P. A. Bancel

Bursary Grant 308/14: Final Report

A Study of Heterogeneity in Parapsychological Databases

Background to this Report

Achieving reproducibility in parapsychological experiments remains a significant challenge. Although reproducibility issues have been widely discussed, quantitative assessments of reproducibility for different psi effects is hard to come by. One consequence is the difficulty to convey succinctly the evidence for psi effects, particularly as regards statistical evidence from meta-analyses. Another is that it's hard to make reliable effect size estimates, which hampers the designing of new experiments, as well as replications. The reasons for this quandary include poor reproducibility, and the related issue of heterogeneity in effect sizes. Heterogeneity may itself provide evidence for an effect when it exceeds statistical noise, or be indicative of uncontrolled experimental or methodological factors. Sources include the distribution of psi ability in test populations; differences in experimental procedures; or variable experimenter psi, to name a few possibilities. As is well-know, publication bias and other methodological problems also can give rise to excess heterogeneity, in much the same way that they can generate spurious effect sizes. In meta-analyses, evidence for an effect rests not only on the significance of its estimated size, but also on the degree to which spurious methodological contributions to the effect size and heterogeneity can be controlled for. Thus, for a meta-analytic rejection of the Null hypothesis to be compelling, it must consider the Null in conjunction with a wide array of potential methodological deficiencies. These days, it is common to refer to these collectively as "Questionable research practices", or Qrps.

The original goal of the Project was to assess the heterogeneity in four databases from published meta-analyses, and for the database of the Global Consciousness Project (GCP). The meta-analyses are of published results for: the Ganzfeld telepathy protocol; micro-PK effects on random number generators; presentiment studies; and replications of Daryl Bem's precognition protocols. Several statistical tools were to be used (funnel plot analysis, trim-and-fill and zero-mean filedrawer estimates, the recent P-curve test and its adaptation to a filedrawer estimate). However, as the Project progressed, it became evident that a more complete treatment of all methodological issues was necessary if any real progress was to be made. In late 2015, Dick Bierman and colleagues published a Monte Carlo study of Qrps for the Ganzfeld database. A decision was made to follow this approach, and considerable development of the method was done for the Project. The Monte Carlo method was extended to include heterogeneity. A number of important improvements in power and simulation speed were implemented. The Bial Scientific Board approved these changes to the original Project goals on February 19, 2018. The Board also approved limiting of analyses to the Ganzfeld, micro-PK and GCP data, given the extent of development time and effort.

Introduction

In many scientific disciplines, inferential evidence for effects often relies on the statistical results of meta-analyses. In recent years, several meta-analyses of psi protocols have appeared in mainstream journals (Bösch, Steinkamp, & Boller, 2006; Storm, Tressoldi, & Di Risio, 2010; Mossbridge, Tressoldi, & Utts, 2011; Bem, Tressoldi, Rabeyron & Duggan, 2015). The analyses are among the strongest arguments for psi *vis-à-vis* the broader scientific community since they use familiar methods and employ best practices. This has allowed not only for publication in journals that reach an audience beyond that of, say, *The Journal of Parapsychology*, but also for published debate on the merits of each case. Although critiques have been vigorous, the open exchanges are welcome and necessary for disseminating what we know about psi and engaging colleagues in other fields.

A second feature of meta-analysis that parapsychology shares with other research fields is a heightened awareness of its weaknesses. It has long been known that publication bias can compromise meta-analytic conclusions and that remedies such as file-drawer estimates and trim-and-fill procedures are inadequate fixes. This also follows for approaches that try to model selection effects in the publication process (Jin, Zhou & He, 2014). Regression techniques such as PET-PEESE (Stanley & Doucouliagos, 2012) can test for a real effect beyond the presence of bias, but are subject to limitations of power and regression assumptions.

The elephant in the room, however, is that other methodological problems, collectively known as questionable research practices (Qrps), also adversely impact meta-analyses. Examples are reporting unplanned analyses, the rounding of P-values and extending data collection in an effort to achieve significance. Recent work has sought to identify Qrps and determine their usage among scientists via survey studies (John, Loewenstein & Prelec, 2012; hereafter, JLP). JLP identify 10 Qrps and find that researchers self-report engaging, at least once, in most of these Qrps with frequencies ranging roughly from 25% to 65%. Other surveys (Fiedler & Schwarz, 2016) emphasize that the real frequencies of Qrps may be considerably lower. The bottom line, however, is that Qrps are widespread, probably depend in unknown ways on discipline and culture, and are difficult to estimate with any reasonable accuracy. Given this state of affairs, it is unlikely that the multiple issues affecting meta-analytic conclusions will be resolved by a single statistical test, no matter how inventive. The recent introduction of p-curve analysis (the p-curve is the distribution of significant P-values in a meta-analytic database) is an attempt in this direction (Simonsohn, Nelson, & Simmons, 2014), but its utility, which assumes a restricted set of Qrps, is diminished if a broader Qrp set is considered (Bruns & Ioannidis, 2016). These considerations highlight the urgency, felt today across disciplines, to adhere to norms of pre-registered, well-powered studies when making inferences about real effects. Effects previously established from meta-analysis face the burden of replication, but this is in turn complicated by the difficulty of reliably estimating effect sizes, and hence determining an appropriate replication power.

One possible way out of the quandary is to go all in and make a brute force estimate of the effect of Qrps on meta-analytic databases via Monte Carlo simulation. The goal is to determine whether any combination of Qrps, at free-ranging levels of frequency, can account for the distribution of effect sizes in the data. In essence, this amounts to augmenting the Null hypothesis to a continuum of Nulls, each with its own Qrps, and seeing if the ensemble of Nulls can be rejected by the data. If so, the evidence for a real effect is strengthened, at least to the extent that the modeled Qrps are representative of the variety of Qrps that researchers actually engage in. The procedure can also simulate a real effects in the presence of Qrps thereby yielding not only evidence for an effect, but also a lower bound on its size. This provides conservative estimates for well-powered replications.

The first attempt at a full Qrp simulation was reported recently, and applied to the Ganzfeld database (Bierman, Spottiswoode & Bijl, 2016; hereafter BSB). The database comprised 79 studies from the period 1985 to 2010. BSB found that Null Qrp models could indeed provide a good fit to the data, but only with unreasonably high prevalences of Qrp usage. They suggest that this strengthens the evidence for psi. Models that included real effects and with more reasonable Qrp prevalences (that is, within the ranges reported in JLP) yielded good fits with hit rates of 27%. This is considerably lower than the raw and meta-analytic hit rates of ~31% (the Null hit rate being 25% for the Ganzfeld protocol). Thus, Qrps may inflate the Ganzfeld effect size by as much as a factor of three. Consequently, replications based on this conservative estimate of hit rates would need nearly ten times more data, for an equivalent power, than a replication based on the hit rate from meta-analysis. Two aspects of BSB's work are worth emphasizing: 1) Moving forward, it will be important for parapsychology to have an answer to the question: How sure can we be that Qrps cannot explain effects? Providing that answer can clarify issues of inference and replicability, and buttress evidence for psi; it also demonstrates that parapsychology remains at the cutting edge of methodology and analysis; 2) Future replications will need input from Qrp analyses to avoid under-powered designs; 'gold standard' replications that are inadvertently under-powered can have serious negative impacts on progress (a case in point is the much noted Consortium replication of the PEAR RNG-PK database (Jahn et al., 2000); widely cited as a replication failure and indicative of elusive psi, the study was in fact simply under-powered and hence inconclusive (Varvoglis & Bancel, 2014)).

