirmation for associations that were anticipated but not precisely defined as hypotheses. The goal is "to let the material speak". The researcher will try to avoid "hineininterpretieren", he will try to avoid contaminating the variables between which he seeks associations, he will be on his guard for spurious correlations; but nevertheless he still attempts, by means of a procedure that consists of searching, trying, and selecting, to "extract from the material what is in it".

Of course, this means that he will also extract that which is in there *accidentally*.

As a warning, in principle this last remark could suffice. It is nevertheless worthwhile to examine the state of affairs more closely.

The researcher proceeds by trying and selecting. *Trying* in the sense that he experiments with (associations between) several variables, with several operational definitions (coding schemes, classifications) for the same variable, with several subgroupings of the entire material, and/or with several association norms and statistical tests, etc. *Selecting*, in the sense that he does not execute, according to some sort of system, all possible processing methods but instead executes only those that "promise something", "appear to show something". This selection occurs *ad hoc*, i.e. partially connected to "what the material shows", so partially connected to outcomes expected or provisionally obtained on the basis of those materials.

Suppose he uses the 5% level. A first inspection and preprocessing of the materials leads him to assess 20 associations, that, at first sight, "promise something". These 20 associations, however, are perhaps 20 out of (e.g.) 200 that he could have investigated had he not let the material partly guide his choice, but instead proceeded according to some sort of objective system of possible variations. Now when it happens that out of these 20 associations there are 10 that yield "positive results", we cannot register this as 10 successes from 20; they are 10 successes from 200. N is not 20 but 200, in this example; using $\alpha = 0.05$ yields $\alpha N = 10$. This means that $n = \alpha N$. The 10 "positive results" together are therefore insufficient to reject the joint null hypothesis that all N ($=200$) null hypotheses are true; statistically they do not mean anything.

The real difficulty is that when one explores — when the researcher lets himself be guided by presumptions and ideas that originated partially *ad hoc* — one does not know how large a number to assign to N. As soon as one starts to try and choose ad hoc, N becomes *undetermined*; an exact interpretation of the meaning of "positive results" is no longer possible.

5. Exploration of the behavior of a die

By neglecting this reasoning one can obtain results that are no different from a product of "capitalizing on coincidences". How easy this is can be clarified by the following report of an experiment on chance with a single die. This experiment served as a parapsychological investigation: the purpose was to study the ability for "psychokinesis" of a possibly paranormally gifted participant. This participant tried to concentrate continually on the 6, while a different participant in an adjacent room used a cup to throw a die 300 times; the hope is that the 6 would show up more often than expected according to the null hypothesis.

Afterwards the participant explained that it had been effortful to concentrate on the 6; he had the feeling, that he did have some influence on the state of affairs, but that this influence could possibly have turned out slightly differently than just solely on the frequency of the 6. Furthermore he had the feeling, that sometimes it "went well"; on other time points less well.

An exploration of the 300 throws (which had been divided in 5 series of 60) resulted in the following:

(1) The 6 did occur more often than 50 times in the 5 series together; the difference, however, was not significant (at the 5% level).
(2) In series 2 and 3, taken together, the 6 did occur significantly more often than expected according to chance.
(3) In the second half, throws with an even number of pips (2, 4, and 6) occurred considerably more often than expected according to chance ($P = 0.02$).

Doesn't this suggest that "something did work out"? The participant said he felt that it sometimes went well and sometimes not; so in series 2 and 3 it apparently went well. The participant said that the influence, which he thought he exerted, could have turned out a little differently than just solely on the 6; well, the surprisingly high ("significant") frequency of even numbers (including the 6) in the second half provides an indication; perhaps the evenness of the 6 did contribute there after all?

The uselessness of such interpretations is easy to demonstrate.

