PLOS ONE

Testing for Questionable Research Practices in a Meta-Analysis: An example from Experimental Parapsychology

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Article
Full Title:	Testing for Questionable Research Practices in a Meta-Analysis: An example from Experimental Parapsychology
Short Title:	SIMULATION OF QUESTIONABLE RESEARCH PRACTICES
Corresponding Author:	Dick Jan Bierman, Ph.D. Universiteit van Amsterdam Amsterdam, NETHERLANDS
Keywords:	Questionable Research Practices; Meta-analyses; replication; experimental parapsychology; Ganzfeld-telepathy; deception; simulation; fit using Genetic Algorithms
Abstract:	Meta-analyses with P-values between 10^-6 and 10^-16 have been cited as evidence of paranormal anomalies. We use simulations of a meta-analysis of the Ganzfeld-telepathy protocol to assess the extent to which these experimental results could be explainable by Questionable Research Practices (QRPs). Our simulations used the same numbers of studies and trials as the original meta-analysis. Results of both meta-analysis and simulations were characterized by 4 metrics, two describing the trial and mean experiment hit rates (HR) of around 31%, where 25% is expected by chance, one the correlation between sample-size and hit-rate, and one the P-value distribution of the database. A genetic algorithm optimized the parameters describing the QRPs, and the fitness of the simulated meta-analysis was defined as the sum of the squares of Z-scores for the 4 metrics. Assuming no anomalous effect a good fit to the empirical meta-analysis was found only by using QRPs with unrealistic parameter-values. Restricting the parameter space to ranges observed in studies of QRP occurrence, under the untested assumption that parapsychologists use comparable QRPs, the fit to the published Ganzfeld meta-analysis with no anomalous effect was poor. We allowed for a real anomalous effect where the HR ranged from 25% (chance) to 31%. With an anomalous HR of 27% the fitness became F =1.8 (P =0.47 where F=0 is a perfect fit). We conclude that the very significant probability cited by the Ganzfeld meta-analysis is likely inflated by QRPs, though results are still significant (P=0.003) with QRPs. Our study demonstrates that quantitative simulations of QRPs can assess their impact. Since meta-analyses in general might be polluted by QRPs, his method has wide applicability outside the domain of experimental parapsychology.
Order of Authors:	Dick Jan Bierman, Ph.D.
	James Spottiswoode
	Aron Bijl
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure Please describe all sources of funding that have supported your work. A complete funding statement should do the following: Include **grant numbers and the URLs** of	The authors received no specific funding for this work

any funder's website. Use the full name, not acronyms, of funding institutions, and use initials to identify authors who received the funding. **Describe the role** of any sponsors or funders in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. If they had no role in any of the above, include this sentence at the end of your statement: "*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript*." If the study was **unfunded**, provide a statement that clearly indicates this, for example: "*The author(s) received no specific funding for this work*." * typeset	
Competing Interests You are responsible for recognizing and disclosing on behalf of all authors any competing interest that could be perceived to bias their work, acknowledging all financial support and any other relevant financial or non-financial competing interests. Do any authors of this manuscript have competing interests (as described in the PLOS Policy on Declaration and Evaluation of Competing Interests)? **If yes**, please provide details about any and all competing interests in the box below. Your response should begin with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests:* **If no** authors have any competing interests to declare, please enter this statement in the box: "*The authors have declared that no competing interests exist*." * typeset	The authors have declared that noi cvompeting interests exist
Ethics Statement	Since this study used public data obtained by other researchers no explicit Ethics

	committee approval is required.
You must provide an ethics statement if your study involved human participants, specimens or tissue samples, or vertebrate animals, embryos or tissues. All information entered here should **also be included in the Methods section** of your manuscript. Please write "N/A" if your study does not require an ethics statement.	
Human Subject Research (involved human participants and/or tissue) All research involving human participants must have been approved by the authors' Institutional Review Board (IRB) or an equivalent committee, and all clinical investigation must have been conducted according to the principles expressed in the Declaration of Helsinki. Informed consent, written or oral, should also have been obtained from the participants. If no consent was given, the reason must be explained (e.g. the data were analyzed anonymously) and reported. The form of consent (written/oral), or reason for lack of consent, should be indicated in the Methods section of your manuscript.	
Please enter the name of the IRB or Ethics Committee that approved this study in the space below. Include the approval number and/or a statement indicating approval of this research.	
Animal Research (involved vertebrate animals, embryos or tissues) All animal work must have been conducted according to relevant national and international guidelines. If your study involved non-human primates, you must provide details regarding animal welfare and steps taken to ameliorate suffering; this is in accordance with the recommendations of the Weatherall report, "The use of non-human primates in research." The relevant guidelines followed and the committee that approved the study should be identified in the ethics statement.	
If anesthesia, euthanasia or any kind of animal sacrifice is part of the study,	

please include briefly in your statement which substances and/or methods were applied. Please enter the name of your Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board, and indicate whether they approved this research or granted a formal waiver of ethical approval. Also include an approval number if one was obtained. **Field Permit** Please indicate the name of the institution or the relevant body that granted permission.	
Data Availability PLOS journals require authors to make all data underlying the findings described in their manuscript fully available, without restriction and from the time of publication, with only rare exceptions to address legal and ethical concerns (see the PLOS Data Policy and FAQ for further details). When submitting a manuscript, authors must provide a Data Availability Statement that describes where the data underlying their manuscript can be found. Your answers to the following constitute your statement about data availability and will be included with the article in the event of publication. **Please note that simply stating 'data available on request from the author' is not acceptable.** ***If, however, your data are only available upon request from the author(s), you must answer "No" to the first question below, and explain your exceptional situation in the text box provided.*** Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	Yes - all data are fully available without restriction
Please describe where your data may be found, writing in full sentences. **Your answers should be entered into the box below and will be published in the form you provide them, if your manuscript is accepted.** If you are copying our sample text below, please ensure you replace any instances of XXX with the appropriate details.	All relevant data are within the paper and its Supporting Information files

If your data are all contained within the paper and/or Supporting Information files, please state this in your answer below. For example, “All relevant data are within the paper and its Supporting Information files.” If your data are held or will be held in a public repository, include URLs, accession numbers or DOIs. For example, “All XXX files are available from the XXX database (accession number(s) XXX, XXX)." If this information will only be available after acceptance, please indicate this by ticking the box below. If neither of these applies but you are able to provide details of access elsewhere, with or without limitations, please do so in the box below. For example: “Data are available from the XXX Institutional Data Access / Ethics Committee for researchers who meet the criteria for access to confidential data.” “Data are from the XXX study whose authors may be contacted at XXX.” * typeset	
Additional data availability information:	

September 14, 2015

To: The Editor PLOS ONE

Dear Editor,

I am enclosing a submission to *PLOS ONE* entitled "Testing for questionable research practices in a meta-analysis: An example from experimental parapsychology". The manuscript is about 40 pages long (~10000 words) and includes 3 tables embedded in the text (included in the page and word count).
The 5 figures are given in separate EPS files.

Two recent papers have disturbed psychology. The frequency of questionable research practices (QRPs) in psychology has been surveyed by John et al[1] and a surprisingly low replication rate has been described by the Open Science Collaboration[2].

Our paper reports the first study (that we know of) which simulates and quantifies the impact of QRPs on meta-analyses (MAs), which are ubiquitous in psychological, medical and other fields, and whose conclusions have substantial social and scientific consequences. Our method is generally applicable to MAs outside of the specific one we have analyzed.

This is our first interaction with PLOS regarding this paper.

The possible PLOS editor to oversee the reviewing:
Jelte M. Wicherts, Tilburg University, NETHERLANDS

Possible reviewers (we are unaware of any supporting or opposing views)
1. Jonathan Schooler, UCSB, Email: jonathan.schooler@psych.ucsb.edu
2. Eric-Jan Wagenmakers, University of Amsteram, email: EJ.Wagenmakers@gmail.com
3. Daniël Lakens , Eindhoven University of Technology, email: D.Lakens@tue.nl
4. Simon Thorpe, CNRS, email: Thorpe@cerco.ups-tlse.fr
5. Jacob Jolij, University of Groningen, email: j.jolij@rug.nl

First and corresponding author: Dick J. Bierman, University of Amsterdam, Weesperplein 4, 1018 XA,Amsterdam, The Netherlands. Phone +31 20 5256727. Email: d.j.bierman@uva.nl

Second author: James Spottiswoode, LFR, Palo Alto

---

[1] John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci 2012 May 1;23(5):524-532.
[2] Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). Doi: 10.1126/science.aac4716

Third author: Aron Bijl, University of Amsterdam.
Aron was a student assistant of the first author and in the mean time finished his masters degree.

All authors have agreed to the byline order and to submission of the manuscript in this form. I have assumed the responsibility for keeping my coauthors informed of the progress through the editorial review process, the content of the reviews, and they will be involved in any revision that mat be required. I understand that, if accepted for publication, a certification of authorship form will be required that all contributors will sign,

Sincerely,

Prof. dr D.J.Bierman, PhD
University of Amsterdam

Testing for Questionable Research Practices in a Meta-Analysis: An Example from Experimental Parapsychology

Dick J. Bierman [1*], James P. Spottiswoode [2], Aron Bijl [1]

[1] Dept. of Brain & Cognition, University of Amsterdam, Amsterdam, The Netherlands

[2] LFR, Palo Alto, CA, USA

***** Corresponding author
E-mail: d.j.bierman@uva.nl (DJB)

Abstract

Meta-analyses with P-values between 10^{-6} and 10^{-16} have been cited as evidence of paranormal anomalies. We use simulations of a meta-analysis of the Ganzfeld-telepathy protocol to assess the extent to which these experimental results could be explainable by Questionable Research Practices (QRPs). Our simulations used the same numbers of studies and trials as the original meta-analysis. Results of both meta-analysis and simulations were characterized by 4 metrics, two describing the trial and mean experiment hit rates (HR) of around 31%, where 25% is expected by chance, one the correlation between sample-size and hit-rate, and one the P-value distribution of the database. A genetic algorithm optimized the parameters describing the QRPs, and the fitness of the simulated meta-analysis was defined as the sum of the squares of Z-scores for the 4 metrics. Assuming no anomalous effect a good fit to the empirical meta-analysis was found only by using QRPs with unrealistic parameter-values. Restricting the parameter space to ranges observed in studies of QRP occurrence, under the untested assumption that parapsychologists use comparable QRPs, the fit to the published Ganzfeld meta-analysis with no anomalous effect was poor. We allowed for a real anomalous effect where the HR ranged from 25% (chance) to 31%. With an anomalous HR of 27% the fitness became F =1.8 (P =0.47 where F=0 is a perfect fit). We conclude that the very significant probability cited by the Ganzfeld meta-analysis is likely inflated by QRPs, though results are still significant (P=0.003) with QRPs. Our study demonstrates that quantitative simulations of QRPs can assess their impact. Since meta-analyses in general might be polluted by QRPs, his method has wide applicability outside the domain of

experimental parapsychology.

