 Mind & Matter Vol. 3(1), pp. 85–132

Meta-Analysis of Mind-Matter Experiments: A Statistical Modeling Perspective

Werner Ehm
Department of Theory and Data Analysis
Institute for Frontier Areas of Psychology and Mental Health
Freiburg, Germany

Abstract

Are there relationships between consciousness and the material world? Empirical evidence for such a connection was reported in several meta-analyses of mind-matter experiments designed to address this question. In this paper we consider such meta-analyses from a statistical modeling perspective, emphasizing strategies to validate the models and the associated statistical procedures. In particular, we explicitly model increased data variability and selection mechanisms, which permits us to estimate "selection profiles" and to reassess the experimental effect in view of potential other effects. An application to the data pool considered in the influential meta-analysis by Radin and Nelson (1989) yields indications for the presence of random and selection effects. Adjustment for possible selection is found to render the, without such an adjustment significant, experimental effect non-significant. Somewhat different conclusions apply to a subset of the data deserving separate consideration. The actual origin of the data features that are described as experimental, random, or selection effects within the proposed model cannot be clarified by our approach and remains open.

1. Introduction

Among the manifold aspects of mind-matter research the following question has received particular attention: are there interrelations between the "state of mind" of a conscious subject and the "state of matter" of a physical system? In numerous experiments directed at this question, states of matter have been measured as some output characteristics of a physical system implementing a random event generator (REG). States of mind, being less easily measurable, have been operationally defined as an intention the subject is instructed to "have" with respect to the physical system. The question then becomes: are there experimentally verifiable relations between subject intention and the output of a physical REG?[1]

[1]The original motivation for studying such relations seems to derive from particular indeterministic aspects of quantum theory suggesting the possibility of mental causation upon material systems. The prototype of corresponding mind-matter experiments can be traced back to Schmidt (1970) and Jahn (1982).

A comprehensive systematic review (meta-analysis) of the results of all related experiments known until 1987 was published by Radin and Nelson in 1989 (subsequently R&N). R&N's conclusion was that there is a small but highly significant experimental effect. The article ends with the following summary:

> The overall effect size obtained in experimental conditions cannot be adequately explained by methodological flaws or selective reporting practices. Therefore, after considering all of the retrievable evidence, published and unpublished, tempered by all legitimate criticisms raised to date, it is difficult to avoid the conclusion that under certain circumstances, consciousness interacts with random physical systems. Whether this effect will ultimately be established as an overlooked methodological artifact, as a novel bioelectrical perturbation of sensitive electronic devices, or as an empirical contribution to the philosophy of mind, remains to be seen.

R&N's analysis and the experience of the Princeton Engineering Anomalies Research (PEAR) laboratory with such experiments encouraged a large-scale replication study that was carried out by a Mind/Machine Interaction Consortium in three different laboratories according to a strict design. The primary result of the study was that the overall summary statistics failed to be statistically significant (see Jahn *et al.* (2000) for a detailed account). This discrepancy in the outcomes of two major efforts of an experimental verification of mind-matter relationships[2] calls for further investigation.

In the present paper we approach the problem by reconsidering (the data of) R&N's meta-analysis from a statistical modeling perspective. Since 1) experimental and control runs are not matched, ruling out a direct use of the latter in the analysis of the former, and 2) the control data appear essentially compatible with the null-hypothesis of pure randomness, comparisons are made against the theoretically expected behavior under control conditions. Figures 1 and 2 in the article of R&N show clearly that the z-score[3] distribution under experimental conditions differs from the (empirical and theoretically expected) one under control conditions. Under experimental conditions there is apparently a shift to a higher mean, and an increased dispersion. The question to be dealt with in the first place then is *what kind of mechanism might have generated the experimental data.* More specifically, we ask for explicit statistical models describing random mechanisms[4] that are

[2] See also Radin (1997), Radin and Nelson (2002), and others.

[3] A z-score is a summary statistic comprising the results of a study in a single numerical value; cf. Section 2.

[4] "Random *mechanism*" is understood in the sense of probability theory, which

a) interpretable in terms of the experimental setting and
b) capable of producing data similar to those considered by R&N.