Ad (2): The 6 "works" in series 2 and 3 together — but the choice of this subgroup of measurements has occurred *ad hoc*. One can just as well take together 1 and 2, or 3 and 4, or 4 and 5; or compare the first half against the second half; or consider the series separately. Besides, one could also compare e.g. series 1, 3, and 5 against 2 and 4; the participant did say after all, that his ability to concentrate was "sometimes good, sometimes poor"? And finally one can also change the division in series of 60. Why not consider 3 series of 100 or 12 of 25? The division in 5 series of 60 was completely arbitrary; perhaps a different division is "more adequate for the course of the psychological process"?

In any case, the researcher chose a single subgroup of observations for a test of significance, because it "promised something"; he did not investigate other possible subgroups because they did not "promise something". For the latter it is safe to assume that the frequency of 6 does not differ significantly from what was expected under H_0. Hence we are confronted with *1 positive result out of N*. This N cannot be determined exactly, but it is rather large: from the perspective of hypothesis testing, the "positive result" has no meaning whatsoever.

Ad (3): In the first place, the same holds here as for (2): the choice for the second half of the series of observations has occurred *ad hoc*. Moreover, one can postulate numerous hypotheses other than the one that states that the psychokinetic effect has expanded from the 6 to the other even numbers: "expansion" to the high numbers (4, 5, and 6) — to the numbers divisible by 3 (3 and 6) — to the extremes (1 and 6) — the 5 "works" (also a high number, indeed the adjacent one) — the 6 occurs less often, not more often, than expected ("blocking", or a "negative psychokinetic effect" or something similar), etc. Here also we are confronted *with one "positive result" from a number* (N), that cannot be determined exactly, but is very large. The positive outcome of the statistical test does not have any meaning in terms of hypothesis testing; an exact interpretation is impossible.

The above "report of an experiment" has been constructed, in the current form, for this occasion. Should the reader feel the urge for a reality check, then he can imitate the writer and experiment on his own with e.g. 120 throws of a die. After some practice he will also not find it difficult to show for any die that it (or the person who throws it) behaves "significantly" exceptionally "somewhere". This claim can

[4] Starting with the case of n out of N, one could speak of *p.s. significances* (p.s. for post selectionem); this is to distinguish them from strictly interpretable significance findings.

always be maintained, the "proof" can always be provided, as long as one does not need to specify in advance, where exactly "somewhere" is located.

6. Conclusions

If the processing of empirically obtained material has in any way an "exploratory character", i.e. if the attempt to let the material speak leads to ad hoc decisions in terms of processing, as described above, then this precludes the exact interpretability of possible outcomes of statistical tests.

This conclusion is not new. Often, however, it is only stated that one "is not allowed to" make ad hoc decisions if one desires to test hypotheses with statistical means, or that one "is not allowed to" use statistical tests after making ad hoc decisions. By contrast, here the reasons behind these prohibitions have been illuminated, by making the connection to the case of a hypothesis testing study where *n* out of *N* postulated and investigated hypotheses yielded positive results.

Prohibitions in statistical methodology really never pertain to the calculations themselves, but pertain only to drawing incorrect conclusions from the outcomes. It is no different here. One "is allowed" to apply statistical tests in exploratory research, just as long as one realizes that they do not have evidential impact. They have also lost their forceful nature, regardless of their outcomes. The researcher can take them or leave them, as he wishes: he can follow them — e.g. by at least not mentioning the non-"significant" associations when the results are interpreted — or he can not follow them. He has a certain freedom in this, because this is not yet strict *hypothesis testing*, but merely a judicious form of *hypothesis generation*. If he keeps this latter point in mind and therefore realizes that it is still essential for these hypotheses to be precisely formulated and tested, then there is no reason to prohibit the calculation of *P*-values — even though they cannot be interpreted exactly.

A different question is whether it invariably makes sense to calculate P-values. In a material-exploration, statistical tests are of course less important than in a hypothesis testing study. Whether one wants to apply them and how one wishes to use them — in order to arrive at a judicious interpretation and/or hypothesis generation — is for the most part a matter of taste. It is nonetheless striking that so little is said about significance for a technique such as factor analysis, which is primarily suited for exploratory purposes. This is not just due to the fact that it is so difficult to pin down, but also due to the fact that it is not strictly necessary as long as one uses factor analysis in an exploratory manner.