Introduction

Recently, it has become clear that experimental research, most notably in social psychology and cognitive neuroscience, but also in the medical literature, is plagued by questionable research practices (QRPs) (1). John, Loewenstein & Prelec (2012, hereafter referred to as JLP) surveyed 2,000 psychologists and determined the frequency or prevalence of QRPs involving post hoc selection of studies, dependent variables, hypotheses, statistical techniques, stopping rules, data-transformations (like outlier treatments) or conditions. QRPs result in inflated P-values and some authors have suggested that some published research findings are actually incorrect due to QRPs, which bias the findings (2). According to Simmons et al. (2011), undisclosed flexibility in data collection and analysis makes it theoretically possible to present too many datasets as 'significant' (3). They reported simulations of these practices substantiating this claim. It is important to quantitatively evaluate how much inflation can occur due to QRPs, allowing an estimate of its impact, especially for meta-analyses (MA). MAs are sensitive to small systematic errors that may accumulate. Most MAs have only investigated how many studies must go into a file drawer to render meta-analytic effect estimates non-significant, but have refrained from analyzing quantitatively the potential contribution of other QRPs.

Very recently the replication rates in Psychology were assessed in a collaborative project (4). While 97% of the original 100 studies were statistically significant, only 36% of the 100 replications were. No explanation is given for the failures to replicate but no signs of deception or methodological errors were found in the original reports.

We consider QRPs in the context of a meta-analysis database of Ganzfeld–telepathy experiments from the field of experimental parapsychology. The Ganzfeld database is particularly suitable for this study, because the parapsychological phenomena it investigates are widely believed to be nonexistent. On the assumption that they are, the dataset would be a good example of how purely random data can be manipulated into the appearance of significance. While we do not claim that assumption is valid, it provides a unique test case for simulations.

Experimental Parapsychology

Experimental parapsychology uses generally accepted scientific methods to study alleged anomalous phenomena such as telepathy, clairvoyance, precognition and psychokinesis. The field is relatively small, partly because there is little funding available for parapsychological research and because publishing on these controversial topics can be detrimental to a scientific career. The publications generated by serious parapsychological researchers during the last 150 years correspond to only about 3 months of research by current experimental psychologists (5). Nonetheless, researchers have amassed large databases of experimental outcomes for each of the anomalous phenomena mentioned above, and each of these databases, taken at face value, strongly suggests the existence of an anomaly. Most proponents of parapsychology claim that these meta-analytic results are the ‘best evidence’ for paranormal phenomena (6,7).

This interpretation has been criticized on several grounds. Firstly, individual experiments have been criticized on grounds of the potential for sensory leakage and lack of effective control of normal explanations. Over the years this type of criticism

has resulted in improving study quality. Nowadays, simple explanations like sensory leakage and inadequate randomization are generally no longer applicable.

Secondly, meta-analytic datasets, showing larger than 6-sigma results, have been criticized because the interpretation of the databases should take into account unpublished studies (the file drawer).

The main question that we will investigate is whether the results from the individual studies, which contribute to these meta-analytic results, could be quantitatively explained by QRPs.

JLP conclude: "… Some questionable research practices may constitute the prevailing research norm…." Many of their surveyed respondents did indeed consider these practices to be acceptable. Most researchers in the field of parapsychology are psychologists and it could be argued that their attitudes should be no different from those often found among other experimental psychologists. There are, however, some potential reasons to expect differences. Parapsychologists may be more sensitized to these issues due to strong scrutiny, even hostility, to which their work is subjected (8). Additionally, publication policy in the parapsychology field is such that it is possible to publish non-significant results, and a non-significant outcome is not a danger to the career of the parapsychologist. On the other hand, many parapsychologists may be more driven by a non-materialist or spiritual world view. The cautious approach therefore is to assume that the prevalence values of the use of QRPs in parapsychological research are similar to those measured by JLP in experimental psychology.

The Ganzfeld telepathy experiment

In a Ganzfeld (GF) trial the subject (the 'receiver') is stimulated with white noise and a homogeneous non-patterned red light, and reports his/her experiences,

while at a distant location somebody (the 'sender') attempts to 'send' information relating to a randomly selected target (a picture or movie clip). At the end of the experiment the subject has to select an unhandled copy of the actual target from a set of 4 images or clips (1 target + 3 decoys). When the subject selects the target it is called a 'hit'. Thus one subject generally contributes one data point and simulation of a GF experiment using this protocol is a simple and straightforward matter. The Mean Chance Expectation (MCE) hit rate (HR) is 25%. The average GF study reports a HR around 31% rather than the expected 25%.

The Ganzfeld database

History of the database

GF experiments started in the 1970s and in 1985 claims were made that the 44 studies published to date had a mean HR of 35% and constituted strong evidence for a true anomaly (9). Hyman (1985) disagreed and presented several weaknesses in these data, including some practices that we today would label as QRPs. Most importantly, Hyman demonstrated that almost all studies used multiple outcome measures, i.e. different variables that would be a measure for telepathy. In the early GF work different researchers had used 5 different indices: the direct HR, the binary HR, the sum of ranks, the normalized rating and a 10-bit binary scoring method. If the conservative Bonferroni correction had been applied, studies that reported a P-value of 0.05 actually should been considered to have a P-value of 0.25. However, these measures are highly correlated and therefore the Bonferroni correction is over-conservative. In order to estimate the required true correction to the P-values, Hyman ran simulations of a GF experiment with a fixed number of trials. He concluded that when a P-value of 0.05 was claimed, then in reality this P-value must have been close

to p = 0.10. Honorton (1985) responded by evaluating the 28 studies (of the 44) where the direct HR was presented or could be inferred (9). He claimed that this analysis was now free of the QRP of using multiple indices and he reported still a highly significant combined (Stouffer) Z-score of 6.6 ($p= 2.06 * 10^{-11}$). Honorton also responded to Hyman's list of procedural weaknesses; for example, that about half of the studies failed to use duplicate target sets and hence sensory leakage through marks on the target picture could have occurred.

As a result of the controversy between this skeptic and this parapsychologist, Hyman and Honorton came together and wrote a joint communiqué (10). They promoted pre-registration, and had some explicit recommendations for GF protocols. Following the communiqué, GF experiments were much more standardized, often automated and had hardly any procedural weaknesses and almost all used a single outcome measure.

The most complete published meta-analysis of GF experiments today pools the 28 studies before the joint communiqué with 80 studies published since 1985 (11). Storm et al removed 6 outliers (3 at the high HR side and 3 on the low HR side) in order to produce a homogeneous dataset. They were left with 102 studies, 24 from the pre-1985 period and 78 from the period from 1985 till 2010. The *mean* study HR in this database was 32.3% (MCE=25%) and the compound over-all HR is 31.5%. The difference is due to a slightly larger HR in smaller studies. Storm et al. (2010) report the over-all meta-analytic results as follows:

"…The homogeneous database consists of 102 studies: mean z= 0.81 (SD =1.23; range: □2.30 to 4.32), … and Stouffer Z = 8.13 ($p = 10^{-16}$). … With Rosenthal's (1995, p. 189) file drawer formula, there would have to be approximately 2,414

unpublished and non-significant papers in existence to reduce our significant Stouffer *Z* to chance…".

Note the extremely small P-value of $p \sim 10^{-16}$. The goal of the current study is to determine how much of this apparently significant result may be explained by QRPs.

Focusing on the modern dataset

While the pre-1985 dataset could be affected by the QRP of multiple outcome measures, this is not applicable to the post-1985 database. If that were the only difference we could have simulated the whole combined data set by allowing one extra QRP on the early studies. The procedural weakness of using a single target-set in the pre-1985 studies is impossible to simulate but had apparently no effect on the results because the mean HR for the single target-set studies were smaller. However, Hyman also noted other procedural weaknesses such as 'inadequate randomization' and 'inadequate security', rejecting any study that used only one experimenter. In such studies there is a risk that the experimenter who is present at the judging is aware of the target or has a reasonable guess at it because (s)he is also involved in the target selection procedure. In the judging procedure at the end of a GF session, the experimenter generally plays an active role and can easily influence the subject's choice in any direction. A skeptical observer who visited Eysenck's parapsychology lab in 1980 and attended a number of GF sessions noted irregularities in the (very complex manual) randomization procedure that potentially would allow for one of the experimenters to get information about the target (Blackmore, 1987). The case was investigated by a committee appointed by the professional organization of parapsychologists, the Parapsychological Association (PA). The committee, consisting of PA members and a member of the professional Skeptical

organization, concluded that no indications of deception could be found. Their report however never became public, leaving an unwarranted taint.

Given the weaknesses and criticisms of the pre-1985 studies, we will focus on the post-1985 dataset for our further simulations. In that post-1985 dataset there were two contributions by the lab that had been accused of irregularities in the randomization procedure. We removed these two studies from the database, and will return to this issue when we discuss how to simulate 'deception'. We included the studies that were removed to obtain homogeneity, as they were run after 1985. Finally, we removed one study that used a judging procedure with 7 decoys rather than the standard 3 decoys. Thus our final data set included 78 studies which is publicly available and is supplied as supporting information

Sample size distribution of the modern dataset

Fig. 1 shows the actual distribution of sample sizes in the GF database. The mean number of participants (and hence data points) in the GF studies in this database was 44.8. Since for some of the QRPs it can be argued that the practice is dependent on the sample size, we used the actual distribution of experimental sample sizes in our simulations.