Qrp simulation can thus provide informative updates to the status of parapsychological databases. This report builds on BSB's approach by addressing several drawbacks and limitations. First, inferences that depend on establishing whether Qrp prevalences are 'reasonable' or not must be tempered by the acknowledgement that

these assessments are subjective. Of course, this is a problem for many statistical inferences (consider the 0.05 P-value significance criterion), but given that Qrps are multiple and that prevalences depend on many human factors, compelling inferences will want to be robust against the highest (i.e., most conservative) prevalences possible. The remedy is to increase the simulation's power. This can be achieved by straightforward adjustments to BSB's approach. Second, the Monte Carlo simulations are CPU intensive and this can hamper the modeling (the BSB simulation took upwards of 30 hours, running 6 processors in parallel). Generally, and for replication design in particular, the need to test a variety of Qrp codings may conflict with the availability of computer resources. To address this, the Monte Carlo strategy is reformulated, allowing a significant gain in simulation speed. Finally, extending the Qrp simulations to other distribution parameters, such as effect size heterogeneity, can in certain cases resolve ambiguous evidence for effects. An example is given in the Discussion.

Methods

Basics of Qrp analysis

In the simplest terms, the problem for meta-analysis is that Qrps can shift the distribution of measured effects to produce a spurious effect size estimate. However, each Qrp has a unique impact on the distribution of effects, beyond the mean shifts, and these signatures can in principle be distinguished from a real effect and from other Qrps. The strategy, then, is to devise models that have a good power for distinguishing real effects from Qrps. The models include all combinations of Qrps, and allow their prevalences to vary freely, or over some restricted range. The models' goodness-of-fits to the data (usually called the Fitness, and perhaps expressed as P-values) then quantify how well Qrps do, or do not, account for the data.

Qrps can be roughly grouped into 3 types: data-hacking, p-hacking and publication bias. Data-hacking Qrps make a free choice, that is, one not stipulated by protocol or algorithm, to accept or reject a small subset of data that has been previously examined. Examples are removing an unsuccessful trial that "didn't go well", or restarting an experiment with a protocol modification after a few trials initially give poor results. P-hacking involves decisions made at or toward the end of an experiment that try to lower the final reported P-value. Examples are extending an experiment that comes close to $P = 0.05$ in an attempt to attain significance, or performing multiple analyses and reporting the lowest (or a low subset). Publication bias includes any actions, by any actors, before or during the publication process, for which the study P-value influences the probability of publication. In Qrp analysis, first a listing of Qrps appropriate for a meta-analysis is decided on, and then the Qrps are implemented by a coding scheme for Monte Carlo simulation.

The Fitness function combines separate tests that are sensitive to Qrps. Tests include portions of the p-curve most affected by Qrps, the overall effect size and other parameters such as small study effects (correlations of effect size with study N), or the heterogeneity. The tests can be combined using standard methods, such as the Stouffer Z or Fisher's chi, to render an overall Fitness value.

Qrp simulations then search, via repeated Monte Carlo simulation, for the Qrp model that produces the best fit or fits to the data. Qrps may be rejected as an explanation of the data if best fits have a low probability (in a frequentist picture), or retained as a viable hypothesis if fits fail to exceed a designated significance value.

Simulated Qrps

Eight Qrps are used in the simulations, and a ninth is treated separately. The simulated Qrps are designed to return the exact number of trials reported for each study of the meta-analysis under review. Three pairs of Qrps are simulated exclusively (not together). These are Stopping and Extension, a pair of Multiple analysis Qrps and two modes of publication bias. Simulations thus encompass 108 possible Qrp combinations. The modeling of simulated Qrps is sketched below.

Optional re-start. An optional re-start occurs when the distinction between a pilot and a confirmatory study is not respected. This is a data-hacking Qrp that allows to reject early data if it does not tend toward hypothesis confirmation. The Qrp was modeled by restarting the simulation if the P-value of the first 10 trials was at or

above 0.6, and subsequently accepting the new trials. By design, the Qrp is activated 40% of the time when the Qrp prevalence is at 100%.

Trial replacement. During a study, trials may be dropped for reasons not clearly specified by the protocol (such as a subject reporting extreme discomfort). If this is done with knowledge of the trial outcomes it is a data-hack. Trial removal was modeled by selecting trials with negative outcomes and accepting the outcome of replacement trials. The replacement was done for a total of 5% of all trials.

Optional stopping. A study may be halted near its end if the overall P-value falls below the nominal 0.05 significance level. This p-hacking Qrp was modeled by examining the last 15% of trials and stopping the first time the study P-value fell below a target value. The target P-value was allowed to vary randomly between 0.05 and 0.06. In order to produce an output with the actual reported study size of the meta-analysis, the model returns a vector of trials with the reported N of trials, but with re-scaled values to yield the P-value obtained by stopping.

Optional extension. Similarly to stopping, a study may be extended in an effort to attain significance. The extension Qrp continues data collection of up to 15% more trials if the P-value first terminated below 0.15. The extension stops at the first instance of attaining a target P-value (determined as for optional stopping). The simulation returns a scaled vector of trials with the original N that has the P-value from extension, whether or not the target is reached.

Weak multiple analysis. This Qrp simulates multiple analysis of strongly correlated tests (such as performing Spearman and Kendall rank correlations), and reporting the more significant outcome. The Qrp was simulated by comparing the study P-value with P-values calculated with 5% of trials truncated at the start and then at the end of the trial vector, and retaining the lowest. A switch caused more aggressive p-hacking when the initial P-value was below 0.20. This was done by increasing the truncation to 10% of trials.

Strong multiple analysis. Reporting analyses of unplanned data splits or other weakly correlated moderators is an example strong p-hacking. This was implemented in a similar fashion to the weak multiple analysis, but with truncations of 17% and 33%. Simulations used either weak or strong multiple analysis and did not run both in parallel.

Publication bias. Two different acceptance functions for publication bias were simulated. Both increase monotonically towards 1 with decreasing P-value. The first (used by BSB) is a smooth function that increases sigmoidally between P-values of 0.30 and 0, from a base value of ~0.50:

$$0.65 + 0.40 \, \mathrm{Tanh}[\, 2\text{-}10 * \mathrm{pval} \,]$$

The bias function places roughly 50% of studies with P-values greater than 0.3 in the file drawer and accepts studies with lower P-values with increasingly higher probability. On average, 59% of studies are published so that the ratio of file drawered to published studies is less than one. The bias function is intentionally moderate; publication bias is among the most impactful Qrps and too strong of a bias is quickly rejected in simulations.

A step-wise bias function was also simulated, with acceptance levels of .30, .60, .80 and 1 that had step points at P-values of .50, .10 and .05. The stepped bias function impacts p-curves more strongly and the more conservative smooth bias was used as a default.

N-dependent publication bias. This Qrp allows for stronger bias for small studies. The bias base level (for insignificant studies) was set at 60% acceptance for N > Mean(N), and decreased proportionally towards zero for smaller N.

Fraud. It is difficult to model fraud in a simulation. Following BSB, fraud was handled by removing a small percentage of significant studies from the experimental data. Fraud is thus treated in an inverse manner: rather than adding it to the simulation, a prophylactic measure is applied to the experimental data. From JLP and BSB, fraud percentages were set at 4% and 6%.