There is the saying: *lies – damned lies – statistics*. Apparently, this saying does not just "hold" for the classical statistical procedure. The modern inductive-statistical aids are not immune either to the danger that arises, when an incapable or dishonest user applies them incorrectly and "lies".

Is it however the case that numbers in particular are treacherous? Is it the fact that we explore *statistically* which enables us to demonstrate that any die shows peculiar "behavioral patterns" — behavioral patterns in this case that are "significantly" different? Is the danger of "hineininterpretieren" or "capitalizing on coincidences" present only in quantitative research?

Of course, this is not the case. The errors of reasoning and irresponsible conclusions from the quantitative case originate from acts, which in principle are performed exactly the same in qualitative research. In that case, one attempts to "let the material speak" as well, by ordering the findings *ad hoc* and by systematizing and associating them by means of perspectives that were obtained *ad hoc*; likewise, one often uses the findings to draw conclusions that purport to generalize to other situations.

That we can say of statistical measures — or rather of (poor) statisticians — that they sometimes "lie", implies a compliment for statistics; namely, that in this discipline lies and truth are distinguishable. Indeed, the difference between quantitative and qualitative exploration methods and interpretation techniques rests mainly in that fact that in quantitative research, errors in reasoning and irresponsible conclusions are eventually *demonstrable*.

The commonalities and differences between quantitative and qualitative methods of reasoning can of course be further elaborated upon; but this exceeds the scope of this article. It suffices to point out that the moral of the above — with a small twist: "lies – damned lies – ad hoc interpretations" — befits qualitative research as least as much as it does quantitative research.

Amsterdam, July 1956

Appendix A

Translators' annotations

General comments

This is a translation of:

De Groot (1969). De betekenis van "significantie" bij verschillende typen onderzoek. *Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden*, 11, 398–409.

In the translation, we tried to stay as close as possible to De Groot's style of writing without sacrificing clarity. Syntactic differences between Dutch and English occasionally forced us to alter the sequence of sentence fragments, or, in extreme cases, split a single sentence in two. Throughout the translation, we retained the generous use of italics and hyphens that typified the style of De Groot.

The goal of this translation is to make the paper available to the scientific community and rescue it from obscurity; according to Google Scholar (27-9-2013) the paper has been cited only twice, once in 1960 and once in 1967. This is unfortunate, especially given that this short paper anticipates the crisis of confidence in psychology (Pashler & Wagenmakers, 2012) and closely connects to the current-day discussion on questionable research practices and preregistration.

Specifically, De Groot makes three important interconnected points. The first point is that exploratory analyses invalidate the standard interpretation of outcomes from hypothesis testing procedures. "Exploratory investigations differ from hypothesis testing in that the canon of the inductive method of testing is not observed, at least not in its rigid form. The researcher does take as his starting-point certain expectations, a more or less vague theoretical framework; he is indeed out to find certain kinds of relationships in his data, but these have not been antecedently formulated in the form of precisely stated (testable) hypotheses. Accordingly they cannot, in the strict sense, be put to the test." (De Groot, 1969, p. 306). Indeed, in exploratory work: "The characteristic element of 'trying out *whether* …' is present (…), but in such a way that the researcher's attitude in fact boils down to 'let us see what we can find.' Now what is 'found' — that is, selected — cannot also be tested on the same materials (…)" (De Groot, 1969, p. 307).

This same point has recently been been raised by many other researchers (e.g., Goldacre, 2008; John, Loewenstein, & Prelec, 2012; Kerr, 1998; Kriegeskorte, Simmons, Bellgowan, & Baker, 2009; MacCallum, Roznowski, & Necowitz, 1992; Simmons, Nelson, & Simonsohn, 2011; Vul, Harris, Winkielman, & Pashler, 2009), but the work of De Groot is almost never mentioned. For example, Wagenmakers, Wetzels, Borsboom, van der Maas, and Kievit (2012, p. 633) state (without acknowledging that the founding father of their department had espoused the same ideas more than half a century earlier):

> "At the heart of the problem lies the statistical law that, for the purpose of hypothesis testing, the data may be used only once. So when you turn your data set inside and out, looking for interesting patterns, you have used the data to help you formulate a specific hypothesis. Although the data may still serve many purposes after such fishing expeditions, there is one purpose for which the data

are no longer appropriate—namely, for testing the hypothesis that they helped to suggest. Just as conspiracy theories are never falsified by the facts that they were designed to explain, a hypothesis that is developed on the basis of exploration of a data set is unlikely to be refuted by that same data. Thus, one always needs a fresh data set for testing one's hypothesis."