Fig 1. Distribution of Sample Sizes (= number of trials) in the database

Hit Rates vs sample size

The mean HR (mHR) over the 78 studies is 31.15 %, but if all experiments are pooled there are in total 1083 hits in 3494 trials, a total HR of 30.99 %. This difference is partly due to the negative correlation between HR and number of trials (N). Note that HR can be employed as an effect size (ES) as it is independent of N. The Cohen's *d* ES is defined as $d = Z/\sqrt{N}$ and *Z* is given by

$$Z = \frac{N(HR - 0.25)}{\sqrt{N \cdot 0.25 \cdot 0.75}}$$

from which follows the (linear) relation between ES and HR:

$$ES \approx 2.13(HR - 0.25)$$

The effect size (d) of this database is ~ 0.138. In the simulations and further presentation of results we will use HRs throughout.

It can be observed that non-significant and negative results have been published because parapsychological journals do not reject negative findings (fig 2). According to a regression analysis, as suggested by Egger (12), the asymmetry in the funnel plot is not significant (p = 0.25).

Fig 2. Funnel plot of hit rates in the database

We obtained a table converting the P-value of the regression into % of studies in the file drawer by simulation (table 1). From this table we may conclude that the file drawer is certainly smaller that 1 unpublished study per published study because the actual P-value of 0.25 is in the range between 0.10-0.30.

Table 1. Simulated estimate of the file drawer as a function of funnel plot asymmetry

P-value of regression analysis	Unpublished per published
0.004	180%
0.04	132%
0.10	105%
0.30	60%

.

P-value distribution

The binomial P-value distribution (Fig. 3) has many more significant studies than would be expected by chance. The empirical P-value distribution is significantly different from the chance null distribution (chi^2 = 25.4, df=9, P < 0.003). A closer look at the empirical distribution indicates that the intervals with the largest differences from the null distribution are $0 < p < 0.1$ and $.7 < p < 1$. In simulations of single and multiple QRPs we will report the chi^2 comparing this empirical P-value distribution with the simulated one.

Fig 3. Distribution of binomial P-values. Upper pane: the experimental GF database. Lower pane: Simulated Null-distribution.

Correlation between hit rate and sample size

In principle HR should be independent of sample size. In the current database the non-parametric correlation between these two variables is -0.112. This is a quite small and non-significant relation. However we feel justified in using this correlation as one of the fitting parameters because in the total GF database of 102 studies the negative correlation, Spearman's rho, is -0.206 (p = 0.038, two-tailed). Furthermore these negative correlations have been found in many meta-analyses within parapsychology. as well as within psychology in general, so we conclude that this is a real aspect of the data (13). It is interesting to note that by removing the part of the database that contains the less rigorous pre-1985 studies the negative correlation has dropped. This suggests that the negative correlation may be associated with the use of QRPs.

We will require our simulations to replicate not only the empirical HRs but also the internal effects like the negative correlation between HR and sample size and the peculiar P-value distribution.

Questionable Research Practices

Non applicable QRPs

There are a few known QRPs that are not applicable in the post-1985 database because of the specific GF research paradigm and its standardization. We describe these QRPs below.

The paper fails to report all dependent variables (JLP QRP 1)

From the discussion of the GF database above, it is straightforward to see that in the post 1985 GF analyses there is no freedom to select a dependent variable from many. In all studies in the database the dependent variable was the 'number of hits'. We suspect that in a few studies independent judging was employed but the subject scoring was also presented as primary measure.

Failing to report all conditions (JLP QRP 3)

In the case of evidential GF there may be more conditions because the research question could go beyond the simple question of the reality of parapsychological (or psi) phenomena. For all studies that had more than one condition, the results were collapsed over all conditions and the meta-analysis only used the *over-all* HR of a study thereby reducing the number of conditions to one.

Rounding off P-values (JLP QRP 5)

We calculated P-values directly from the numbers of hits and trials.

Claiming larger generalizability (JLP QRP 8 & 9)

Reporting an unexpected finding as having been predicted, and claiming larger generalizability than is justified, are irrelevant here, because this practice does not have any impact on the values of the relevant dependent variable, HR.

Applicable QRPs

These QRPs are still applicable to post-1985 research. For each of the applicable questionable research practices there are two or three parameters (see table 2). For each practice, 'prevalence' is the relative number of experimenters engaging in the practice. JLP's estimates are given in table 1 as well as the interval of prevalences that we allow if we run the constrained fitting procedure. The other parameters of each QRP are discussed in the relevant paragraph.

Fraud (JLP QRP 10)

Fraud was surprisingly admitted to by about 0.6% of the respondents in the JLP study. After application of the Bayesian Truth Serum algorithm (BTS), that takes into account what the respondents think about what *other* researchers are doing, this practice is estimated to be present in 1.7% of the reported studies. Of all QRPs, application of the BTS has by far the largest impact on this specific practice of falsifying data. It is difficult to assess if this prevalence also holds for psi researchers. In the past 50 years, two high profile fraud cases by parapsychologists have been reported, the Levy case and the Soal case (14). Soal had been accused after his death

of post hoc editing raw data. It should be mentioned in all fairness that there are currently doubts that these accusations against Soal are correct. Both these cases concerned researchers known for their long years of work in the field who had published numerous papers. The total number of psi researchers with similar publishing records is estimated to be 200. It therefore seems that the percentages of fraudulent researchers in the field of experimental psychology and of experimental parapsychology are similar., i.e. between 1% and 2%. It is impossible to simulate this practice so we had to correct the database by removing studies before we ran the simulations. We decided on the basis of the arguments given above that one senior researcher in the 80 studies post-1985 database 'had to be' guilty of deception. In order to take into account the contribution of deceptive research to the database we thus removed the two studies of one senior researcher, the person that had been implicated in errors in the randomization procedures. These studies were quite significant with HRs of 35% and 41%. In the database there are 29 principal investigators, so removal of one of them (3.4%) exceeds the prevalence of deception found by JLP (1%-2%) and is therefore conservative.

Confirmation to Pilot (CtoP)

We identified a version of optional stopping (JLP QRP 4) that produces a file drawer that is generally not recognized. CtoP occurs when a confirmatory experiment is started but then halted after a few subjects (parameter: Trialnr) when results do not meet expectations, for instance in the observed HR (parameter: P-crit). The experimenter then adjusts some aspect of the protocol and restarts the experiment afresh. The data from the initial unsuccessful sessions are discarded. This data is in fact properly part of the file drawer, though it is frequently not treated as such because in general one thinks of the file drawer as containing finished studies. This QRP is described in our simulations by its prevalence and two parameters: the number of hits below which the experimenter decides to stop the experiment and the trial number at which the experimenter looks at the cumulated number of hits.

Pilot to Confirmation (PtoC)

This particular QRP has not been extensively discussed in the literature yet. Most labs and researchers, when starting a GF experiment, will run some pilot trials (parameter N-trials) to get experience with the equipment and the rather complex procedure. The intention is to check everything and, if no problems are encountered, to run a confirmatory experiment. If the pilot trials are technically successful (parameter: pCrit) and show a promising HR, then the experimenter may consider the pilot data as a part of the larger trial. This QRP was described by its prevalence and two parameters: the number of trials in the pilot and the cumulated number of hits above which the pilot is added to the confirmatory experiment.

Optional stopping (OS)

Simulation of OS showed that optional stopping has no noticeable effect on the obtained HR (see results table 3). This was somewhat surprising since students are generally taught that this particular flexibility in experimenting practice gives misleadingly inflated scores. Many articles have appeared on the effects of optional stopping or, as it was also called, 'repeated significance testing' on accumulating data (15). Interestingly, one of those articles appeared in the *Journal of Parapsychology*: Feller (16) claimed that most results of Rhine's telepathy card guessing-tests could be explained by the practice of optional stopping. In hindsight this was probably a false conclusion. However OS has an effect on the P-value distribution (17) and the correlation between sample size and effect size. Apart from the prevalence, there is one parameter, the trial number at which repeated significance testing starts. The criterion to stop was always set to $p < 0.05$.

Optional Extension (OE)

This QRP, often considered to be a special form of Optional Stopping, is probably the most well known QRP. If the experiment has not quite reached $p < 0.05$ (parameter: P-crit) a number of extra, unplanned, trials (parameter extra-N) are added in the hope that the final P-value will be less than 0.05. Aside from the prevalence parameter it has two parameters: the P-value interval at which the experiment will be extended and the maximum number of extra trials generally constrained by time or money.

Selectively reporting studies that have significant results: publication bias (PB)

This QRP occurs when experimental questions are addressed using several studies and only the study with significant results is published. In the GF database the research question for each experiment is identical; hence in this context we can consider the file drawer as representing the unreported studies.

Post hoc selection of studies for publication, producing a file drawer of unpublished studies, has been extensively discussed in the literature. The usual way to treat this problem in parapsychology has been to calculate how many unpublished studies would be required to eliminate the reported effect using either fail-safe formula (18) or P-curve analysis (19). For instance, Storm et al (2010) calculated that 2,414 unpublished studies were required to eliminate the overall results of the GF database. It is argued by Storm et al that this number, given the limited resources of the field and the acceptance of publishing negative findings, is unreasonably large. Scargle (2000) pointed out however that when calculating this failsafe number, it is generally incorrectly assumed that the decision not to publish is unbiased. This has been shown to be incorrect.