P-curves for simulated Qrps

The Qrp analysis uses the full range of the P-value distribution, as opposed to just that of significant P-values (as is done for PET-PEESE). Figure 1 shows p-curves for the Null and a real effect in the absence of Qrps. Note the exponential increase in density at low P-values (blue trace), which signals the increase of significant studies

when an effect is present. The p-curve gives the distribution of study P-values of a meta-analysis, but cannot provide a suitable estimation of the effect size because it does not contain information of the study sizes.

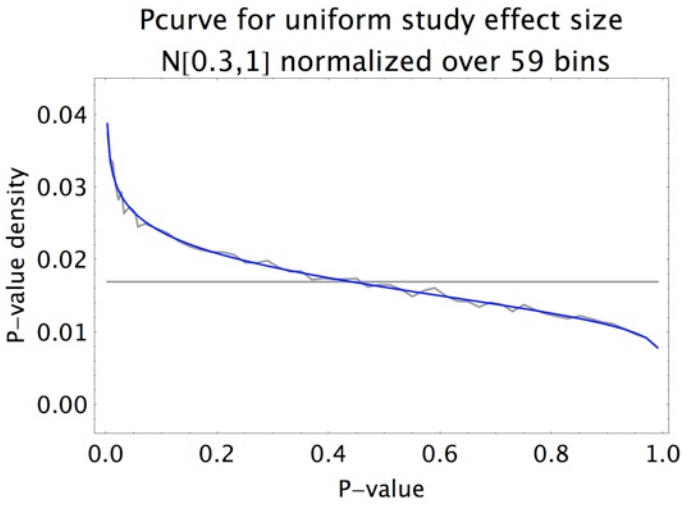

Figure 1. The p-curve for a real effect with effect size of 0.30 over the full P-value interval using normal statistics. The blue trace is an exact calculation and the gray line tracking it is from a Monte Carlo simulation of modest size. The horizontal line shows the flat distribution of the Null p-curve. The vertical scale is normalized to give a density that sums to one.

How Qrps impact the p-curve (with prevalences at 100%) is shown in Figure 2. Several observations are noteworthy. First, the p-curves vary considerably, with structure that follows from the nature of the Qrp. The data-hacking Qrps (Restart and Replacement) closely resemble real effects. This is expected since these Qrps don't depend on P-values, and merely replace a subset of trials in a biased way. The Stopping and Extension p-hacks have the strongest signatures since they narrowly target the p-curve around 0.05. Note that they both show a dip as they promote P-values above 0.05 into a sharp spike, and then decrease quickly and flatten out. These Qrps are not efficient at producing P-values much lower than about half the target P-value. The Multiple analysis Qrps increase the p-curve at low P-value, somewhat in the manner of a real effect, but they also show a promotion dip as hacking becomes more aggressive close to 0.20. The Publication bias Qrps reproduce the bias function. The Publication bias p-curves are distinguished from real effects by the flattening seen in the right-hand half of the curve, and also at low P-values. The two Publication bias curves are the same because the p-curve does not carry information on the study sizes.

Note that large regions of the respective p-curves carry relatively little information about the Qrp. This can lead to a loss of power if the entire p-curve is included in the Fitness function. Selecting p-curve regions that have large dispersion under the influence of Qrp combinations is a reasonable tactic for optimizing the Fitness power.

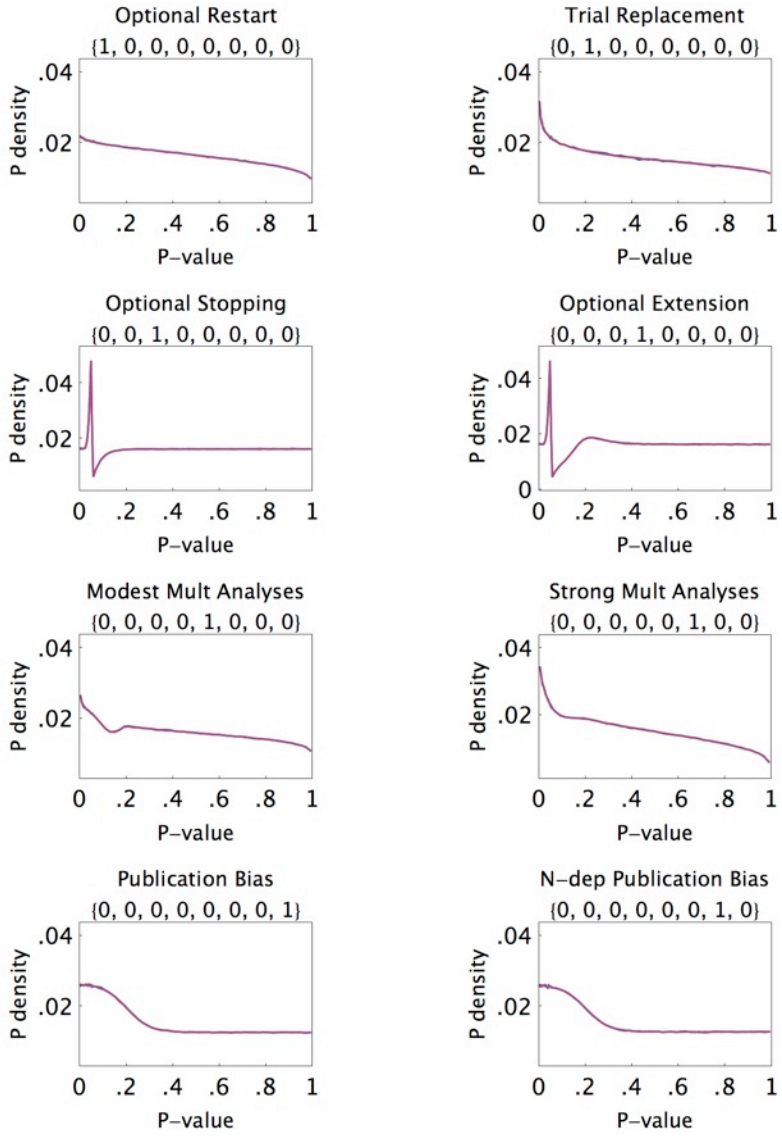

Figure 2. P-curves for the eight modeled Qrps. Prevalence levels are 100%. The curves are relative to a homogeneous Null p-curve aligned at a density of 0.017.

In combination, Qrps interact to produce more complex p-curves. For instance, stopping will be triggered more frequently, and hence have a greater impact, if it is preceded by a data-hack that increases P-values near the stopping target. Examples of p-curves for Qrp combinations are shown in Figure 3. The Qrps represented are identified by the binary vectors on the plots (see Figure 2 for the codings).