As long as a developed theory of mind-matter relations is lacking, an explanation of the data cannot possibly aim at an identification of "the" underlying mechanism; rather the goal is to provide (models of) mechanisms satisfying a) and b). Both requirements are crucial: a) helps to reduce arbitrariness and makes the proposed models amenable to detailed criticism; b) acts as a censor ruling out grossly inadequate models. Questions related to experimental effects are dealt with in the second place, and only in the context of models that are not obviously wrong. This restriction pays attention to the general statistical wisdom that conclusions based on an inadequate model can be misleading.

Empirical studies directed at a particular subject often vary in experimental details and come up with results yielding no coherent picture. Meta-analysis emerged as a tool for settling related issues in the social sciences, medicine, and other fields, by systematically searching the relevant literature and gathering the individual results in an overall statistical analysis. Its major aims are (E) strengthening of evidence, and (G) generalization. Regarding (E) the idea is that an isolated study is often too limited to be conclusive on its own, whereas suitable combinations of the results from many scattered studies should yield strengthened hypothesis tests and statistically more stable "effect size" estimates. (G) aims at clarifying the scope of the results, i.e. the extent to which results obtained with diverse protocols and goals are extensible to more general settings. Clearly, from this perspective diversity of studies has to be considered as necessary. On the other hand, such diversity renders averaging procedures doubtful, and has prompted criticism of meta-analysis on the whole. From the vast literature on meta-analysis we refer to Hedges and Olkin (1985) for methodology, to the Antiplatelet Trialist's Collaboration (1994) for a famous example, to Thompson and Pocock (1991) for criticism, and to Utts (1991) and, from a different point of view, Atmanspacher and Jahn (2003) for the question of reproducibility.

Well-known problems in meta-analysis arise from (V) high data variability and (S) selective reporting practices. (V) is commonly seen as reflecting the heterogeneity of the studies, and accounted for by allowing for extra variation due to study-specific covariates and/or random effects. The most prominent instance of (S) is publication bias, which results from significant studies being more likely to be published than non-significant ones. Methods proposed to assess or deal with selection bias include: Rosenthal's (1979) popular "fail-safe n" method of estimating the number of unpublished studies "resting in the file drawer"; graphical procedures

describes a mechanism phenomenologically by the statistical structure of its output, with no reference to its physical functioning.

and related formal tests such as the one by Egger *et al.* (1997); and diverse variants of explicit selection models proposed in, *e.g.*, Iyengar and Greenhouse (1988) and Copas (1999).

Another specific problem results from the fact that the (raw) source data from single experiments usually are unavailable in a meta-analysis; rather, only an overall summary statistic (*e.g.*, a z-score) is known for each experiment. Therefore, inferences from the z-scores based on models for the source data involve all the difficulties of an inverse problem. See *e.g.* Hedges (1992, Conclusions) and Copas and Li (1997, Introduction) for related cautionary remarks. Such difficulties are in fact unavoidable since the distribution of the z-scores depends on the inaccessible stochastic structure of the source data, of which the z-scores are functions. On the other hand, assumptions about the source data have implications for the statistical properties of the z-scores, hence can be examined in an indirect manner at least. This will suffice to reject some common models as implausible, so that some rudimentary model checking is possible also in meta-analysis.

In this study the source data of an experiment are the single bits of the REG. Accordingly, the statistical modeling starts at the bit level of the REG and introduces step by step "degrees of freedom" that describe possible deviations of the REG's behavior from a strict, "pure chance" Bernoulli scheme. In line with the subject literature, the main experimental effect is modeled by parameter changes corresponding to a mean shift at the level of the z-scores. Further model components account for increased data variability, through unsystematic/random as well as systematic modifications of the main effect and for possible selection, where we adopt a model proposed by Hedges (1992). As a special feature of our study, we trace the choices made in the modeling process, discuss the respective rationale, and check for consistency with the data.

The model to be developed allows, *e.g.*, to determine a "net" experimental effect, by correcting it for the contributions of other effects – in principle. At the same time it offers multiple explanations for observed phenomena that are not easily, or not at all, separable on the basis of the available data. This raises the question whether related statistical procedures are capable of disentangling the single model components and, especially, of distinguishing between experimental (shift) and selection effects. Moreover, there might be a considerable loss of efficiency if in corresponding testing problems adjustment is made for an effect that is confounded with the one to be tested. We investigate these questions through Monte Carlo simulation, by generating surrogate data under parameter configurations that imply or exclude the respective effect.