The second, related, point that De Groot makes is the pressing need to distinguish between exploratory and confirmatory ("hypothesis testing") analyses. De Groot reiterated this point in his book "Methodology": "It is of the utmost importance at all times to *maintain a clear distinction between exploration and hypothesis testing*. The scientific significance of results will to a large extent depend on the question whether the hypotheses involved had indeed been antecedently formulated, and could therefore be tested against genuinely new materials. Alternatively, they would, entirely or in part, have to be designated as *ad hoc* hypotheses, which could, emphatically, not yet be tested against 'new' materials." (De Groot, 1969, p. 52).

Indeed, De Groot believed that it was unethical to blur the distinction between exploratory and confirmatory work: "It is a serious offense against the social ethics of science to pass off an exploration as a genuine testing procedure. Unfortunately, this can be done quite easily by making it appear as if the hypotheses had already been formulated before the investigation started. Such misleading practices strike at the roots of 'open' communication among scientists." (De Groot, 1969, p. 52). This point was later revisited by Kerr (1998) when he introduced the concept of HARKing ("Hypothesizing After the Results are Known"), as well as by Simmons et al. (2011); John et al. (2012), and Wagenmakers, Wetzels, Borsboom, and van der Maas (2011).

The third point that De Groot makes concerns preregistration. De Groot strongly felt that in order for research to qualify as confirmatory (and, consequently, for statistical inference to be meaningful), an elaborate preregistration effort is called for: "If an investigation into certain consequences of a theory or hypothesis is to be designed as a genuine testing procedure (and not for exploration), a precise *antecedent formulation* must be available, which permits testable consequences to be deduced." (De Groot, 1969, p. 69).

Indeed, De Groot proposed a detailed list of requirements such a preregistration document should adhere to. Specifically, De Groot (1969, p. 136) stated:

"Foremost (…) is the recommendation to *work out* in advance the investigative procedure (or experimental design) *on paper to the fullest possible extent*. This 'blueprint' should comprise: a brief exposition of the theory; a formulation of the hypothesis to be tested; a precise statement of the deductions leading up to the predictions to be verified; a description of the instruments – in the broadest sense – to be used, complete with instructions for their manipulation; detailed operational definitions of the variables to be used; a statement about the measurement scales (nominal, ordinal, interval, ratio) in which the respective variables are to be read (…); a clearly defined statement of the respective universes to which the hypothesis and the concrete prediction(s) apply; an exact description of the manner in which the samples are to be drawn or composed; a statement of the confirmation criteria, including formulation of null hypotheses, if any, choice of statistical test(s), significance level and resulting confirmation intervals (…); for each of the details mentioned, a brief note on their rationale, i.e., a justification of the investigator's particular choices."

It is unclear how exactly De Groot would have felt about the fact that virtually all current-day research in virtually all fields proceeds to tests hypotheses without anything resembling the "blueprint" that he envisioned, but from his writings it becomes clear that he considered such a practice suboptimal at best, and deeply misleading at worst. On a more optimistic note, it is encouraging and perhaps even surprising that in the last year alone, several journals — among which *Cortex*, *Perspectives on Psychological Science*,[5] *Attention, Perception, & Psychophysics* — have adopted special sections for preregistration of experiments (Chambers, 2013; Wolfe, 2013). Moreover, founding agencies such as the Laura and John Arnold Foundation have released new guidelines demanding that hypothesis testing research uses preregistration. The involvement of founding agencies was also anticipated by De Groot (1969, p. 138):

"This section is marked by a strongly argumentative, almost propagandistic tone. The reason for this is not hard to find. In research in the social and behavioral sciences, and in Continental research in particular, the importance of thorough preparations is all too often underestimated. There is undue eagerness to 'get the thing started' or to obtain results, and there is often too much fear of criticism, preventing consultations with experts and colleagues. (…) It is of the utmost importance that researchers and future researchers be trained in the techniques of experimental design described here. In addition, research promoters — foundations, government agencies — who provide the financial wherewithal, should be aware of their significance."