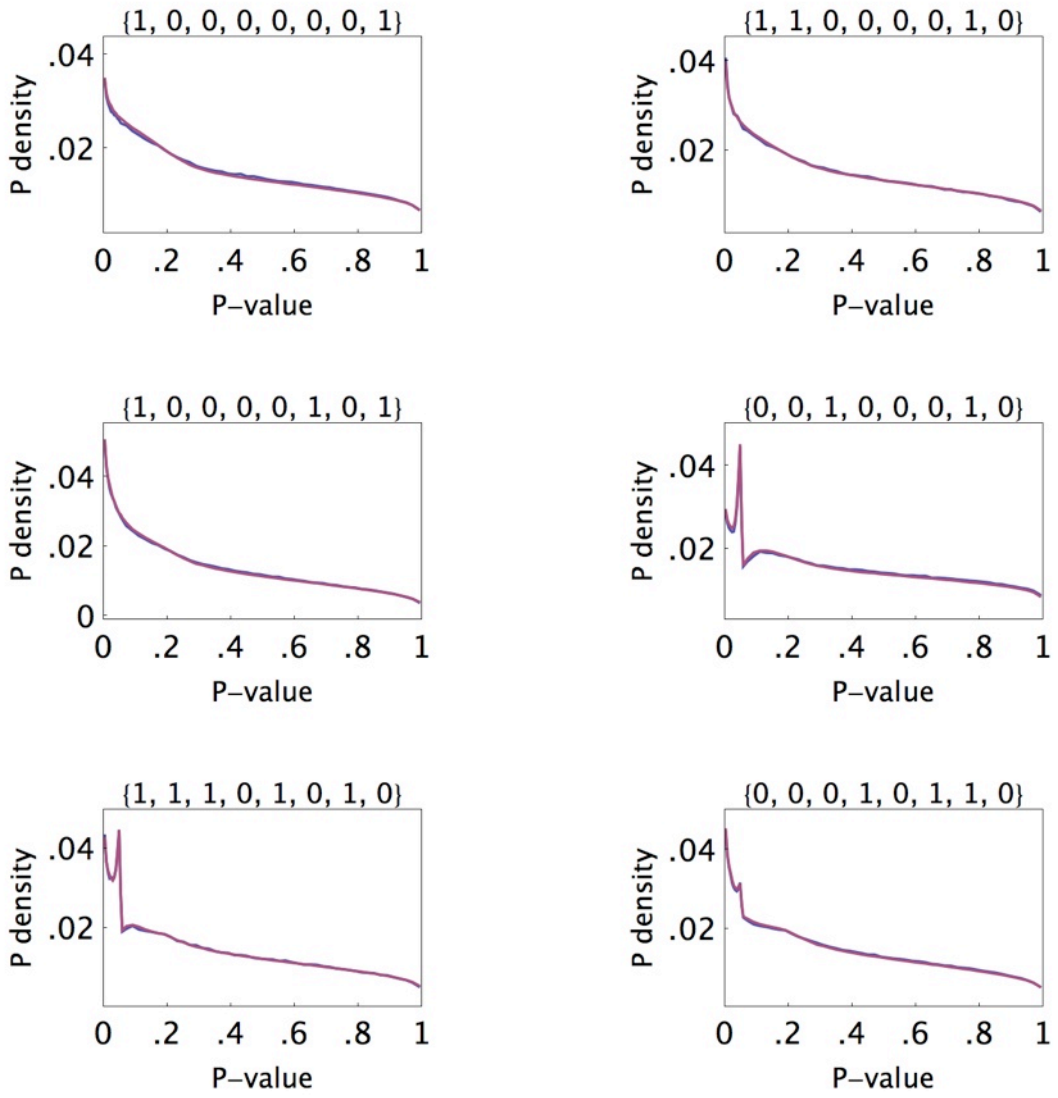

Figure 3. Examples of p-curves subject to multiple Qrps. Here, the Qrps operate at 100% prevalence. The binary vectors above each plot code the operative Qrps, and refer to the basis vectors in the plots of Figure 2. Plots overlay 2 traces: P-curves from direct Monte Carlo simulations (blue) and those using a linear basis method (red).

Qrp effects on distribution parameters

As mentioned above, Qrps affect other parameters of the study effect size distribution, and these can be simulated and included in the Fitness function. Four parameters of interest are: the fixed-effects model effect size (FEM); the Spearman rank correlation of effect size and study N (indicative of small study effects); the heterogeneity; and the skewness. Clearly, the FEM effect size is the most salient since its estimation is the main focus of meta-analysis. But the others are impacted by Qrps and thus may provide additional power for the Fitness function. Generally, a decision is made before simulation about which parameters to include, depending on the Qrps to be modeled and the experimental parameter values.

Figure 4 displays the simulated parameters for combinations of Qrp effects from a meta-analysis of 50 studies with study sizes ranging from 25 to 100 trials. Error bars are 96% confidence intervals (CI) and the models are all 108 combinations of Qrps at 100% prevalence. In the Figure, the Null Qrp is in red and the 8 individual Qrps in blue. Gray points are models that combine two or more Qrps (models are sorted in increasing order for clarity, but the sorting is different for each plot).

Several observations are in order. First, a large dispersion of values suggests an advantage for the Fitness since this allows for a better discrimination of models. Including the correlation, heterogeneity and certainly the skewness may, in fact, lower the Fitness power due to their low dispersion. Second, the simulations provide approximate bounds to values attainable with Qrps. Experimental values that exceed the bounds thus may indicate a real effect, or perhaps a missing Qrp. For all parameters, optional stopping and extension have negligible impacts. This is counter intuitive, but it was also noted by BSB. Publication biases are consistently among the most important contributors. However, all Qrps other than stopping and extension contribute substantially to the effect size dispersion.