As in any statistical analysis, the conclusions and interpretations for concrete data sets depend on the model assumptions and can fail if the latter are inappropriate. In order to diminish this risk considerable em-

phasis is laid on model checking. Still, the model assumptions cannot formally be verified with the given data, and there should be no ambiguity about the conditional character of our results. Final answers are not the intention of this work.

The basic framework of this paper is set in Section 2. In Section 3 the generalized Bernoulli paradigm is introduced and the random coefficients regression model developed from it; Section 5 extends the model by incorporating selection effects. Sections 4 and 6 present the results of an analysis of the R&N data on the basis of these models. The main results appear in Section 7, which is devoted to a critical assessment of the various statistical procedures under conditions relevant to the R&N data. A brief summary in Section 8 concludes the main part of the paper. Three appendices provide details about a mathematical derivation, the data base, and implementation issues.

2. Basic Framework

The basic material constituent of each experiment included in the meta-analysis of R&N is a device serving as the source of intrinsic randomness. Its prototypical form is a physical random event generator (REG) producing a stream of bits coded as 0s and 1s. This REG is intended to mimic perfect randomness so that, ideally, the bit stream generated under control conditions should behave like a sequence of independent and identically distributed (i.i.d.) Bernoulli trials. Therefore, *the individual bits* $Y_1, \ldots, Y_N$ *are supposed to represent (the realization of) i.i.d. random variables* Y_j *assuming the values 1 and 0 with probabilities* p_0 *and* $1 - p_0$*, respectively.* Here N denotes the total number of Bernoulli trials (bits) collected in the experiment, and p_0 is the "by-chance hit probability", a characteristic of the REG. Subsequently we shall refer to this assumption as the *strict Bernoulli paradigm* (sBp).

The summary statistic entering the meta-analysis is

$$Z \;=\; \frac{S - Np_0}{\sqrt{Np_0(1-p_0)}}\,, \qquad \text{where} \;\; S \;=\; \sum_{j=1}^{N} Y_j$$

denotes the total number of 1s ("hits") among the N bits. What counts as a hit depends on the experimental condition. We come back to this in Section 3.4.3.

Under the sBp, S has the binomial distribution $\mathcal{B}(N, p_0)$, and Z represents a *z-score* for the experiment: it has expectation zero and variance one under the sBp, and is approximately normally distributed if N is not too small. Usually N is large, and then the normal approximation $\mathcal{N}(0, 1)$ to the exact distribution of Z is sufficient for all practical purposes.

Passing now from a single experiment to the set of all n experiments considered in a meta-analysis, let any quantity associated with the i-th experiment be equipped with the index i. (Conversely, if the index i is missing that means that attention is temporarily focused on a single experiment.) The basic meta-analytic data consist of all triplets $(Z_i, N_i, p_{i,0})$, $i = 1, \ldots, n$, with Z_i the z-score, N_i the sample size (total number of bits), and $p_{i,0}$ the by-chance hit probability in the i-th experiment.

Under experimental conditions, the REG does not run on its own but in the presence of a conscious subject, called agent (or operator). The question is whether in this way correlations are induced between agent intention and the output of the REG. The basic experimental idea is to check whether a prescribed intention, *e.g.*, HI or LO, yields that 1s occur with increased or decreased frequency, respectively. On the basis of the data triplets $(Z_i, N_i, p_{i,0})$ available for the meta-analysis, such effects can only be detected if they imply a perturbation of the distribution of the z-scores. The traditional manner to characterize such changes is by way of *effect size* measures. Although this notion is widely used in meta-analysis without much discussion, from the perspective of an explicit modeling of alternatives it is much less straightforward and requires reconsideration; see Section 3.4.2.

In the following we propose statistical models intended to describe the stochastic behavior of REGs under experimental conditions. In compliance with the common conception of "the experimental effect" as a mean shift, we shall represent experimental effects within these models by (changes of) model parameters that describe translations of the z-score distributions.