For instance Franco et al. (2014) analyzed the file drawer effect in a group of experiments known as TESS (Time-sharing Experiments in the Social Sciences), known for its quality and consisting of a known population of studies with full accounting of both published and unpublished ones. They classified the studies both on results (null, mixed & strong results) and on publication status. Of the total number of studies investigated (221), about half were published. Only 20% of those with null results were published, while studies with strong and mixed results had publication rates of 60% and 50% respectively. A survey of all finished studies at the Koestler Parapsychology Unit of the University of Edinburgh revealed that 15% of the non-significant and 70% of the significant studies were reported (20). These figures seem

to confirm the publication bias observed in the TESS studies. Thus, in both fields there was a strong bias, the probability of publication increased by about 40%-55% when comparing strong results to null ones. When TESS researchers were asked why they didn't publish null results, 15 out of the 26 respondents reported abandoning the project because of the low publication potential (even if they found these results interesting). A smaller group (9) reported decreasing the priority of writing up null results, in favor of other projects. When the filing to the drawer is biased in this way a significantly smaller number of studies residing in the file drawer can compensate the overall published results. We will use a publication probability function derived from the empirical findings of Franco et al. To avoid using a likely unrealistic discontinuous PB function, we fitted the Franco step function with a continuous function:

$$pubprob = \frac{25 + 40\left(\tanh\left(2 - 10p\right) + 1\right)}{100}$$

where *pubprob* is the probability that the study will be published and *p* is the P-value of the study. This QRP has no free parameters (except of course the prevalence for that practice).

Deciding to exclude data post hoc (RmSs)

As formulated, this QRP appears basically fraudulent. However JLP report a defensibility rating of 1.61 (0= indefensible, 1= possibly defensible, 2= defensible). Most probably, researchers are thinking here about removal of outliers, not of subjects. In the case of GF research there are no outliers, since there is but one data point per subject. In a lot of psychological research it is common practice (but still questionable) to remove subjects on the basis of reasons that weren't specified *fully* in advance. For example, if a subject falls asleep during an experimental test, complains extensively about the temperature, or is late, researchers could argue that this subject should be removed. The problem is that if such a subject has results that conforms to the cherished hypothesis, the experimenter may be non-consciously less inclined to make the subjective decision to remove the subject. If subject removal happens blindly before inspection of the data, then the practice will not introduce a bias. However in GF research experimenters are generally not blind to the outcome of a session.

This QRP was modeled by its prevalence and the percentage of subjects that was removed. The minimum of removed subjects was set to one (if the QRP occurred at all determined by a random decision based on its prevalence). The idea is that even in small studies where a fixed percentage would result in a number below 1, an experimenter generally can find a situation that 'justifies' removal of one subject.
Removing a larger percentage than 5% of subjects (parameter '% of N') with misses will make the *post hoc* arguments for removal that an experimenter has concocted more and more artificial and will basically turn this practice into fraud (see 'Fraud').
This QRP is akin to the asymmetric behavior of experimenters with respect to experiments that are close to p=0.05. If p is slightly over 5% they will check their data

and methods while when p is slightly smaller than 0.05 they might be less inclined to do so.

Method

Simulation of QRPs

Simulation software was written in Real Basic (2011, Release 4.3), and developed independently in R (21). The software is publicly available in the Supporting Information. Each trial in a simulated experiment had the probability of a hit preset to 25% when simulating no real GF telepathy effect. Simulations generally proceeded in 300-500 sets of 78 experiments for a total of ~30,000 simulated experiments per run. The sample sizes for the simulated experiments were selected from the actual distribution of sample sizes in the database in such a way that after one set all sample sizes of the meta-analysis had been used. For each simulated experiment, the QRPs were applied probabilistically. For each QRP there was a prevalence figure representing the probability that an experimenter would be 'using' this particular QRP. Then in each simulated experiment a random decision was taken whether to apply each QRP with the probability equal to that QRP's prevalence. This can be conceptualized as a simulation of the experimenter followed by a simulation of the experiment run by this experimenter. Because this was done 300-500 times we could assess the standard deviations of the outcome measures with the set (~meta-analysis) as a unit of analysis. Software was validated by comparing the outcomes of the simulations written in *RealBasic* and the simulations written in *R*.

Finding the optimal fit with the empirical data

Associated with each QRP are a prevalence figure and some free parameters. These are described in the section 'description of QRPs'. These parameters are considered free parameters when a simulation is sought that produces results that fit best to the empirical data. This is an example of an optimization problem where finding the maximum, or minimum, of some function is desired. In this case we wish to find those values of the QRP parameters that result in a simulation of the meta-analysis which most closely resembles the empirical results. To this end we defined a fitness parameter which is small when the simulated and empirical results are close and increases as the results differ. The fitting itself was performed by a Genetic Algorithm (22,23). Genetic Algorithms, although being very computationally intensive, are preferable to hill-climbing algorithms because the latter may get stuck on a suboptimal fit.

The fitness parameter was defined as the sum of the squares of Z-scores that combined the normalized deviations of four metrics comparing the simulated meta-analyses to the actual one. Two of these quantities were derived from the HR by calculating the HR per study and averaging those to a mean HR (mHR) and the HR over all 3,494 trials comprising the 78 experiments (HR). These two HR measures were different and we computed the mean (mHR) and difference (DHR) of them. These two quantities were converted to Z-scores using their means and standard deviations obtained from the simulations. The remaining two Z-scores that went into the fitness were the Z transformed sample-size vs effect-size Spearman's rho correlation (Z_r) and the Z-score (Z_c) derived from the P-value distribution fit (expressed as a Pearson's chi^2 transformed into a Z-score). That is:

$$Fitness = Z_{mHR}^2 + Z_{DHR}^2 + Z_r^2 + Z_c^2$$

where,

$$Z_{mHR} = \frac{\overline{mHR_{Sim}} - mHR_{MA}}{std\left(mHR_{Sim}\right)}$$

$$Z_{DHR} = \frac{\overline{DHR_{Sim}} - DHR_{MA}}{std\left(DHR_{Sim}\right)}$$

$$Z_r = \frac{\overline{r_{Sim}} - r_{MA}}{std\left(r_{Sim}\right)}$$

$$Z_c = \overline{ZOP\left(c^2\left(D_{Sim}, D_{MA}\right)\right)}$$

and overbar denotes taking the mean, std the standard deviation, and χ^2 is the Chi^2 test whose arguments are the distributions of counts D_1 and D_2 and which returns the probability of the resulting c^2 and *ZOP(P)* is a function that returns the Z-score derived from the probability *P*.

It should be remarked that Z_r and Z_c are correlated. This implies that the number of degrees of freedom required to calculate P-values from the fitness value is reduced. We therefore calculated P-values on the basis of Monte-Carlo simulations of the 4 component sum-of-squares where 2 of the components were forced to correlate in the simulation.

The Genetic Algorithm was run using a population size of 60 chromosomes until for 20 consecutive generations no further decrease in the fitness function defined above was obtained. Each chromosome was a binary representation of one possible set of values of the 14 parameters that describe the QRPs being modeled and a 15th parameter for the assumed anomalous HR. The algorithm evaluated the fitness of each chromosome by decoding it into a set of 15 values for these parameters then running 200 Monte-Carlo simulations of the meta-analysis using those values. At the completion of a generation, the next generation of chromosomes is produced by a set

of rules that combine the chromosomes of the current generation favoring those with the best solution, in this case with the lowest fitness values. Further details on how Genetic Algorithms derive one generation from the next are available in the papers referenced (22,23) and are beyond the scope of this paper.

Results

The simulated effect of each QRP in isolation

First we simulated all the applicable QRPs in isolation assuming that every experimenter intended to use the practice whenever possible. Note that this assumed 100% prevalence does not imply that experimenters always do use that particular practice because, for many practices, an extra condition must occur in the data before it can be applied. For example, optional stopping can only be applied if the running P-value becomes smaller than 0.05. Later, we will present results of simulated combinations of QRPs. A review of the results with typical values for the QRP parameters is given in Table 2. We also give the fit of the resulting mean MA simulation with the actual MA data as 4 Z-scores representing the deviations of the 4 key metrics. The value of the over-all fitting variable is a sum of the four Z^2 scores. The fit is perfect when Z=0.

Table 2. Review of applicable QRPs and their parameters

Description	**QRP**	**JLP prevalence**	**Reasonable Prevalence**	**Parm1**	**Parm2**
Confirmation to Pilot	*CtoP*	22.5	20-50	P-crit*	Trialnr**
Pilot to Confirmation	*PtoC*	-	20-50	P-crit	Trialnr
Optional Stopping	*OS*	22.5	20-50	P-crit	StartN
Optional Extension	*OE*	58	40-60	P-crit	extra N
Publication Bias	*PB*	50	40-60	pubprob***	-
Biased removal of Ss	*rmSs*	43	35-55	% of N	-
Deception	*Fraud*	1	~ 1		

* 'P-crit' is the critical P-value above or below that value the QRP is considered by the experimenter. **'Trialnr' is the trial at which the experimenter considers engaging in the QRP. *** pubprob is the publication probability as a function of P-value and rmSs is an acronym for removing subjects post hoc.

As can be seen from the fitting values, each of the QRPs with the exception of optional extension does nudge the data into a direction with a better fit (Table 3).

Table 3. The effect of each of the applicable QRPs in isolation on the simulated variables.

QRP	HR pooled	mHR	Rho	Fit HR	Fit - dHR	Fit - Rho	Fit - pdistr	Overall-Fit
Database	30.99	31.1	-0.11	0	0	0	0	0
None*	25.0 (0.74)	25.01 (0.9)	0.00 (0.12)	-7.8	-0.2	1.16	4.52	83
CtoP** P>.6 @ 10	27.5 (0.7)	28.2 (0.9)	-0.16 (0.11)	-4.2	1.0	-0.4	1.6	21.9
PtoC** P <.4 @ 6	25.7 (0.7)	25.8 (0.9)	-0.01 (0.1)	-7.0	0	0.91	2.78	57.7
OS ** P< .05 @ >15	25.0 (0.8)	25.9 (1.0)	-0.12 (0.11)	-6.5	1.15	-0.11	3.21	54
OE *** P <0.2 N+20	24.9 (0.6)	24.6 (0.9)	+0.07 (0.11)	-8.1	-0.89	2.05	8.06	135
PB* Franco	28.2 (0.8)	28.8 (1.0)	-0.11 (0.11)	-3	0.82	0.02	0.45	9.9
RmSS 5%	26.1 (0.8)	26.3 (0.8)	-0.01 (0.11)	-5.83	0.03	0.89	2.21	39.7

* The second entry gives the theoretical (25% HR, no QRP) fit to the database.
** P >.6 @ 10 means that this QRP is used when the running P-value is larger than 0.6 at trial 10. *** The experiment is extended with 20 trials if P < 0.02 at the end of the planned experiment.