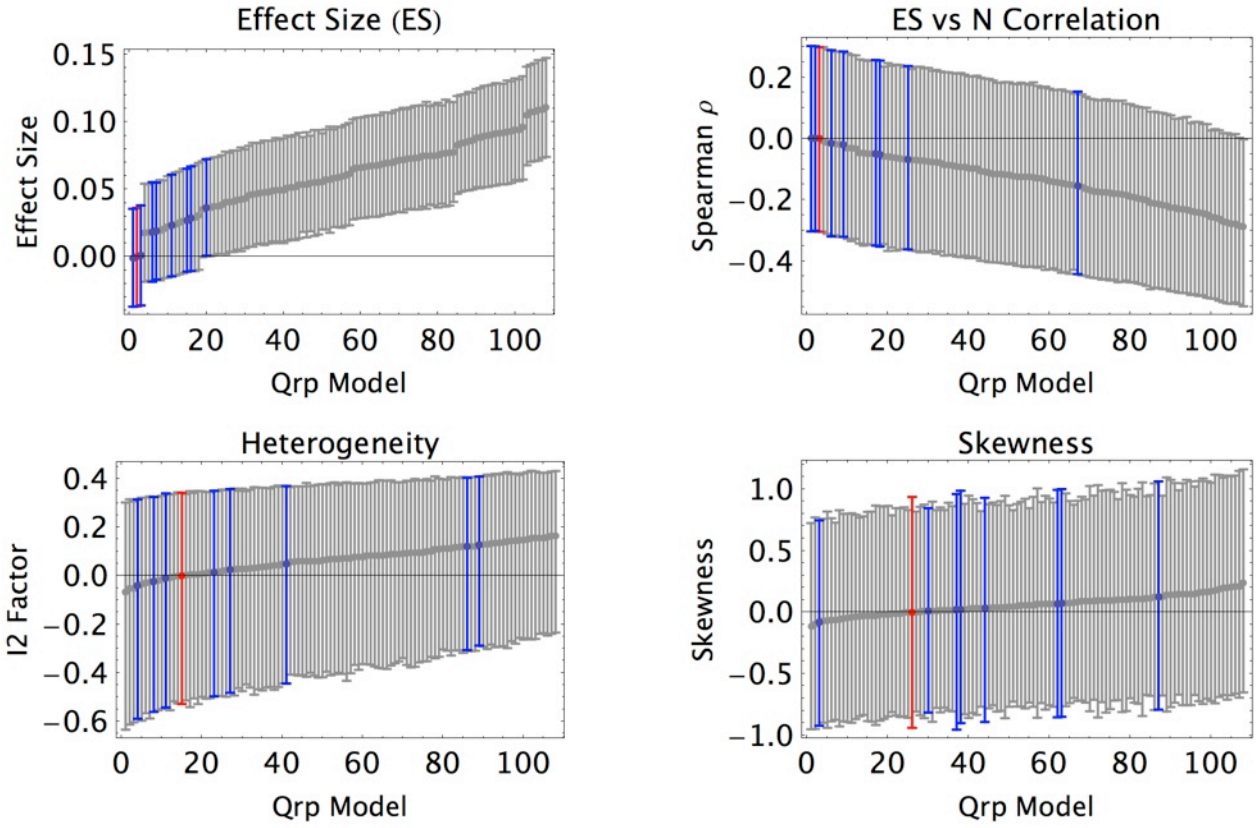

Figure 4. Parameter values and CIs for a Null effect with 108 different Qrp combinations. The Qrp prevalences are set to 100%. Red points are for no Qrps and blue points are single Qrps. Gray points are combinations of 2 or more Qrps. Error bars give the 96% CI.

Simulation Fitness function

The Fitness function provides a goodness-of-fit for combined tests of model p-curves and parameter values relative to the experimental data from a meta-analysis. For p-curves, the interval, 0 to 1, is divided into bins and the model expectancy for each bin is compared to the actual occupancy of P-values from the meta-analysis. A P-value for each bin's occupancy is then determined by a binomial test. For parameters, the parameter mean value and its distribution are determined via model simulations. This allows the estimation of a P-value for the experimental parameter. The set of test P-values are then combined to give a Fitness. Herein, the Fitness is taken as the Fisher Chi of P-values (that is, the sum of P-value logarithms, multiplied by -2).

BSB binned the p-curve into 10 bins of equal width. It is a sub-optimal choice since Qrp effects on the p-curve are not smoothly distributed (see Figures 2 and 3). The Fitness power can be improved by choosing bins with a high dispersion under different Qrps. Six bins were used in these simulations. The bins, which exclude about half of the p-curve, are: 3 low P-value bins of width 0.01 (centered at 0.005, 0.015 and 0.035), a bin just above the 0.05 P-value (from 0.05 to 0.20), a broad bin from 0.5 to 0.7 and a bin at high P-value (from 0.9 to 1).

The last bin helps distinguish Qrps, but is also useful in detecting negative effects, such as psi-missing. It was decided to retain only the effect size parameter since it has the highest dispersion.

A linear basis method for Qrps

A full simulation searches the space of all Qrp models for those that minimize the Fitness. The space is very large since it comprises all 108 Qrp combinations, across all possible prevalences. Simulations covering the entire model space are hardly feasible and BSB used a genetic search algorithm to render the problem tractable. The algorithm nevertheless imposes heavy computational demands. An alternate, faster method can be constructed by observing that the Qrp effects on both the p-curve and the parameters combine in a nearly linear fashion. It is thus possible to construct a method that performs a Monte Carlo simulation for each individual Qrp only once, and then uses these as a basis set for generating models with different combinations and prevalences. The method reduces the number of Monte Carlo simulations from tens or hundreds of thousands to a handful.

The basis method for the p-curve is straightforward. A Qrp simulation tracks the evolution of study P-values and generates a transfer matrix that models how P-values should be updated under the Qrp's influence. P-curves with combinations of Qrps are calculated by multiplying together the transfer matrices, in their order of occurrence. The basis set of p-curves then allows for direct calculation of the p-curve for any Qrp model. Tests that feed into the Fitness are had by fitting the p-curve to the experimental data. When prevalences are less than 100%, the matrices are replaced by the weighted sum of the transfer matrix and the identity matrix (here, the fractional prevalence is denoted as α):

$$M(\alpha) = \alpha M(1) + (1-\alpha)I.$$

The basis method produces highly accurate estimates of model p-curves. This can be seen by comparing the basis method p-curves with those generated by direct Monte Carlo simulation. Figure 3 overlays the basis method and simulated p-curves. The tiny visible discrepancies are statistical and controllable by adjusting the size of the basis simulations (there is evidence for some systematic discrepancies, but they are too small to impact the Fitness).

Parameters can be estimated by α-weighted linear combinations of their Qrp basis values. For parameters, interactions between Qrps cause deviations from the linear estimates. Interactions are strongest when publication bias is present and are most severe for the small study correlations and especially the heterogeneity. However, the estimates are quite good for the effect size (it is possible to put in ad hoc second-order corrections, if deemed necessary). Figure 5 compares the linear estimation of 3 parameters with those determined by full simulations. The plots show the evolution of 108 Qrp models as the α level is varied uniformly (red traces are Qrp combinations that include publication bias). Interactions between Qrps cause non-linear deviations in the traces. The Figure shows that interactions are small for the effect size and quite severe for the heterogeneity.

Next, an estimation of the parameter P-value is needed for the Fitness function. It is devised by observing that the width of confidence intervals is quite stable across Qrp models (see Figure 4). Further, it can be verified that the simulated values closely follow a Normal distribution (this does not hold for the heterogeneity, however). P-values can thus be estimated by comparing experimental parameter values to a Gaussian that is centered on the linear estimate and with a width determined by the weighted sum of basis widths. The CDF of the Gaussian then gives the parameter P-value.

The linear basis model provides a fast method for finding Qrp models that minimize the Fitness function. An initial simulation of precision Qrp bases requires 10 to 20 minutes of CPU time. The space of Qrp models is then sampled to return a Fitness value for each model. Finally, a precise P-value for the Fitness is obtained by a single simulation of the model that minimizes the Fitness. With the current implementation, calculating the Fitness for 100,000 models takes about 3 minutes. The basis method gives a speed-up of roughly 10,000 over direct Monte Carlo simulation and a 1,000-fold speed-up when simulations are aided by efficient search algorithms.