In sum, throughout his career in methodology De Groot made a compelling case against the meaningful interpretation of hypothesis tests for exploratory research, and in favor of (1) a strict distinction between exploratory and confirmatory research; and (2) preregistration to preserve and guard that distinction. To the best of our knowledge, De Groot's 1956 article is the first in which he clearly makes these points, points that later become pivotal in his monograph on the empirical cycle (De Groot, 1969). His consideration of the function of statistical practices in a broader statistical and philosophical framework is, to our knowledge, one of the earliest of its kind. The 1960s and 1970s would witness a greatly proliferating literature on conceptual issues in methodology, sparked by the work of American methodologists such as Paul Meehl, Lee Cronbach, Thomas Cook, and Donald Campbell. De Groot is perhaps one of the earliest methodologists to take this route.

Annotations to the introduction

Despite a litany of complaints, *p* value null hypothesis statistical testing is still the dominant methodology in the psychological literature. It should also be noted that De Groot, together with Hofstee and Drenth, strongly advocated the use of statistical methods in psychological research.

Annotations to Section 1: Hypothesis testing research versus material-exploration

The empirical cycle was the centerpiece of De Groot's later monograph "Methodology", first published in Dutch in 1961. This book was influential in The Netherlands, but its English translation had much less impact and is cited only 146 times according to Google Scholar (1-10-2013).

In the second paragraph, De Groot lists several ways in which the observed data can be re-organized in a search for "significant" effects. What is striking here is that De Groot acknowledges that although "the general intention (…) was probably determined beforehand", this intention alone does not safeguard the analysis against the dangers of exploration — De Groot's examples demonstrate how the researcher retains freedom to explore and analyze the data in many different ways. The comments by De Groot about the search for significance anticipate the recent articles by Simmons et al. (2011), showing how "undisclosed flexibility in data collection and analysis allows presenting anything as significant", and John et al. (2012), showing the prevalence of

[5] See http://www.psychologicalscience.org/index.php/replication.