3. Modeling Variability: Generalized Bernoulli Paradigm and the Random Coefficients Regression Model

3.1 Adding Parameter Variation

Suppose that in addition to the basic data triplets $(Z_i, N_i, p_{i,0})$, $i = 1, \ldots, n$, further characteristics are known for every study. Let these data be represented numerically as vectors $x_i = (x_{i1}, \ldots, x_{ip})$ whose individual components x_{ik}, called *covariates,* encode pieces of the additional information available for the i-th study. It is conceivable that a possible effect due to mind-matter relations is not universal but depends on, or is modified by, factors related to those encoded in the covariates. If this were the case one would expect that the distribution of z-scores Z_i varies in a systematic manner with the corresponding covariate vector x_i. The variability of the covariates would then explain some of the variability of

the z-scores. This is the basic idea of regression analysis. Further unsystematic variation not attributable to the measured covariates will also be accounted for, by allowing for random effects.

In setting up a concrete model for the REG's stochastic behavior under experimental conditions we pick up, and elaborate, an approach that underlies, explicitly or implicitly, all previous meta-analyses in the field. The idea is to suppose that in the presence of an intentionally focused mind, hit probabilities are slightly perturbed at the bit level compared to the by-chance hit probabilities $p_{i,0}$ applying under control conditions. We refer to such a setting as the *generalized Bernoulli paradigm* (gBp). The gBp involves the following assumptions.

(B1) Z-scores Z_i are independent across experiments and of the form

$$Z_i = \frac{S_i - N_i p_{i,0}}{\sqrt{N_i p_{i,0}(1-p_{i,0})}} \tag{1}$$

where $S_i = \sum_j Y_{ij}$ is the total number of "hits" in N_i Bernoulli(p_{ij}) trials.

(B2) Bernoulli variables Y_{ij} are conditionally independent across trials given the hit probabilities p_{ij}.

(B3) The p_{ij}s ($j = 1, \ldots, N_i$) are, themselves, random variables of the form

$$p_{ij} = s(\theta_{i,0} + \epsilon_{ij}), \qquad \epsilon_{ij} = x_i(\beta + \rho_{ij}), s(\theta_{i,0}), \tag{2}$$

where s is a scale function, $\theta_{i,0}$s are such that $s(\theta_{i,0}) = p_{i,0}$ for every i, and ρ_{ij}s are p-dimensional, centered column random vectors.

(B4) Bit-level effects ϵ_{ij} are uniformly small, and total numbers N_i of bits are large.

This requires explanation. First, a *scale function* is a smooth, strictly increasing function s from some interval J to the (open) interval $(0, 1)$, representing probability $p \in (0, 1)$ as a function $p = s(\theta)$ of some parameter $\theta \in J$. Assumption (B3) then means that bit-level effects ϵ_{ij} are additive in the θ-scale, i.e., in the domain of the scale function s. For example, in the linear scale corresponding to $s(\theta) = \theta$, $J = (0, 1)$, one has the common additive relation $p_{ij} = p_{i,0} + \epsilon_{ij}$. Schmidt (1975) suggested a scale function $s(\theta) = 1/(1 + e^{-\theta})$, $J = (-\infty, \infty)$, which corresponds to the "log-odds" (or logistic) scale widely used in the analysis of binary data. The name derives from the function inverse to $p = s(\theta)$, namely $\theta = \log(p/(1 - p))$, where the ratio $p/(1 - p)$ represents the odds of a "hit" compared to a "miss". With this scale function, bit-level effects are

defined as the difference of the log-odds under experimental and control conditions,

$$\epsilon_{ij} = \log(p_{ij}/(1-p_{ij})) - \log(p_{i,0}/(1-p_{i,0}))\,.$$

For more information about scale (or "link") functions see McCullagh and Nelder (1989).

Secondly, according to (B3), the bit-level effects ϵ_{ij} can be split into a systematic, non-random part $x_i\beta$ plus a random perturbation $x_i\rho_{ij}$. Here $x_i\beta$ denotes the inner product of row vector x_i with column vector β, so that $x_i\beta = \sum_{k=1}^{p} x_{ik}\beta_k$ is a linear function of the study-specific covariates ("regressors") x_{ik} with unknown regression coefficients β_k. This term is constant across bits within each study, but may vary across studies. The other term, the inner product $x_i\rho_{ij}$, adds unsystematic random variation at the bit level – "unsystematic" because ρ_{ij}s are assumed to be centered, i.e., to have expectation zero. From a slightly different perspective, the ρ_{ij}s may be regarded as (bit-level) random perturbations of the regression coefficient vector β, and $\beta_{ij} = \beta + \rho_{ij}$ as (a vector of) random coefficients. In the sequel we refer to the ρ_{ij}s as *random effects.* Suitable assumptions about their (joint) distribution are made below.