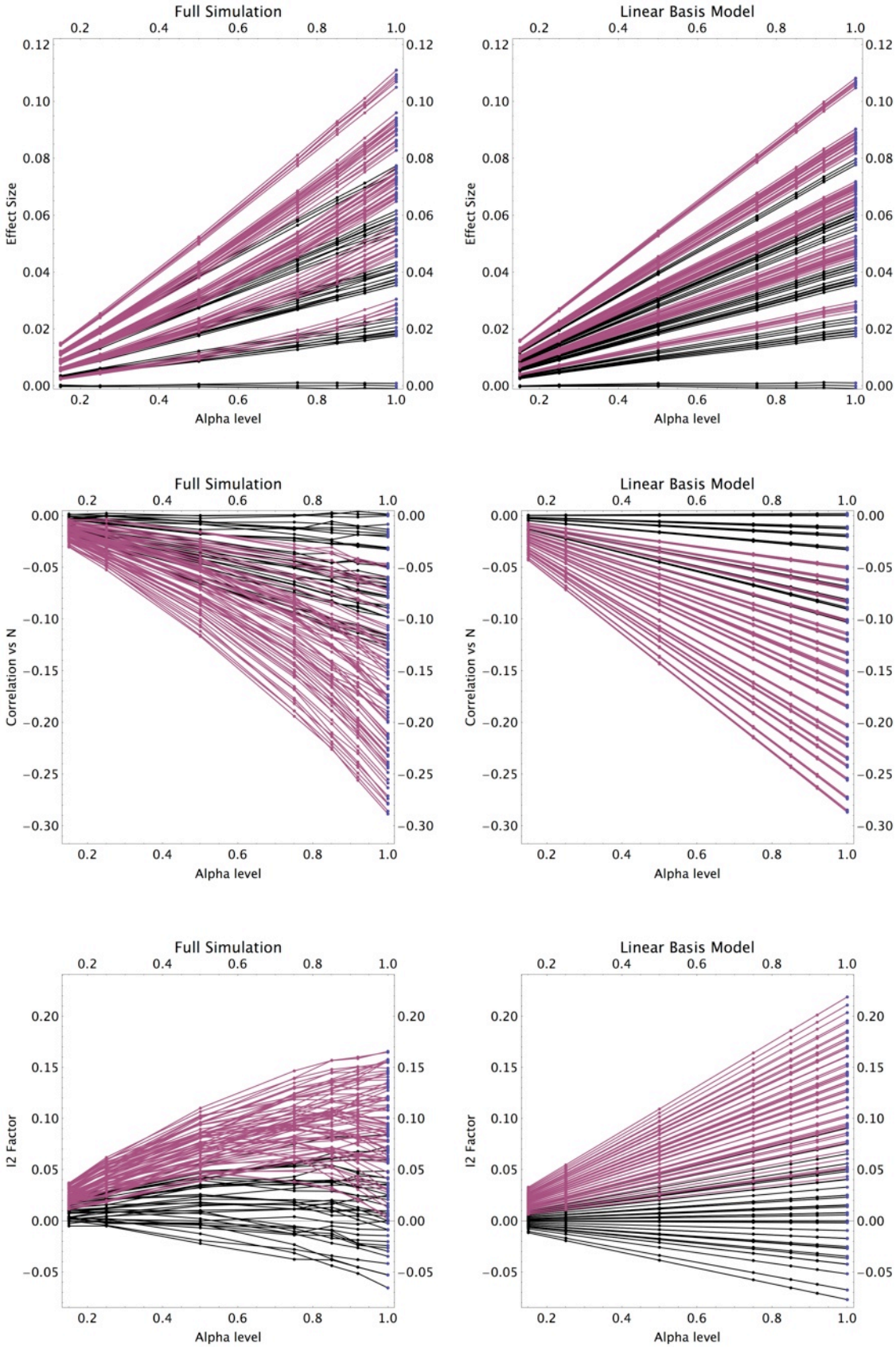

Figure 5. Comparisons of the linear basis method to full simulations for 3 parameters. Each of 108 Qrp combinations are evolved with uniformly varying prevalences. The red traces indicate the presence of the publication bias Qrp.

Discussion and Results

The steps of Qrp simulation for meta-analyses can be summarized as follows:

1. Determine an appropriate set of Qrps and devise a model for each.
2. Define the Fitness function and specify its test parameters (p-curve, effect size, etc.).
3. Optimize binning for the p-curve.
4. Calculate a Qrp basis set via Monte Carlo simulation.
5. Calculate the Fitness over a sampling of Qrp combinations and prevalences.
6. Identify the model that minimizes the Fitness and run a final simulation to estimate its P-value.
7. Make inferences about the viability of the Qrp hypothesis for the meta-analysis.

How Qrp analysis works in practice is demonstrated by applying it to three databases: the Ganzfeld database analyzed by BSB, and the micro-PK meta-analysis (Bösch, Steinkamp & Boller, 2006) and that of the GCP.

The Ganzfeld meta-analysis

For the Ganzfeld database, all 79 studies report the total number of trials and hits, and derive P-values from the binomial hit rates. Because the unique hit rate is the only reported effect size for all studies, the multiple analysis Qrps are not retained in the analysis. This is an example of how Qrp sets are adapted to specific meta-analyses. To handle the fraud Qrp, 3 studies were dropped from the database (4%). These were the most significant study and two others with P-values just above and below the 0.05 level.

The model space was searched in α steps of 0.025. The best fit returned a Fitness of 21.77. Its model had an α of 0.95 for Replacement, and α's of 1 for the Restart, Publication bias and Extension Qrps. Thus, the best fit for the Ganzfeld has all Qrps operating near 100% prevalence. The Fitness P-value is $P = 0.053$ which indicates that Qrps give a poor fit of the Ganzfeld data, even under the extreme scenario of maximal prevalences. The Trial Replacement Qrp contributes strongly to the Fitness minimum. Recall that this Qrp is modeled aggressively as the biased replacement of 5% of all trials. The near uniform adoption of such aggressive Replacement appears implausible, yet this level of application is required for the best Qrp fits. If analyses use a higher level of fraud, by removing the next two most significant studies, the Fitness P-value becomes 0.083. Thus the Fitness minimum changes only slightly with higher fraud rates, and it again requires all prevalences at or near 100%.

The Fitness function changes smoothly within the model space. Generally, better fits are had when Publication bias is at 100% and Restart and Replacement Qrps have prevalences of above 70%. Figure 6 plots the Fitness P-values for a range of Restart and Replacement prevalences. The diagonal contour steps are due to the strong correlation between these two Qrps.

Finally, the heterogeneity of the best fit model is $I2 = 0.10$ (90% CI: -0.16, 0.30). This can be compared with the experimental values in the absence of fraud, and at the two fraud rates of 4% and 6%. The I2 values are, respectively, 0.37, 0.30 and 0.21. Thus, the best fit accounts for the heterogeneity only if fraud is included in the model.

Several conclusions can be drawn. First, optimizing the Fitness increases the analysis power. BSB found acceptable fits when Qrp prevalences were unrestricted, but here, all fits are poor even with prevalences ranging up to 100%. Second, this result provides a quantitative basis for evaluating the Qrp-Null hypothesis. The mapping of the Fitness in model space can help estimate the levels of hacking and Publication bias required for a Qrp that can fit the data. Third, the basis method is accurate and fast. It makes it easy to extend the scope of analyses and quantify alternate hypotheses. A large Monte Carlo simulation for the one model that gave the best fit converged to a P-value of 0.053. This compares well to basis method estimation of 0.047. Yet, the basis method runs more than 10,000 times faster than a full simulation of all models. Fourth, the results show which Qrps have the greatest impact. This can inform inclusion criteria for meta-analyses. Lastly, the analysis reinforces the evidence for a Ganzfeld effect. It provides an answer to the question: How well do we know it's not Qrps?

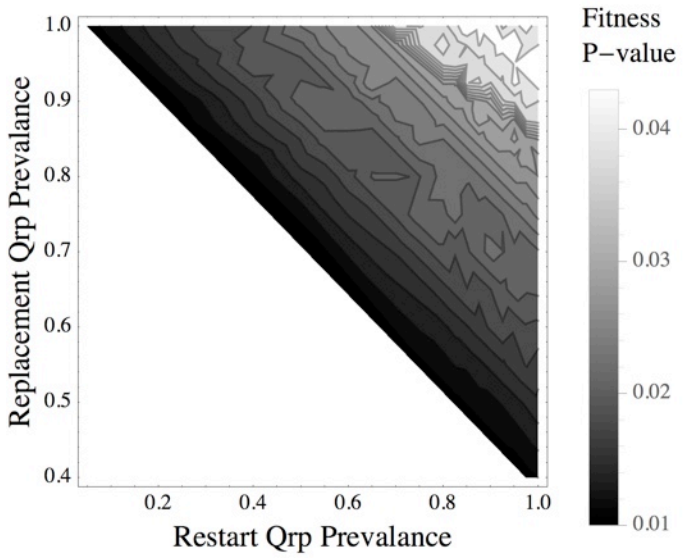

Figure 6. Contour plot of Fitness P-values for variable Replacement and Restart (R&R) Qrp prevalences (the Publication bias is set at 100%). The best fit, with P=0.053, has α's for R&R of (0.95, 1). Note that P-values decrease with α (indicating poorer fits), especially when α falls below 0.90. P-values fall below 0.01 precipitously in the triangular lower left half of the plot (in white).