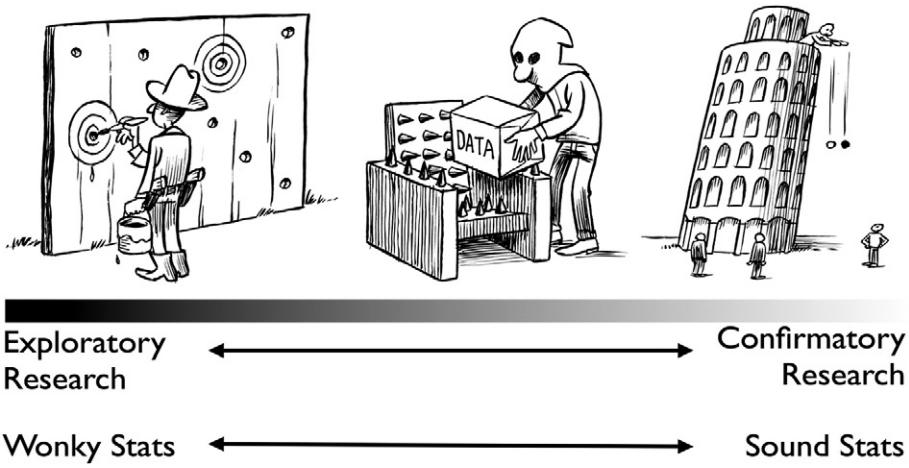

Fig. 1. A continuum of experimental exploration and the corresponding continuum of statistical wonkiness (Wagenmakers et al., 2012). On the far left of the continuum, researchers find their hypothesis in the data by post hoc theorizing, and the corresponding statistics are "wonky", dramatically overestimating the evidence for the hypothesis. On the far right of the continuum, researchers preregister their studies such that data collection and data analyses leave no room whatsoever for exploration; the corresponding statistics are "sound" in the sense that they are used for their intended purpose. Much empirical research operates somewhere in between these two extremes, although for any specific study the exact location may be impossible to determine. In the gray area of exploration, data are tortured to some extent, and the corresponding statistics are somewhat wonky.
Figure downloaded from Flickr, courtesy of Dirk-Jan Hoek.

"questionable research practices" that often express themselves through an undisclosed selection process.

Another important remark is that, for statistical tests to be meaningful, the hypotheses under consideration must have been "precisely formulated"; in addition, it is essential that "the details of the testing procedure (…) have been registered in advance". This plea for preregistration is reiterated in Section 5: "This claim can always be maintained, the 'proof' can always be provided, as long as one does not need to specify in advance, where exactly 'somewhere' is located." As mentioned above, in his later work De Groot argued for preregistration even more explicitly.

The dangers of exploration and the associated need for preregistration are hotly debated in the current literature. Wagenmakers et al. (2012) presented a continuum of experimental exploration, replotted here in Fig. 1. All the way on the left is the Texas sharp shooter, who fires randomly at the side of a barn and later paints the targets around his shots. This popular metaphor highlights the destructive power of post hoc theorizing and *ad hoc* data manipulation on the veracity of statistical results.

In August 2013, the Laura and John Arnold Foundation (LJAF) released new guidelines for research. These guidelines state, amongst others, that studies funded by LJAF have to preregister their statistical analysis plan. LJAF Director of Research Stuart Buck explained this with an analogy very similar to the Texas sharp shooter: "Pre-registration is like drawing a target publicly and then shooting an arrow at it; if you hit the target, it is a meaningful display of your shooting ability. Without pre-registration, however, there is so much flexibility with statistical analysis and dataset manipulation that you might as well shoot an arrow into the air and then later draw a target around wherever it landed."

Another similar analogy was drawn by Neuroskeptic (2012) in his article "The nine circles of scientific hell", where, in the Third Circle, the one reserved for researchers guilty of post-hoc storytelling, "Sinners (…) must constantly dodge the attacks of demons armed with bows and arrows, firing more or less at random. Every time someone is hit in some part of their body, a demon proceeds to explain at length that it was aiming for that exact spot all along." (p. 643)

Perhaps the only reliable weapon to combat the Texas sharp shooter, as De Groot suggests, is preregistration. This practice is mandatory for clinical trials (but see Goldacre, 2008, 2012):

> "For approval of a new drug through clinical trials, thanks to the close regulation exerted by government agencies, both the full experimental design and all endpoint measures to be used as proof of efficacy need to be specified *a priori*. This controlled experimental environment strikes a sharp contrast with most research in basic science as well as preclinical animal research, where there are no requirements to nominate primary endpoint measures, nor specify other design details. There is therefore a risk that exploration of the various endpoint measures that are available to the experimenter may reveal an effect that was not particularly expected." (Brunner, Balci, & Ludvig, 2012, p. 190)

Annotations to Section 2. Hypothesis testing research for a single hypothesis

In Footnote 1, De Groot refers to Spitz (1956). The Spitz article discusses the meaning of the significance level and points out that the proportion of hypothesis tests that are falsely rejected depends not only on the significance level but also on power and on the proportion of true null hypotheses in the population of hypotheses under consideration. This line of argumentation is similar to that in the famous article "Why most published research findings are false" (Ioannidis, 2005).

Annotations to Section 3. Hypothesis testing research for multiple hypotheses

The remarks about multiple comparisons are still relevant. If anything, the widespread availability of statistical software programs and the increasing ease of data collection have compounded the problem. Entire books can be written about this issue, but we feel that the problem is succinctly captured by one of Randall Munroe's better-known cartoons, "significance" or "do jelly beans cause acne?": http://xkcd.com/882/.