Finally, (B4) reflects the expectation that any changes due to the experimental conditions should be small. This plays an important role in the derivation of the random coefficients regression (RCR) model from the gBp.

3.2 Modeling the Bit-Level Random Effects

A particular problem in modeling the stochastic structure of the (bit-level) random effects ρ_{ij} ensues from the fact that in meta-analysis data are available only at the top level (*e.g.*, z-scores), but not at the bit level (*e.g.*, Y_{ij}s). At least, one knows in some cases, and may expect for others, that the individual bits Y_{ij} are collected in a certain hierarchical manner. For definiteness, suppressing the study index i for the moment, let us assume that:

> There are n_a agents participating in the study.
> Each agent performs n_s sessions.
> Each session consists of n_r runs.
> Each run consists of n_t trials.
> Each trial is a sum of n_b binary variables Y_j.

Then there are $N = n_a n_s n_r n_t n_b$ individual bits $Y_j = Y_{asrtb}$, whose sum

$$S = \sum\nolimits_{a,s,r,t,b} Y_{asrtb}$$

gives the total number of hits. For ease of reference, let us speak of the various levels in this hierarchy as *aggregation levels* (because individual

bits are collected along these stages). Random effects ρ_j may accrue at every aggregation level, from agent effects, session effects, etc. This suggests to decompose ρ_j as

$$\rho_j = \rho_{asrtb} = R^{(1)}_a + R^{(2)}_{as} + \ldots + R^{(5)}_{asrtb}\,, \tag{3}$$

where $R^{(1)}_a$, $R^{(2)}_{as}$, etc represent the respective contributions to the random effect ρ_j that are due to agent #a, to session #s of agent #a, etc. Note that if there were no such contributions (i.e., $\rho_j = 0$), then according to (B3) the bit-level effect would be constant across all bits of the study, meaning that all agents "perform" equally well, during all sessions, runs, and so forth, down to the bit level. Conversely, the above approach allows for differences between agents, for variability between different sessions of the same agent, and so forth. Concerning the single components of ρ_j in (3) let us make the following assumption.

(B5) All $R^{(l)}_{\ldots}$s are statistically independent, with expectations $E\,R^{(l)}_{\ldots} = 0$ and covariance matrices $\mathrm{Cov}(R^{(l)}_{\ldots}) = \Omega^{(l)}$ $(l = 1, ..., 5)$ depending on aggregation level l, but constant otherwise.

This assumption will have to be strengthened further in order to make the model tractable. Nevertheless, the decomposition (3) helps to gain insight into the possible impact of random effects at various aggregation levels.

3.3 From the GBP to the RCR Model

As shown in Appendix A, expectations and variances of the z-scores under the gBp are approximately of the form

$$E\,Z_i \;\simeq\; \gamma_i\, x_i \beta\,, \tag{4}$$

$$\mathrm{Var}(Z_i) \;\simeq\; 1 \;+\; \gamma_i^2\, x_i\, \Gamma_i\, x_i^T\,, \tag{5}$$

respectively. The quantities γ_i and Γ_i, related to the informativity of the i-th study and the dispersion of the random effects, are defined as

$$\gamma_i \;=\; \left(\frac{N_i}{p_{i,0}\,(1-p_{i,0})}\right)^{1/2} s'(\theta_{i,0})\,, \tag{6}$$

$$\Gamma_i \;=\; \mathrm{Cov}\Big(N_i^{-1} \sum\nolimits_j \rho_{ij}\Big)\,, \tag{7}$$

respectively. The transpose x_i^T of x_i is a column vector, and the term $x_i \Gamma_i x_i^T = \sum_{k,l} \Gamma_{i,kl}\, x_{ik} x_{il}$ in (5) is a quadratic form. For the RCR model it is assumed that the z-scores Z_i are *normally distributed* with these moments. This assumption can be justified at least in part, and appears

reasonable in view of the aggregation of the random effects; see Appendix A.