The micro-PK meta-analysis

The micro-PK case provides an example of how Qrp analysis is useful even when simulations are difficult to run. In RNG-PK experiments, the effect size is often taken as the measured deviation of bit-probabilities. However, when the number of bits in experimental trials vary, bit-probabilities are not suitable for comparing effect sizes across experiments (Jahn et. al., 2000). In the Bösch et al. database examined here, the bits per trial vary by orders of magnitude across studies and combining them in the meta-analysis is troublesome for precisely this reason. Nevertheless, the Bösch et al. meta-analysis of micro-PK concluded (somewhat speculatively, and to sweep much under the rug) that publication bias might indeed account for the apparent effect. A particularity of the micro-PK database is that, aside from exhibiting small, yet significant, FEM and REM effects the heterogeneity of effect sizes is extremely high. A standard heterogeneity measure, the inconsistency factor, I2, has a value of 0.75 for the database (I2 measures the fraction of variability in effect sizes that cannot be explained by, or is inconsistent with, statistical noise). Importantly, their attempt to model publication bias only accounted for about half of the heterogeneity. The excessive heterogeneity, then, may give evidence of an effect, independent from problematic estimations of the effect size.

The absence of a satisfactory effect size definition means we can't simulate Qrps for the whole database. This is further complicated because Qrps involve interventions on the part of the experimenter down to the level of individual trials (the smallest datum unit available for manipulation), and this information is not always available in published reports. However, a subset of studies in the micro-PK database does provide enough information. Of the 377 studies in the database, 86 employed homogeneous protocols with one intervention and one bit per trial.

Qrp analysis for the subset of studies gives a Fitness minimum of 42.4. A Monte Carlo simulation of the minimum model yields a Fitness of 44.0, corroborating the estimate from the linear basis analysis. The simulation value corresponds to a z-score of 3.72, indicating that the best Qrp fit can be strongly rejected as an explanation for the data. While we can't extend this conclusion to the full micro-PK database, the heterogeneity and correlation with study N indicate that the subset of studies is nevertheless representative of the whole. The I2 and Spearman's rho for the subset are 0.68 and -0.38. These compare well with the values for the full database: 0.75 and -0.33. Lastly, the subset I2=0.68 far exceeds that of the best fit simulation (I2=0.33), and lies outside the 99.99% CI of the simulation estimate. The Qrp analysis thus likely explains why Bösch et. al. could not reproduce the heterogeneity: maximal combinations of Qrps will not generate the extreme heterogeneity seen in the micro-PK database. Indeed, the large heterogeneity supplies independent evidence for an effect since neither statistical noise nor Qrps can account for it.

The GCP database

The GCP hypothesizes that the occasional collective emotion or attention of large populations, such as when people react to significant world events in real time, will impact the output of RNGs. Since 1998, the Project has maintained a worldwide network of synchronized RNGs to test its hypothesis. Through December 2015, the GCP tested 496 'global events' (such as large terror attacks or New Year's celebrations). The average z-score of data deviations corresponding to events is 0.33, which rejects the null hypothesis by 7 standard deviations.

The experiment is interesting for Qrp analysis because it pre-registers each test in a public online repository. The GCP is thus an example of a protocol that precludes Qrps: tests and analyses are publicly pre-registered and all raw data is available for download. Figure 7 shows that the GCP data accords with the theoretical p-curve for the measured effect size. Additionally, the heterogeneity I2 value is 0.15, a low value that is consistent with statistical noise.

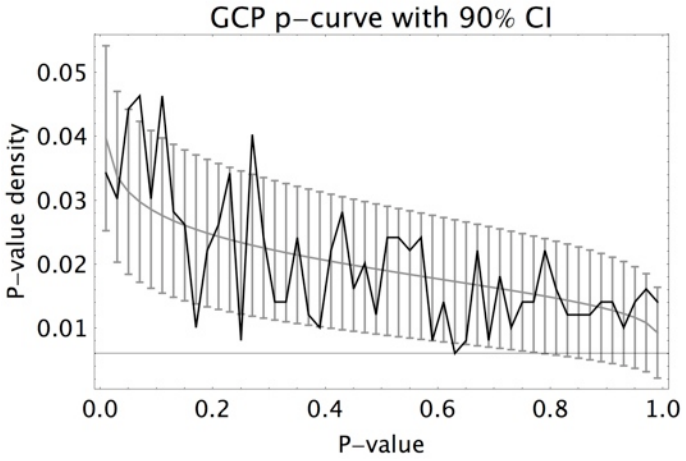

Figure 7. The GCP p-curve (black trace) and the expected p-curve for a homogeneous effect given the experimentally determined effect size (gray trace). The p-curve is from 496 registered 'events' that test the GCP hypothesis. The error bars represent 90% CI's for bins of 0.02 widths in the P-values. The data are a good fit to the theoretical curve: 6 of 50 points slightly exceed the 90% intervals, which is statistically consistent with the theoretical curve.

The GCP, then, provides strong evidence for an anomalous effect. But does the GCP result support the global consciousness hypothesis, or is it indicative of more familiar psi effects such as those evidenced by the Ganzfeld telepathy protocol? A detailed analysis, undertaken for this Bursary grant, unequivocally favors the latter (Bancel, 2016). The analysis tests the GCP result against the alternate hypothesis of experimenter psi, for which intuitive psi informs the experimenter's choice of test parameters (such as the free selection of start and end times for the timing of events made by the experimenter when event tests are pre-registered). The analysis shows that the psi selection hypothesis can be tested by devising an operational definition of goal-oriented psi effects, of which psi selection is an example. Tests on the data all confirm the selection hypothesis and refute a causal explanation due to global consciousness (for details see (Bancel, 2016), a copy of which accompanies this report). The conclusion that the GCP result is a real psi experimenter effect due to selection, and not a consequence of global consciousness, leads immediately to a further testable hypothesis of the data. The selection hypothesis states that choices of start and end times of data test periods will favor the inclusion of natural, statistical data fluctuations that deviate in the direction of hypothesis confirmation. A consequence of the model is that the excluded data lying just *outside* the designated test periods will necessarily deviate in the opposite direction. The hypothesis is testable because the GCP database contains a continuous record of the RNG network's output, including all data outside of test periods. Figure 8 compares the test periods of GCP events with the proximate data 12 hours before and after the events. The GCP measures correlations among RNGs in the global network and the hypothesis states that the correlations deviate positively when an effect is present. To visualize the data, 349 events (those that allow free choices of both start and end times of test periods) are summed together. The data are integrated so that trends are easily visible; an effect is then seen as a positive slope of the data trace. The black line is the formally registered event data and the gray lines are the sums of 12 hours of data just before and after events. The positive slope of the black line represents a 5.9 sigma

deviation for the data. This is the GCP result. It is easily seen that the proximate data deviate strongly in the opposite sense and cancel the cumulative deviation of the event data. This provides a stark visual confirmation of the selection model and is contrary to predictions of the global consciousness hypothesis.