Annotations to Section 4. Material-exploration: N becomes unspecified

In this section De Groot suggests that the effect of exploration is similar to that of conducting multiple tests, except that in exploration the number of tests conducted or intended remains undetermined — consequently, it is impossible to correct statistically for the fact that a test was exploratory rather than confirmatory; "an exact interpretation of the meaning of 'positive results' is no longer possible".

It is interesting to consider the Bayesian perspective on multiple comparisons and exploratory data selection techniques (Dienes, 2011). In Bayesian statistics, one conditions on the observed data and, by the

likelihood principle, the intentions of the researcher are irrelevant; hence, the evidence that the data provide for a specific hypothesis is the same, regardless of whether that hypothesis was conceived *a priori* or post hoc, and regardless of whether that hypothesis was tested in isolation or together with many different hypotheses. What is relevant for Bayesian reasoning is not the number of tests that were executed or planned, but rather the prior belief in a particular hypothesis (Scott & Berger, 2010; Stephens & Balding, 2009). Thus, the adage "do not use the data twice" also holds for Bayesians: one should not use the data first to increase the prior odds of an unlikely hypothesis so that it becomes a viable candidate for testing, and second to test that hypothesis using those same data. Hence, preregistration is also useful from a Bayesian perspective, as it identifies the hypotheses that have appreciable prior odds, that is, the hypotheses that are worth testing.

Annotations to Section 5. Exploration of the behavior of a die

Extrasensory perception has always been a favorite subject among methodologists and statisticians. The scenarios sketched by De Groot resonate with our own concerns about a recent demonstration that people can look into the future (Bem, 2011; but see Wagenmakers et al., 2011 and also Galak, LeBoeuf, Nelson, & Simmons, 2012; Ritchie, Wiseman, & French, 2012; Wagenmakers et al., 2012).

The dangers of overly flexible, *ad hoc* data analysis techniques can also be illustrated with examples from modern-day neuroscience (Kriegeskorte et al., 2009; Vul et al., 2009). For instance, Carp (2012) reviewed 241 fMRI articles and concluded: "Many studies did not report critical methodological details with regard to experimental design, data acquisition, and analysis. Further, many studies were underpowered to detect any but the largest statistical effects. Finally, data collection and analysis methods were highly flexible across studies, with nearly as many unique analysis pipelines as there were studies in the sample. Because the rate of false positive results is thought to increase with the flexibility of experimental designs, the field of functional neuroimaging may be particularly vulnerable to false positives." (p. 289)

Annotations to Section 6. Conclusions

In this section, and throughout the entire article, De Groot anticipates the current "crisis of confidence" in psychology and other fields (Pashler & Wagenmakers, 2012). When exploratory research is statistically treated no different from confirmatory research, it may not come as a surprise that many research findings are not reproducible (i.e., as is the case in cancer research, see Begley & Ellis, 2012; Prinz, Schlange, & Asadullah, 2011; Wadman, 2013). But De Groot also points out that, in principle, quantitative research does allow one to get to the truth eventually: by replication research that does ensure that the analysis plan has been specified in advance (Chambers, Munafo, et al., 2013).[6] It is exactly these replication and preregistration efforts that are now supported by the Open Science Framework (Nosek & Bar–Anan, 2012; Nosek, Spies, & Motyl, 2012).

In the final paragraphs of this section, De Groot touches upon the difference between quantitative and qualitative research. De Groot himself preferred qualitative research, but only if it followed strict methodological rules.

In sum, De Groot was ahead of his time. His arguments about preregistration are based on a model of scientific research — the empirical cycle — that demands a strict distinction between exploratory and confirmatory research. The next years will have to show whether or not De Groot's ideals can form the basis of an academic revolution that is here to stay.