Unlike what is suggested in the literature (24), the simulation of optional extension of the experiment with extra subjects if the cumulated result at the planned N is just above 0.05 does not yield any measurable advantage, neither in the mean HR nor in the overall HR. The respondents in the research of JLP found this practice very defendable (1.79 on a scale from 0 to 2). The simulations show that indeed application of 'Optional Extension' does not turn experiments with just random data into something significant and the practice does no harm under the condition that the extension is limited to a realistic number of subjects. The practice has also a marginal positive effect on the correlation between sample size and HR.

The largest improvement of the fit when using a single ORP is produced by the publication bias. Using only the file drawer produces a fit of 9.9, implying that the simulation results and the empirical results do still differ significantly ($p = 0.05$). Also about 1.6 studies are going into the file drawer for each published study and this figure is larger than the estimate we obtained from the asymmetry in the funnel plot

From Table 3 we can see that none of the QRPs in isolation produces results that fit the observed data. However, given that 5 of the 6 QRPs all cause the simulations to more closely parallel the published data, we investigated if combinations of QRPs could completely explain the results.

The results of combined use of QRPs

In order to see if it is possible to get an acceptable fit of simulations with empirical results when experimenters use several QRPs simultaneously we first ran the GA fitting procedure with no constraints on the prevalence figures. The results show that a reasonable fit of $F = 2.01$ is now reached after 35 generations. However, a few of the QRP prevalence figures converge to very high values. For instance:

applying the publication bias QRP is required to occur for 93% of the experimenters, and 77% of experimenters are required to remove about 5% of the subjects when results do not confirm to the experimenters expectation. These prevalence figures are well above the prevalence figures from the literature and above what we would expect on the basis of our funnel plot analysis.

The realistic prevalence of different QRPs

In order to construct a realistic model of the use of QRPs in GF research, we have to take into account that not every QRP is applied by every researcher. Only the JLP study has measured how often researchers use specific practices. We use their prevalence figures to define an interval of reasonable prevalence figures if we have no other information. However, sometimes there is additional information. For instance, when trying to simulate the effect of fraud we used the prevalence figure from John but we also used information about detected fraud in the field of experimental parapsychology. Those two sources of information provided converging figures giving more confidence in the estimate by John et al. (2012).

Three of the QRPs in isolation do produce negative correlations between sample size and HRs similar to the correlation observed in the database (Table 3). If we combine these QRPs, assuming a prevalence of 100% for each of them, this correlation becomes -0.58, dissimilar from the observed empirical value of -0.10. This suggests that in order to simulate a realistic correlation the prevalences of these 3 QRPs are likely considerably reduced, probably to values around 45% as reported in JLP.

With regard to optional stopping, a glance at the database shows sample sizes ranging from 5 to more than 100. About 45% of the sample sizes seem to be set in advance (N=10, 20, 30 etc. with frequencies around 10) but the other sample sizes are

unusual and appear only 1 or 2 times. No explicit statements in the publications such as ‘using a power-analysis to determine sample size’ are available. Thus we may conclude that the QRP of not specifying sample size in advance has been used. The prevalence of these QRPs has been estimated by John et al. (2012), using a Bayesian Truth Scoring algorithm, to be 58% for collecting more data than planned and 22% for stopping prematurely if the results reach significance. These figures are in line with our estimate of ~40% based upon the publication of a round number of trials.

The practice of abandoning initially unsuccessful studies results in a hitherto not much discussed file drawer. In most file drawer analyses the file drawer contains data of finished studies and not of studies that were prematurely ended. For instance Franco et al (2014) report that 55% of the researchers responded that they favoured starting a new project over writing up null-results. This suggests that the prevalence for this QRP should be set indeed to around 50%.

Fitting using combinations of QRPs and reasonable QRP prevalences

As described in the method section, the reasonable parameter intervals were given to a Genetic Algorithm allowed to search for the optimal fit in the QRP parameter space under the restriction that the value of parameters were not allowed to go outside the reasonable parameter interval.

It can be seen that the final best fit-value of 10.15 due to reasonable application of combinations of QRPs is about the same as in the case of the unreasonable 100% prevalence file drawer effect (Fig.4). Application of the Genetic Algorithm results in the prevalence parameter for publication bias converging to 58%. This effectively

results in a realistic figure of 49.9 % unpublished completed studies. The values of the other prevalences that are obtained through the GA are as follows:

C2P= 49 %, P2C = 47%, OS = 32%, OE = 44% , PB = 58%, RmSS = 41%

Fig 4. The fitting value as a function of generation of the Genetic Algorithm The QRP-parameters are kept within a reasonable interval. Circular points show the mean fitness for each generation, with error bars of 1 SE, diamond points show the best fitness per generation.

The fit of 10.15 indicates that the simulated results still differ significantly from chance ($p < 0.05$). The major reason that the fit is still unsatisfactory is that the simulated HRs are about 2.5% too low. If we had restricted fitting to the HR measures then the difference between the simulations and the empirical HRs would have been very significant ($Z = -2.96$, $p << 0.01$). The fit of the simulated and experimental P-value distribution, however, is very good (corresponding $Z = 0.21$) and also the correlation is reasonably simulated (simulation rho = -0.15 and experimental rho = -0.11, $Z = 0.29$)

The other QRP parameters that came out of this simulation were:

C2P is activated if $p > 0.27$ at trial 10 and P2C is activated if $p < 0.29$ at trial 7. OS starts checking if $p<0.05$ at trial 23 and rmSS yields removal of 4.5% of subjects removed post hoc. Giving an average of 0.7 removed subject per study.

Adding a psi component

All the simulations described so far have used a chance HR of 25%. With this restriction we failed to 'explain' about 2.5 % in the HRs. We can also simulate a true telepathic effect by increasing the probability of a hit over 25%.

We therefore repeated the simulations with telepathic HRs of 26, 27, 28 and 29%. The fit improves with increasing true psi HR (Fig. 5). It can be seen that the effect of adding a true psi component is minimal after a true psi HR of 27%. The fit, F = 1.79, for this case and the corresponding prevalences are all well in the reasonable interval.

Fig 5. Fitness values when using reasonable QRP parameters and allowing for a small true telepathic effect (Psi HR). The right Y-axis indicates the probability that simulations and experimental data are the same.

Discussion

During the last decade there have been intensive debates on the issue of poor replicability in psychology, biology and the medical sciences and there has been much discussion as to what measures would improve replicability.

One of the major issues concerns the effect of file drawers of unpublished studies. Many authors have argued that the mean effect sizes reported in the literature are dissimilar to the true effect size due to biased publication probabilities. However, putting a non significant study in the file drawer is just one ‘Questionable Research Practice’; during the last few years many other QRPs have been shown to be used by a large fraction of the psychology research community and others.

Rather than evaluating how much these other QRPs do in fact contribute to the distribution of results, the focus in the literature so far has been on prevention of the QRPs together with multi-lab replication efforts. For instance there is a growing consensus that preregistration of planned experiments could prevent a number of these QRPs, though not all. The only way to remove all options for QRPs, would be

real time raw data storage in read-only format, with reviewers really checking these raw data against the final publication.

Running multi-lab (preregistered) replication efforts of well established or less well established psychological 'facts' has also been promoted. This focus either on pre-registration or on multi-lab replications might suggest that we'd better forget the results that were obtained before much attention was given to QRPs.

However, rather than discarding the old databases that have been polluted to some extent by QRPs, one might also try to calculate how much these QRPs could possibly contribute to reported effect sizes. This is the approach chosen in this paper. It is important to determine QRP-adjusted effect sizes and therefore to calculate the real power to be used in further research.

Some controversial research findings have been 'explained away' without any quantitative evaluation by skeptics who suggest that these controversial results must be due to QRPs. This is similar to a dirty test-tube argument: that is, declaring a whole experiment to be invalid even if the dirty test-tube can never quantitatively 'explain' all the empirical data.

The simulations of applicable QRPs in the controversial paradigm of GF-telepathy that are presented in this article were intended to answer this quantitative question: How much of the controversial result can be explained under the assumption that QRPs have been used to the same degree that has been reported in the general scientific literature? The GF meta-analytic results are generally seen as the best evidence for an anomaly that proponents call telepathy. These meta-analyses have also been discussed extensively in psychology (25-27).

The first issue to discuss regarding our approach is the selection of studies to be

included in the GF database. We started with the largest database available, which included 108 experiments from 1974 to 2010, but decided to exclude studies before 1985 for reasons detailed in the Introduction.

The second issue concerns the way we chose to account for the QRP of fraud. We argued that the prevalence of fraud assessed in the literature of around 1% did correspond well with the prevalence of fraud (uncovered by psi researcher colleagues, not outside skeptics) in the past within the field of experimental parapsychology. Of the total of 108 studies we removed 2 (2%), both of which were by an experimenter accused publicly of deviating from standard procedures. One could argue that using 2% is a bit conservative and we could instead have removed only the most significant claims by this researcher. It turns out that this would not have made any noticeable difference to the outcome.

The third issue that might be seen as debatable is the choice of the QRP-parameter ranges used in the simulation and fitting procedure. We based our allowed parameter intervals on the prevalence data reported by John et al, but depending on the type of QRP we also took into account other free parameters. For instance, for the QRP of 'optional stopping if $p<0.05$', one has to specify at what trial number an experimenter might start to calculate P-values after each subject.

In those cases where we had to decide upon the allowed parameter range it was the first author's extensive experience with this kind of experiment that allowed us to formulate a reasonable choice. For instance, it can be argued that the practice of optional stopping only starts after a number of trials that at the time was considered to be publishable, say after the 20^{th} subject. It should be mentioned that the results of our simulations are not critically dependent on these choices, with one exception: the fraction of subjects that an experimenter would remove from an experiment in a

biased way. Again the choice here was based upon our experience in the field but some proponents of the psi-hypothesis might consider that assumed fraction too large.