Initially, the gBp (*e.g.*, (B3)) implies no restriction on the covariance matrices Γ_i. Given a balanced hierarchical design as described in Section 3.2 along with decomposition (3), the Γ_is assume a special form. Suppressing again study index i, one has

$$\Gamma \;=\; \mathrm{Cov}\Big(N^{-1}\sum\nolimits_j \rho_j\Big) \;=\; \frac{\Omega^{(1)}}{n_a} + \ldots + \frac{\Omega^{(5)}}{n_a n_s n_r n_t n_b}\;. \qquad (8)$$

The last expression shows that the random effects associated with each aggregation level are weighted differently according to the depth of the level: the covariance $\Omega^{(5)}$ associated with random effects at the lowest (bit) level is weighted inversely to $n_a n_s n_r n_t n_b$ = total number of bits N, whereas the covariance $\Omega^{(1)}$ of the random effects at the highest (agent) level has a weight inverse to n_a = number of subjects participating in the study. Consequently, since N may be much larger than n_a, the overall extra variability expected if random effects occur at the agent level may outrange by orders of magnitude the extra variability expected if random effects occur only at the bit level.[5] To simplify matters, let us proceed with the following assumption replacing (B5).

(B5') The covariance matrices Γ_i defined by (7) are of the form $\Gamma_i = \delta_i^2\,\Omega$, with a common covariance Ω and study-specific scaling constants δ_i.

By (8), this special form of Γ_i obtains if for every level l the covariance $\Omega^{(l)}$ either equals the null matrix or is identical with Ω. Difficulties remain even with this simplification because the numbers $n_a, \ldots, n_b$ characteristic of the individual study designs generally are unknown in meta-analysis. However, assuming that random effects accrue at one level at least, one can give upper and lower estimates for the scaling constants that are independent of the unknowns $n_a, \ldots, n_b$. The lower estimate applies if random effects occur only at the (lowest) bit level; then δ_i^2 is not smaller than N_i^{-1} for every i. With random effects at the (top) agent level, δ_i^2 cannot exceed the number of levels l for which $\Omega^{(l)} \neq 0$, hence stays uniformly bounded. This discussion, and the impossibility of estimating all scaling constants from the data, prompts us to resort to the following

pragmatic strategy for choosing the δ_is:
Try both extremes, i.e., either set $\delta_i^2 = 1$ for all i, or set $\delta_i^2 = N_i^{-1}$ for all i, and continue with the choice giving the better fit.

Let us remark that the random effects modeling that is customary in the context of single measurements is often adopted also in meta-analysis

[5] Intuitively, this is easily understood considering that low level random effects are averaged out (much) more heavily than high level random effects, where the averaging involves relatively few items.

without modification (*e.g.*, DerSimonian and Laird 1986, p. 183). It is tantamount to the choice $\delta_i^2 = 1$, which may imply an inappropriate scaling of the random effects.

The random coefficients regression (RCR) model finally emerging from the above may now be stated as follows.

(R1) Z-scores Z_i are independent across experiments and normally distributed as $\mathcal{N}(\mu_i\,,\,\sigma_i^2)$ where

$$\mu_i \;=\; \gamma_i\, x_i\, \beta\,, \qquad \sigma_i^2 \;=\; 1 \;+\; \gamma_i^2\, \delta_i^2\, x_i \Omega x_i^T. \tag{9}$$

(R2) The unknown parameters of the model are the $p \times 1$ vector of regression coefficients β and the $p \times p$ covariance matrix Ω.

(R3) Known parameters are the $1 \times p$ covariate vectors x_i, the informativity quantities γ_i defined in (6), and the scaling constants δ_i (chosen according to the pragmatic strategy).

Each single quantity appearing in the RCR model can be interpreted in terms of its contribution to the data variability expected under the model. The term $x_i\beta$ accounts for systematic effects induced by the variation of the covariates across studies. The term $\delta_i^2\, x_i \Omega x_i^T$ covers unsystematic extra variability modeled by random effects, with scaling constants δ_i adjusting the degree to which random effects are averaged out due to data aggregation. The magnitude of these two effects increases with the scale of the study, which is measured by the informativity quantities γ_i. Finally, the constant 1 appearing in σ_i^2 represents the (normalized) intrinsic randomness contributed by the REG's bit production mechanism.