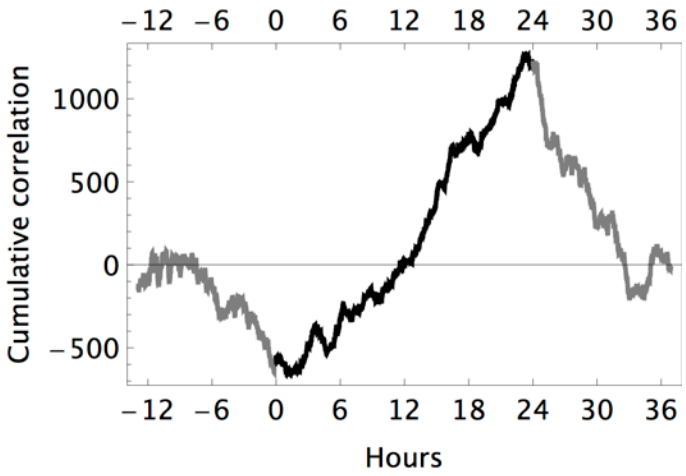

Figure 8. Cumulative correlation of events that freely designate event start and end times. The black trace is the summation of 349 standard analysis events that have been uniformly stretched over a 24-hour period in order to provide a better visualization of the data. The gray traces are proximate data for the same events, extending 12 hours before and after event start and end times. The curve has been shifted vertically to align the endpoints of the proximate data with zero. The rise of the black trace corresponds to a 5.6-sigma deviation. It is entirely cancelled by negative deviations in the proximate data, as predicted by the selection model.

Conclusions

This Bursary grant has developed a fast, efficient method for assessing the influence of Qrps on meta-analytical inference. The method has been applied to three different parapsychological databases, demonstrating how it can be adapted under quite disparate conditions. The Ganzfeld database allows for a full application of the method. The analysis shows that Qrps cannot provide an adequate explanation for the anomalous Ganzfeld effect and thus reinforces the evidence for psi under this well-established protocol. It provides a stronger conclusion than the earlier analysis of Bierman et al. due to significant improvements in the method's power that were devised during the course of this work. The micro-PK database gives an example of how limitations in the quality of literature and our understanding of underlying effects can be addressed by Qrp analysis. This was done by examining a data subset, as well as the database heterogeneity. While the evidence for an anomalous effect is statistically stronger than for the Ganzfeld data, assumptions made for the analysis make the case less direct and probably less compelling for those not intimately familiar with the literature. However, the analysis clearly shows that the conclusion of the original meta-analysis (that publication bias may explain the data) is not viable. It thus clarifies and advances our understanding of this important sub-field of parapsychology. The GCP is a case where the question of Qrp effects is rendered moot by pre-registration. P-curve and heterogeneity analyses are consistent with a real effect. This allows going a step further to ask whether the GCP result confirms its hypothesis. A detailed analysis, which introduced a novel, operational definition of goal-oriented effects, demonstrates clearly that the answer is no. All evidence supports the interpretation of an experimenter selection effect, and contradicts causal explanations based on the hypothesized global consciousness.

There are a number of obvious follow-ups and extensions to this work.

1. The size of the Ganzfeld effect can be estimated by models that include real effects. BSB found a residual hit rate of 27% (compared to the Null value of 25%), after allowing for Qrps. It will be interesting to see if the

improved power of the basis method yields a different value. This will also provide a needed reference point for any future replication of the Ganzfeld protocol.

2. The original Bursary proposal was to analyze pre-cognition and presentiment meta-analyses and the completion of this work sets the ground for applying Qrp analysis to these important databases. In particular, an update of the presentiment meta-analysis has just been completed (P. Tressoldi, private communication) which includes the newest published reports. It augments the meta-analysis size, which is helpful from the point of view of power considerations.

3. It would be good to present the Qrp basis method to a wider audience via publication in a mainstream journal. Because there are efforts these days to replicate established effects in a variety of fields, I suspect that it will be possible to find published meta-analyses outside of parapsychology that have been confirmed or rejected by well-powered, pre-registered replications. Such would provide a verifiable test demonstration for Qrp analysis.

4. The code for this project was written in the Mathematica language. This is not accessible by most researchers in psychological, economic or medical disciplines. Porting the code to a more widespread platform, such as R, would give an added-value to the product of the grant.

5. My intuition is that further improvements in the method's power are possible and this is an avenue that should be explored.

6. A weakness that Qrp analysis shares with any modeling program is that one cannot be sure that models accurately reflect real-life situations. A possible improvement might be the creation of generic Qrps for the three Qrp categories of data-hacking, p-hacking and publication bias. This certainly would be an interesting direction to pursue, particularly in light of a successful completion of point 3 above.

REFERENCES

Bancel, P. A., (2016). Searching for Global Consciousness: A 17-Year Exploration. *Explore: The Journal of Science and Healing*, 13(2), 94-101.

Bem D., Tressoldi P. E., Rabeyron T. & Duggan M. (2015). Feeling the future: A meta-analysis of 90 experiments on the anomalous anticipation of random future events. *F1000Research*, 4, 1188.

Bierman, D. J., Spottiswoode, J. P. & Bijl, A., (2016). Testing for Questionable Research Practices in a Meta-Analysis: An Example from Experimental Parapsychology. *PLoS ONE* 11(5): e0153049.

Bösch, H., Steinkamp, F., & Boller, E. (2006). Examining psychokinesis: The interaction of human intention with random number generators--A meta-analysis. *Psychological Bulletin, 132*(4), 497-523.

Fiedler, K. & Schwarz, N., (2016). Questionable Research Practices Revisited. *Social Psychological and Personality Science*, 7(1), 45-52.

Jahn, R., Dunne, B., Bradish, G., Dobyns, Y., Lettieri, A., Nelson, R., Mischo, J., Boller, E., Bösch, H. Vaitle, J., Houtkooper, J. & Walter, B. (2000). Mind/Machine interaction consortium: PortREG replication experiments. *Journal of Scientific Exploration*, 14, 499–555.

Jin, Z-C, Zhou, X-H, & He, J., (2014). Statistical methods for dealing with publication bias in meta−analysis. *Statistics in Medicine*, 34(2), 343-360.

John, L. K., Loewenstein, G., & Prelec, D., (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524-532.

Mossbridge, J., Tressoldi, P. E. & Utts, J. (2011). Predictive physiological anticipation preceding seemingly unpredictable stimuli :a meta-analysis, *Frontiers in Psychology* 3, 390.

Simonsohn U., Nelson L. D., Simmons J. P., (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, 143, 534–547.

Storm, L., Tressoldi, P. E., & Di Risio, L. (2010). Meta-analysis of free-response studies, 1992–2008: Assessing the noise reduction model in parapsychology. *Psychological Bulletin, 136*(4), 471-485.

Varvoglis, M. & Bancel, P. A., (2015). Micro Psychokinesis. In: Cardeña E., Palmer J., & Marcusson-Claverts, D. *Parapsychology: A Handbook for the 21st Century*. Jefferson, N.C.: McFarland & Co.