We used the publication probability function reported in a study of psychologists in general. As noted, some have argued that the publication probabilities for studies in the field of experimental parapsychology could be totally different, and far more cautious, from those in general psychology research. For instance, some skeptics have argued that the publication probability function was such that only 1 in 20 studies had been published! As a consequence, about 2000 GF studies would have been required, from which about 1900 ended up in the file drawer. Given the costs in sheer time as well as money of these studies, and the editorial policies in parapsychological journals which also publish non significant studies, this seems to be totally unreasonable. An active search by Blackmore of unpublished GF experiments in 1980 yielded 19 yet unpublished versus, at the time, 31 published studies with no significant difference in outcome measures (28). She therefore concluded: "The bias introduced by selective reporting of ESP GF studies is not a major contributor to the overall proportion of significant results." Finally, a crude interpretation of the funnel plot of this database estimated the number of unpublished studies per published one as much smaller than 1.

Given the assumptions we made about the reasonable intervals for the simulation parameters, we conclude that QRPs are capable of explaining away about 60% of the effect size reported in the GF meta-analysis. Simulations allowing for a true telepathic effect confirmed this estimate and suggested that the true psi effect HR could be ~ 27% (or larger) corresponding to a tiny effect size of ~0.06 (or larger).

Another point of discussion is whether we really did discover and simulate all the QRPs that could be used in this type of experiments. Each of the QRPs used in our

simulations (except those for optional extension and optional stopping) gave some 1% extra HR. This could suggest that, under the assumption that telepathy doesn't exist, there are still 2 or 3 QRPs we failed to investigate. One that we didn't implement is the QRP of manual data-entry. Nowadays the practice of hand scoring has largely disappeared due to the fact that most tests are done using a computer where the subject herself enters the responses or other data-items. Error rates are highly context specific (http://panko.shidler.hawaii.edu/HumanErr/Basic.htm) and range from 0.03% for experienced bank machine operators (29) to 1-2% for students performing a table lookup task (30). The automated GF results, where data-entry was always directly into the computer, have however a slightly *larger* over-all HR (30.77%) than the remaining studies where data-entry was generally less sophisticated (30.25%). Therefore we might conclude that manual data-entry probably didn't introduce noticeable effects on the results.

Our exercises show that QRPs can account for large fractions of small effect size phenomena. What holds for this particular GF-telepathy paradigm most probably also holds for small effect size of less controversial effects, especially in paradigms where the experimenters' freedom is larger. Such is generally the case in most traditional paradigms. For instance, in GF telepathy experiments, unlike in many general psychology experiments, there are no outlier corrections and there is no freedom in preprocessing of physiological data.

As we argued above, this kind of quantitative evaluation of the possible contribution of QRPs to meta-analytic databases allows us to *estimate* the true effect size corrected for QRPs and therefore the required power in further QRP-free experiments.

In the case of GF-telepathy research our simulations suggest a true psi HR of

27% if no further QRPs are found. With this small HR, the required number of subjects to be tested for a probability of 80% to obtain a P-value of 0.05 is in the order of 700 subjects. Never in the history of GF-telepathy experimentation, has this large a number of subjects been used in a single study due to extremely high projected costs.

Splitting up this required number of subjects into smaller numbers for a coordinated parallel replication effort could be shown to be equivalent statistically to a single large study but only if QRPs like 'Confirmation to Pilot' or 'Pilot to Confirmation' can be excluded. These two QRPs relate to a single experiment and having many small experiments gives many options to use them while in a large experiment there is only one option, which considerably reduces the impact of these QRPs on the over-all result.

To obtain a reasonable power, the only realistic option is advance selection of 'gifted' subjects rather than using the average freshman psychology student. Comparing 11 free response telepathy studies using a selected population with 80 similar studies using a non-selected population showed that the effect size for the selected subjects was about three times larger than for the unselected subjects (31). For the GF database, the effect sizes uncorrected for QRPs for selected (artistic) population and the unselected population are ~0.5 and ~0.14 resp. (32,33), If we assume that studies with special subjects do not differ from studies with unselected subjects in terms of use of QRPs the estimated true effect size after correction for QRPs for the artistic population would be around 0.43 and a study with around 50 artists, like the Juilliard students used in the Schlitz-Honorton GF study with selected subjects, would have a power of 80% to establish this effect (34).

A consistent finding in the medical literature is that there are large discrepancies between results of meta-analyses and those of large scale randomized controlled trials

(35). The latter found for instance that meta-analyses would have resulted in the adoption of an ineffective treatment in 32% of the cases and in about the same percentage of the cases an effective treatment was rejected! Hence, a satisfactory power analysis result for each study has been suggested as a required inclusion criterion in meta-analysis (36). Muncer relaxes the power requirements for inclusion of a study in a meta-analysis to 0.50 on the basis of the weighted mean effect size of the initial database. In the case of the GF database only 6 studies would qualify. Interestingly these would produce a mean HR of 31.2% ($p < 10^{-4}$). But of course this result assumes that no QRPs were used in those 6 studies.

The methodology used in our simulation and fitting procedure can also be applied to other meta-analytic databases. For each paradigm, one has to determine first what QRPs are possible. For instance one can easily simulate the effect of trying out different outlier correction procedures and picking the one that will give the 'best' results. Other QRPs might be more difficult to simulate and therefore require careful examination of the original materials. This kind of quantitative evaluation of the possible contribution of QRPs to meta-analytic databases allows us to estimate the true effect size corrected for QRPs and therefore the required power in further QRP-free experiments.

ACKNOWLEDGMENTS

We thank Jim Kennedy for his comments regarding the issue of deception. We thank Peter Bancel for his important remarks on earlier drafts of this paper. We thank Damien Broderick for his comments and revisions on the final draft of this paper.

References

(1) John LK, Loewenstein G, Prelec D. Measuring the prevalence of questionable research practices with incentives for truth telling. Psychol Sci 2012 May 1;23(5):524-532.

(2) Ioannidis JP. Why most published research findings are false. PLoS medicine 2005;2(8):e124.

(3) Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol Sci 2011 Nov;22(11):1359-1366.

(4) Open Science Collaboration. Estimating the reproducibility of Psychological Science. Science 2015;349(6251).

(5) Schouten S. Are we making progress. Psi research methodology: A re-examination 1998:295-322.

(6) Broderick D, Goertzel B editors. Evidence for Psi: Thirteen Empirical Research Reports. Jefferson, NC: McFarland & Company; 2014.

(7) Williams BJ. Revisiting the ganzfeld ESP debate: a basic review and assessment. Journal of Scientific Exploration 2011;25(4):639-661.

(8) Hyman R. The Ganzfeld Psi Experiment: a Critical Appraisal. Journal of Parapsychology 1985;49(1):3-49.

(9) Honorton C. Meta-Analyses of Psi Ganzfeld Research: A response to Hyman. Journal of Parapsychology 1985;49(1):52-86.

(10) Hyman R, Honorton C. A joint communiqué: The psi ganzfeld controversy. Journal of Parapsychology 1986;50(4):351-364.

(11) Storm L, Tressoldi PE, Di Risio L. Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. Psychol Bull 2010;136(4):471.

(12) Egger M, Davey Smith G, Schneider M, Minder C. Bias in meta-analysis detected by a simple, graphical test. BMJ 1997 Sep 13;315(7109):629-634.

(13) Slavin R, Smith D. The relationship between sample sizes and effect sizes in systematic reviews in education. Educational Evaluation and Policy Analysis 2009;31(4):500-506.

(14) Kennedy JE. Experimenter misconduct in Parapsychology: Anaslysis, manipulation and fraud. 2013; Available at: jeksite.org/psi/misconduct.pdf.

(15) Armitage P, McPherson C, Rowe B. Repeated significance tests on accumulating data. Journal of the Royal Statistical Society.Series A (General) 1969:235-244.

(16) Feller W. Statistical aspects of ESP. Journal of Parapsychology 1940;4:271-298.

(17) Lakens D. What p-hacking really looks like: A comment on Masicampo and LaLande (2012). Q J Exp Psychol 2015;68(4):829-832.

(18) Radin DI, Nelson RD. Evidence for consciousness-related anomalies in random physical systems. Foundations of Physics 1989;19(12):1499-1514.

(19) Bösch H, Steinkamp F, Boller E. Examining psychokinesis: The interaction of human intention with random number generators--A meta-analysis. Psychol Bull 2006;132(4):497.

(20) Watt C. Research Assistants or Budding Scientists. Journal of Parapsychology 2006;70:355-356.

(21) Team RC. R: A language and environment for statistical computing.R Foundation for Statistical Computing, Vienna, Austria, 2012 2012.

(22) Mebane Jr WR, Sekhon JS. R version of GENetic Optimization Using Derivatives. 2013.

(23) Goldberg DE. E. 1989. Genetic Algorithms in Search, Optimization, and Machine Learning. Reading: Addison-Wesley 1990.

(24) M. Bakker. Good science, Bad science: Questioning Research Practices in Psychological ResearchUniversity of Amsterdam; 2014.

(25) Bem DJ, Honorton C. Does psi exist? Replicable evidence for an anomalous process of information transfer. Psychol Bull 1994;115(1):4-18.

(26) Bem DJ, Palmer J, Broughton RS. Updating the ganzfeld database: A victim of its own success? Journal of Parapsychology 2001;65(3):207-218.

(27) Milton J, Wiseman R. Does Psi exist? Lack of replication of an anomalous process of information transfer. Psychol Bull 1999;125(4):387.

(28) Blackmore S. The extent of selective reporting of ESP Ganzfeld studies. European Journal of Parapsychology 1980;3:213-220.

(29) Klemmer E, Lockhead G. Productivity and errors in two keying tasks: A field study. J Appl Psychol 1962;46(6):401.

(30) Melchers R, Harrington M. Human error in simple design tasks. ; 1982.