3.4 Submodels, and Other Miscellanea

We assume throughout that $x_{i1} = 1$ for every i, so that only the part $\breve{x}_i = (x_{i2}, \ldots, x_{ip})$ of the vector $x_i = (1, \breve{x}_i)$ encodes proper covariate information. Accordingly, the components x_{ik} of x_i with $k \geq 2$ are sometimes called *proper covariates* (and x_{i1} *improper*). This assumption, quite common in regression analysis, guarantees that the linear compound

$$x_i\beta \;=\; \sum\nolimits_{k=1}^{p} x_{ik}\, \beta_k \;=\; \beta_1 + \sum\nolimits_{k=2}^{p} x_{ik}\, \beta_k \tag{10}$$

contains a component that is functionally independent of the proper covariates and stays constant across studies. Hence, it allows to model a constant effect representing the experimental main effect.[6]

[6]The "main effect" is related to, but generally not identical with the constant effect parameter β_1. The latter has to be adjusted for the contribution of the proper covariates to the constant effect in order to yield the net main effect.

3.4.1 Submodels

Submodels of the RCR model give rise to various null hypotheses of interest. These can be stated in terms of conditions imposed on the RCR parameters β and Ω.

A first class of submodels is characterized by the condition $\Omega = 0$ tantamount to the non-existence of random effects. These models will be referred to as *fixed coefficient regression (FCR)* models. Any extra variability beyond the REG's intrinsic randomness is ascribed to systematic effects connected with the variation of the covariates across studies within an FCR model. Restrictions of the β parameter generate further submodels. E.g., $\beta = 0$ gives the *'no effect'* FCR model, and $\beta_k = 0$ for $k = 2, \ldots, p$ gives the *'no covariate effect'* model. The latter is also called the *'constant effect'* model since the bit-level effects then are globally constant, i.e., $\epsilon_{ij} = \beta_1$ for all ij; cf. (B2) and recall that $x_{i1} = 1$.

A second class of submodels is obtained by restricting β as in the FCR models but allowing for random effects. For example, the *'constant effect'* RCR model leaves Ω arbitrary and assumes $\beta_k = 0$ for $k = 2, \ldots, p$. This rules out any systematic variation of the expected values with the proper covariates. However, random variability connected with *all* covariates *is* permitted by the model.

If the model is to admit unstructured random variability only, unconnected to the proper covariates, this can be achieved by supposing that $\Omega_{kl} = 0$ for all matrix elements except Ω_{11}. Then $x_i \Omega x_i^T = \Omega_{11}$ is constant, and we have the standard random effects model which stipulates additive random shocks that are uncorrelated with any measured quantity. Further models can be produced by combining restrictions that apply to both β and Ω. Some limitation being necessary, *we shall restrict attention to the two extreme cases where Ω either is unrestricted (RCR models) or vanishes identically (FCR models).*

3.4.2 Aside on Effect Sizes and Scales

The 'constant effect' model forms the basis, if only implicitly, of the common usage to estimate "the" effect size as the difference between the empirical and the theoretical relative frequencies of a hit. Intuitively underlying the model is the notion that "the" experimental effect is due to something like a law of nature acting uniformly, irrespective of the special circumstances. Its usual implementation (in terms of the Bernoulli paradigm) is tantamount to the assumption that under the experimental conditions the by-chance hit probabilities $p_{i,0}$ are all shifted by the same fixed amount δp, representing the effect size, to hit probabilities $p_{ij} = p_{i,0} + \delta p$. However, this assumption is less natural than it may seem if the $p_{i,0}$s are not identical across experiments. For a shift by $\delta p = 0.05$, say, represents, in a sense, a stronger effect if $p_{i,0} = 0.1$ than if $p_{i,0} = 0.5$. This

factor is not reflected by the implied linear scale, where the bit-level effects are constant no matter what $p_{i,0}$, $\epsilon_{ij} = \delta p$ for all ij; cf. (2). Other than in linear scale, the bit-level effect in the log-odds scale that is necessary to achieve a given probability shift of δp *increases* as the underlying $p_{i,0}$ decreases from 0.5 to 0,[7] in correspondence with intuitive expectations.