[6] Similar arguments are made by researchers on the internet, see for instance the blog posts by Dorothy Bishop (e.g., http://deevybee.blogspot.nl/2013/07/why-we-need-pre-registration.html), Chris Chambers (e.g., http://neurochambers.blogspot.nl/2012/10/changing-culture-of-scientific.html), NeuroSkeptic (e.g., blogs.discovermagazine.com/neuroskeptic/2013/04/25/for-preregistration-in-fundamental-research), and Rolf Zwaan (e.g., http://rolfzwaan.blogspot.nl/2013/01/pre-registration-at-journal-desk.html).

References

Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*, 531–533.

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 407–425.

Brunner, D., Balci, F., & Ludvig, E. A. (2012). Comparative psychology and the grand challenge of drug discovery in psychiatry and neurodegeneration. *Behavioural Processes, 89*, 187–195.

Carp, J. (2012). The secret lives of experiments: methods reporting in the fMRI literature. *NeuroImage, 63*, 289–300.

Chambers, C. D. (2013). Registered reports: a new publishing initiative at Cortex. *Cortex, 49*, 609–610.

Chambers, C. D., Munafo, M., et al. (2013). Trust in science would be improved by study pre-registration. *The Guardian* (http://www.theguardian.com/science/blog/2013/jun/05/trust-in-science-study-pre-registration).

De Groot, A. D. (1969). *Methodology: foundations of inference and research in the behavioral sciences*. The Hague: Mouton.

Dienes, Z. (2011). Bayesian versus orthodox statistics: which side are you on? *Perspectives on Psychological Science, 62*, 74–290.

Galak, J., LeBoeuf, R. A., Nelson, L. D., & Simmons, J. P. (2012). Correcting the past: failures to replicate psi. *Journal of Personality and Social Psychology, 103*, 933–948.

Goldacre, B. (2008). *Bad science*. London: Fourth Estate.

Goldacre, B. (2012). *Bad pharma: how drug companies mislead doctors and harm patients*. London: Fourth Estate.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*, 696–701.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth—telling. *Psychological Science, 23*, 524–532.

Kerr, N. L. (1998). HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review, 2*, 196–217.

Kriegeskorte, N., Simmons, W. K., Bellgowan, P. S. F., & Baker, C. I. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nature Neuroscience, 12*, 535–540.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model modifications in covariance structure analysis: the problem of capitalization on chance. *Psychological Bulletin, 111*, 490–504.

Neuroskeptic (2012). The nine circles of scientific hell. *Perspectives on Psychological Science, 7*, 643–644.

Nosek, B. A., & Bar–Anan, Y. (2012). Scientific utopia: I. Opening scientific communication. *Psychological Inquiry, 23*, 217–243.

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*, 615–631.

Pashler, H., & Wagenmakers, E. J. (2012). Editors' introduction to the Special Section on Replicability in Psychological Science: a crisis of confidence? *Perspectives on Psychological Science, 7*, 528–530.

Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: how much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery, 10*, 712–713.

Ritchie, S. J., Wiseman, R., & French, C. C. (2012). Failing the future: three unsuccessful attempts to replicate Bem's 'retroactive facilitation of recall' effect. *PLoS One, 7*, e33423.

Scott, J. G., & Berger, J. O. (2010). Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics, 38*, 2587–2619.

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

Spitz, J. C. (1956). De Betekenis van het SignificantieniveauDe betekenis van het significantieniveau. *Nederlands Tijdschrift voor de Psychologie en Haar Grensgebieden, 11*, 410–418.

Stephens, M., & Balding, D. J. (2009). Bayesian statistical methods for genetic association studies. *Nature Reviews Genetics, 10*, 681–690.

Vul, E., Harris, C., Winkielman, P., & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*, 274–290.

Wadman, M. (2013). NIH mulls rules for validating key results: US biomedical agency could enlist independent labs for verification. *Nature, 500*, 14–16.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., & van der Maas, H. L. J. (2011). Why psychologists must change the way they analyze their data: the case of psi. *Journal of Personality and Social Psychology, 100*, 426–432.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*, 627–633.

Wolfe, J. M. (2013). Registered reports and replications in attention, perception, & psychophysics. *Attention, Perception, & Psychophysics, 75*, 781–783.