(31) Baptista J, Derakhshani M. BEYOND THE COIN TOSS: EXAMINING WISEMAN'S CRITICISMS OF PARAPSYCHOLOGY. Journal of Parapsychology 2014;78(1):56-79.
(32) Exploring the links: Creativity and psi in the ganzfeld. Proceedings of Presented Papers: The parapsychological Association 40th Annual Convention; 1997.
(33) Schlitz M, Honorton C. Ganzfeld Psi Performance Within an Artistically Gifted Population. Journal of the American Society of Psychical Research 1992;86(2):83-98.
(34) Faul F, Erdfelder E, Lang A, Buchner A. G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behavior research methods 2007;39(2):175-191.
(35) LeLorier J, Gregoire G, Benhaddad A, LaPierre J, Derderian F. Discrepancies between meta-analyses and subsequent large scale randomized controlled trials. New England Journal of Medicine 1997;337:536-618.
(36) Muncer S, Taylor S, Craigie M. Power dressing and meta-analysis: incorporating power analysis into meta-analysis. Journal of Advanced Nursing 2002;38(3):274-280.

Figure1
Click here to download Figure: fig1.eps

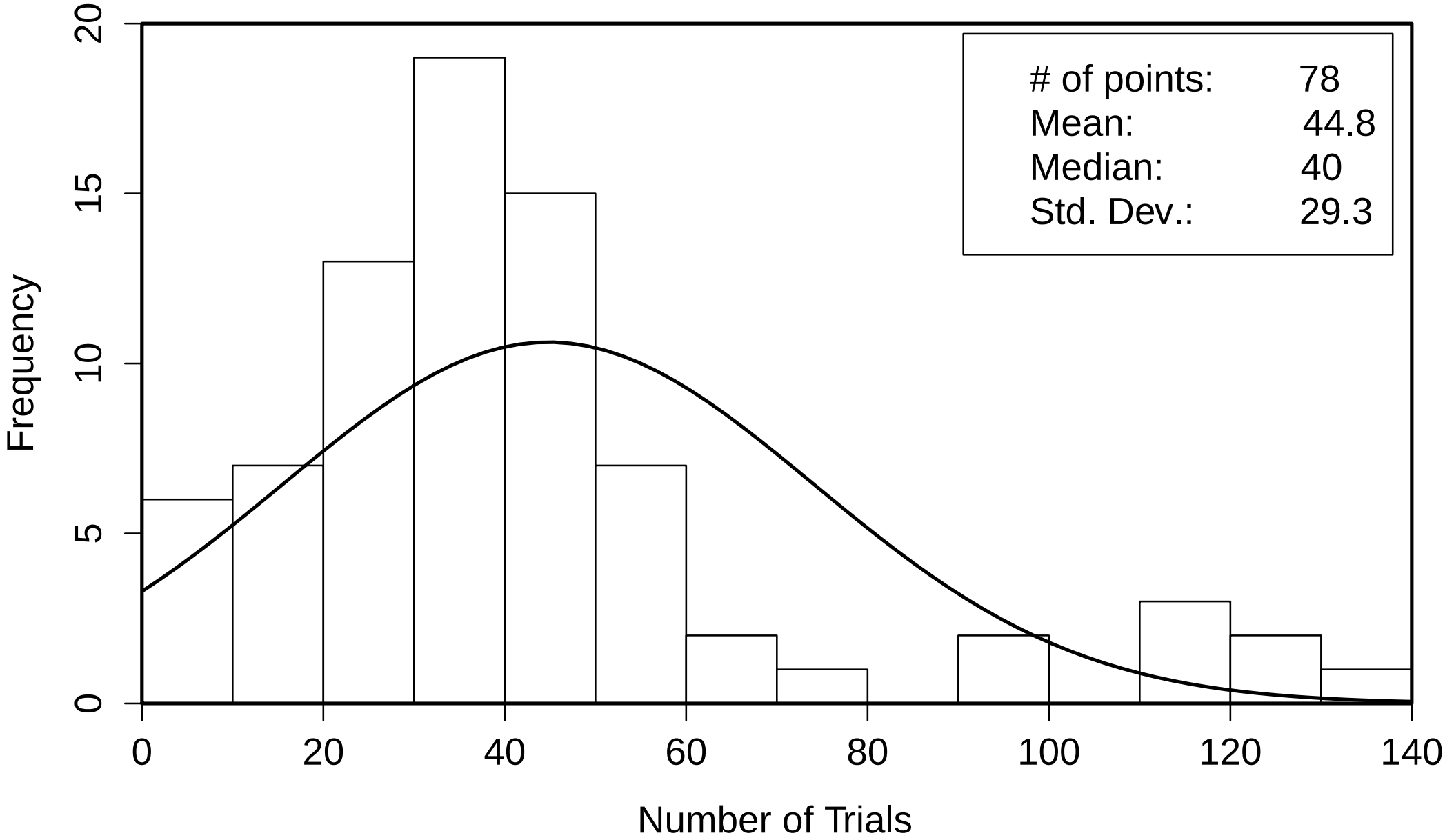


Figure2
Click here to download Figure: fig2.eps

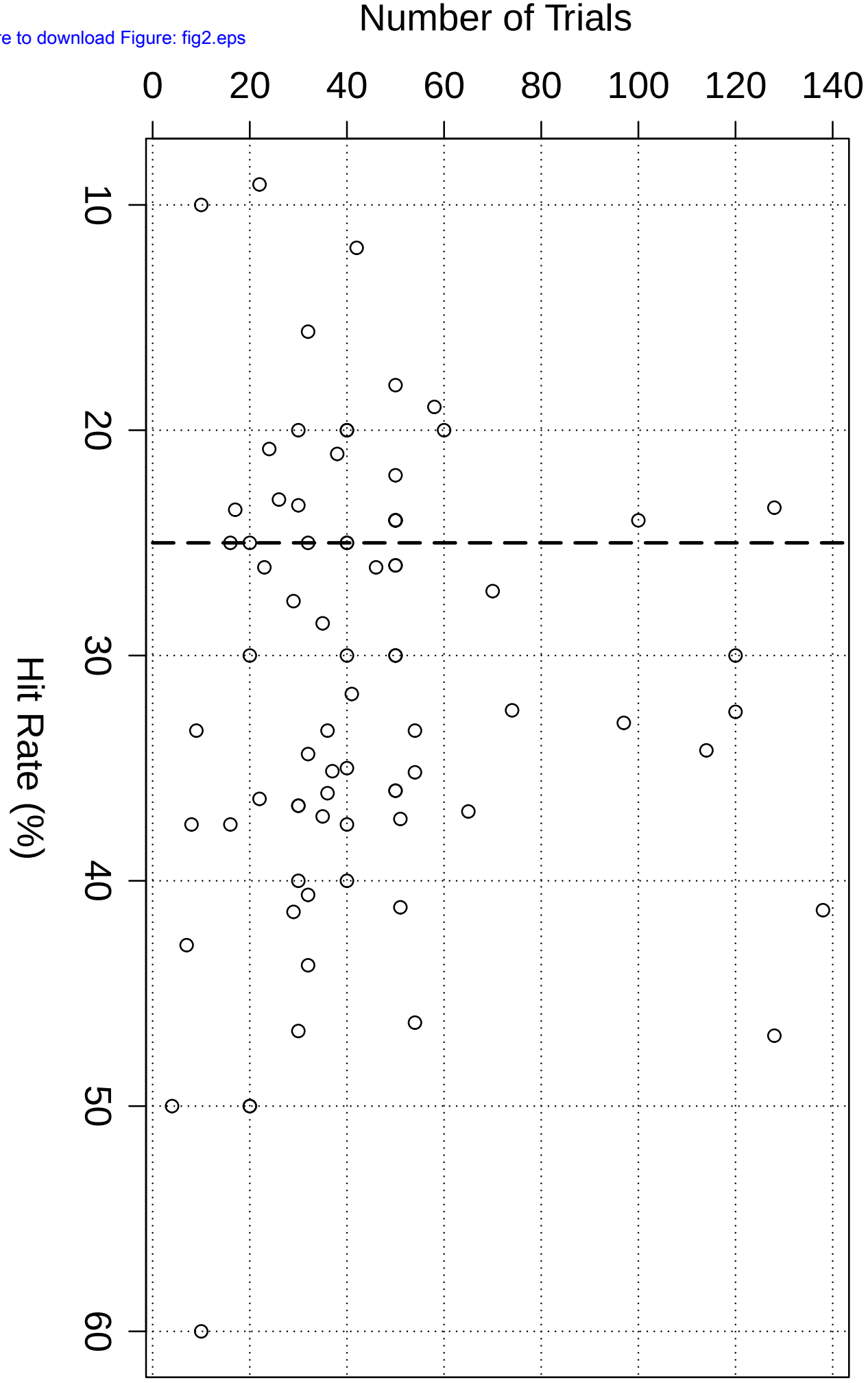


Figure3
Click here to download Figure: fig3.eps

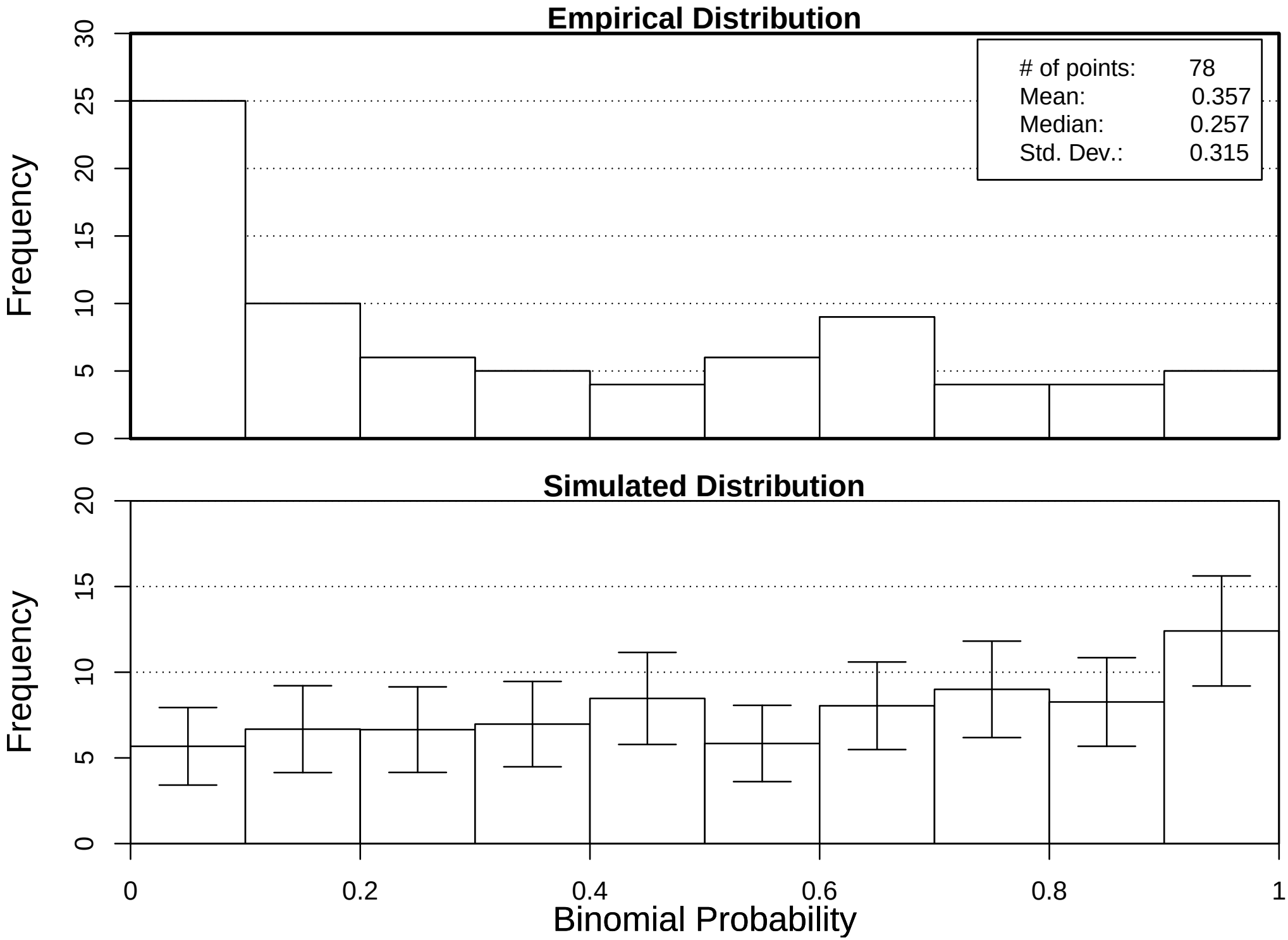


Figure4
Click here to download Figure: fig4.eps

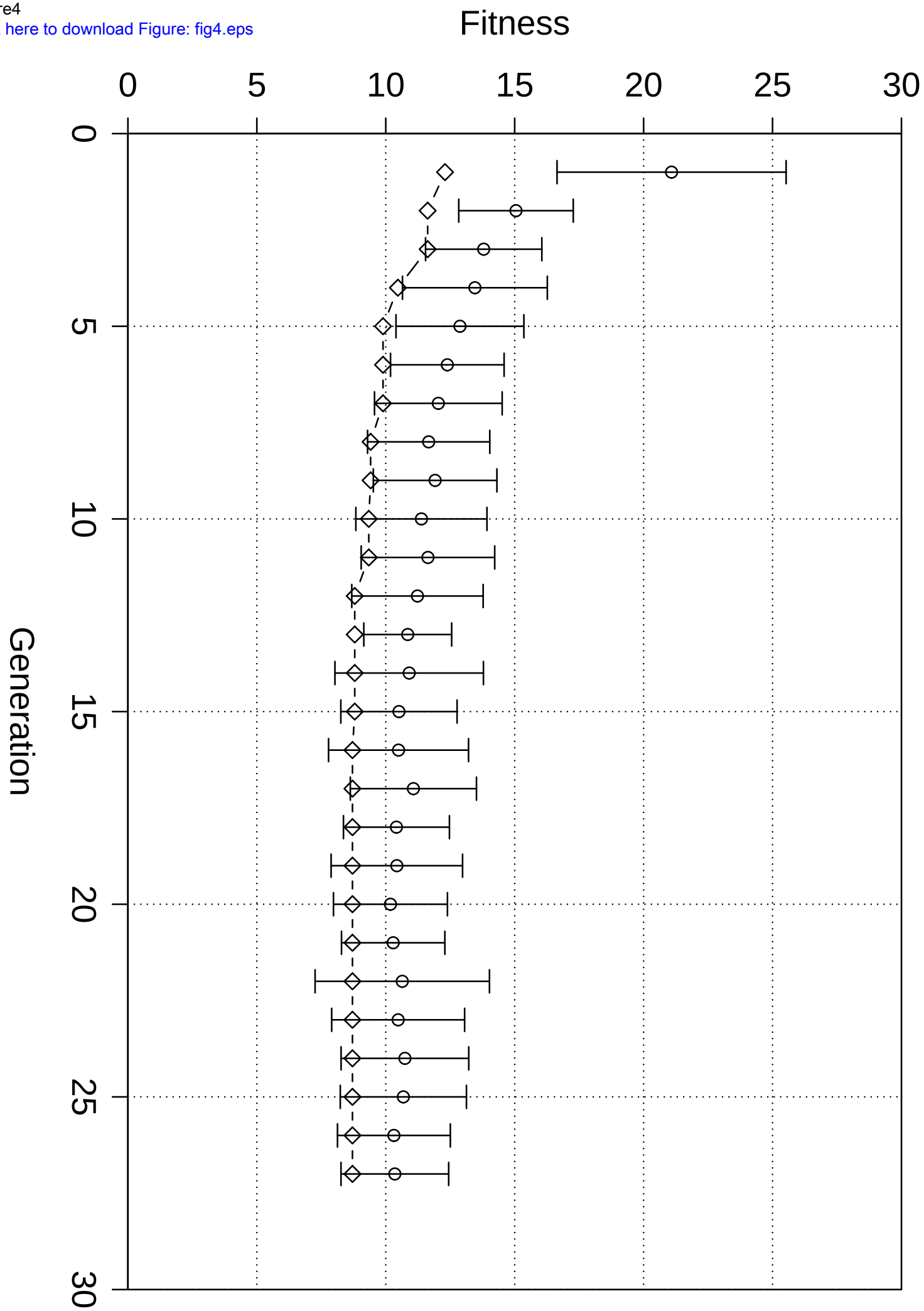
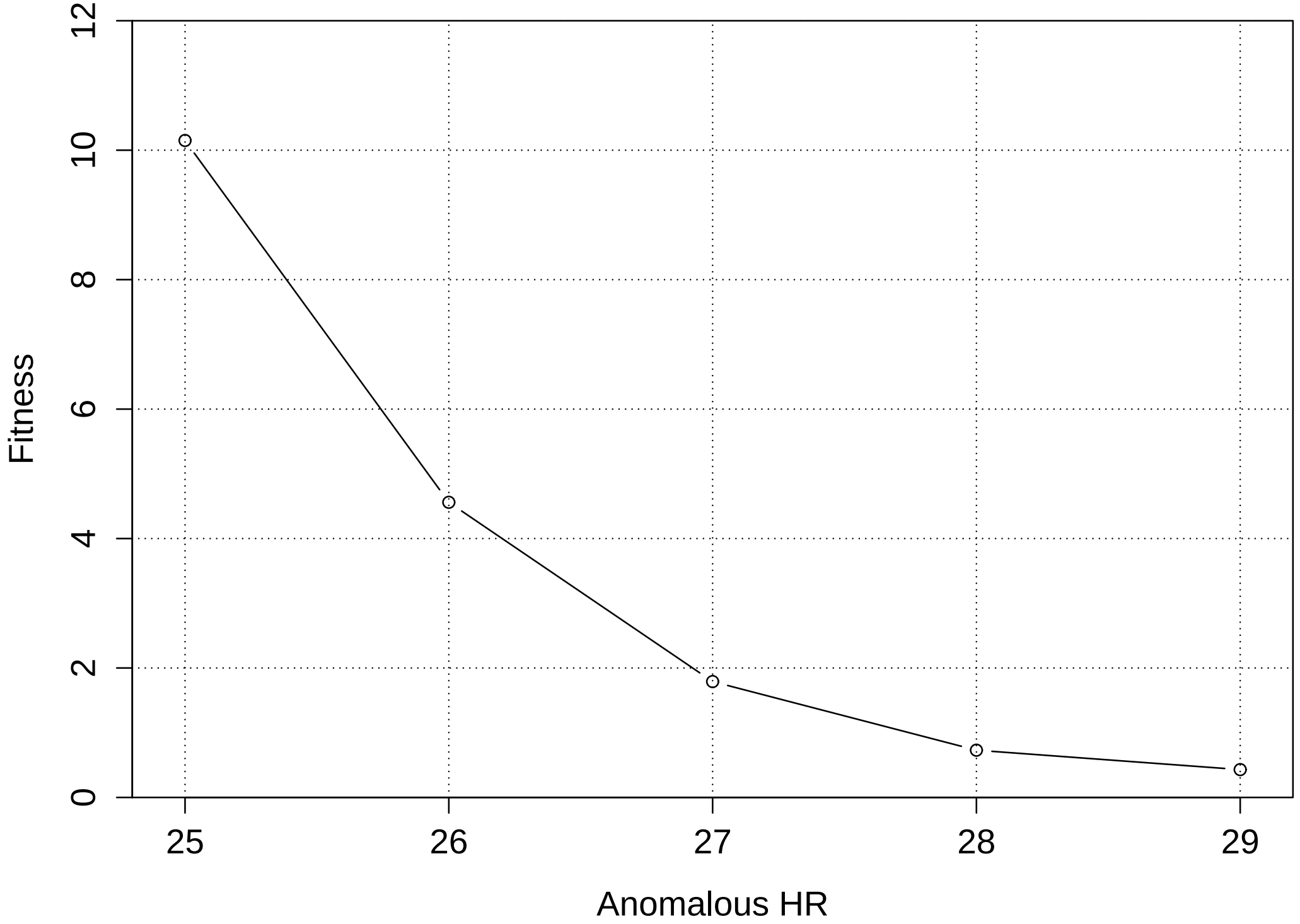


Figure5
Click here to download Figure: fig5.eps



Supporting Information

Click here to download Supporting Information: GF_MA_for_QRP_analysis copy.xlsx

Supporting Information

Click here to download Supporting Information: simu_qrp.rbp