3.4.3 Coding of Hits and Choice of Scale

In the meta-analysis of R&N z-scores are not always computed according to Eq. (1). Specifically, in studies where the intentional prescription is to *lower* the number of 1s, the sign of the z-score is reversed in order to give a uniform meaning to deviations in different directions. At the bit level the sign change corresponds to counting 0s rather than 1s as hits: under the substitution $Y \to 1 - Y$, $p \to 1 - p$ z-scores Z are sent to $-Z$; cf. (1). However, our models describing alternative behavior, *e.g.* the RCR model and the gBp, generally do not enjoy such an equivariance property. They do so only if the scale function is *symmetric* in the sense that the following holds for any $0 < p < 1$: if $\theta_{0,1}$ are such that $s(\theta_0) = p$, $s(\theta_1) = 1-p$, then $s'(\theta_0) = s'(\theta_1)$. We shall subsequently work with the log-odds scale, because a) its scale function is symmetric, b) it reflects the different meaning of "effect size δp" at small and moderate by-chance hit probabilities p_0, c) it complies with a proposal by Schmidt (1975) for roughly the same purpose as here (see next subsection). With this choice of s the square of the informativity quantities (6) reduces to a binomial variance,

$$\gamma_i^2 \;=\; N_i \, p_{i,0} \, (1 - p_{i,0}) \, . \tag{11}$$

3.4.4 Connection with a Model by Schmidt

Schmidt (1975) proposed a statistical model for the output of an REG that may be regarded as a special case of the gBp and its further specifications. For instance, his product rule for different "world histories" is related to the independence assumption (B2), and his "psi axiom" implies that effects are measured in the log-odds scale. Parts of his discussion of the "divergence problem" might be understood as dealing with random effects. However, Schmidt appears to assume that his effect parameter Θ is (globally) constant, so in our terminology he proposes a constant (fixed) effect model. This model, sometimes extended to an FCR model in order to include covariates, seems to underlie all meta-analyses in mind-matter research. Let us emphasize that despite the fact that Schmidt's model is a special case of the RCR model (and its extension in Section 5), we do

[7] This is seen most easily by solving the approximate equation $\delta p = s(\theta + \epsilon) - s(\theta) \simeq s'(\theta)\,\epsilon$ for ϵ, $\epsilon \simeq \delta p \,/ s'(\theta)$, and noting that $s'(\theta) = p(1-p)$ becomes small for $p = s(\theta))$ close to 0 or 1.

not subscribe to the interpretations put forward by Schmidt. We prefer to stay entirely at the phenomenological, descriptive level.

3.4.5 Likelihood and Residuals

Model fitting and statistical inferences will rely on likelihood methods. We parametrize the covariance matrix Ω using the Cholesky factorization $\Omega = \phi\phi^T = \Omega(\phi)$, where ϕ is a $p \times p$ lower diagonal matrix re-arranged here as a vector in $p(p+1)/2$-dimensional space (and made unique by some rule fixing the sign). The parameters of the RCR model then are β and ϕ, and the log-likelihood given observed z-scores $Z_i = z_i$ is

$$\Lambda_r(\beta, \phi) \;=\; \sum_{i=1}^{n} \log\, f_{\theta_i}(z_i) \quad \text{where} \quad f_{\theta_i}(z) = \varphi\left(\frac{z-\mu_i}{\sigma_i}\right)\frac{1}{\sigma_i} \tag{12}$$

denotes the density of Z_i, $\varphi(t) = e^{-t^2/2}/\sqrt{2\pi}$ the standard normal density, and $\theta_i = (\mu_i, \sigma_i^2)$ with $\mu_i = \mu_i(\beta)$, $\sigma_i^2 = \sigma_i^2(\phi)$ from (9). Setting $R_i = (z_i - \mu_i)/\sigma_i$ and ignoring unimportant constants, (12) becomes

$$\Lambda_r(\beta, \phi) \;=\; -\frac{1}{2}\sum_{i=1}^{n} R_i^2 - \frac{1}{2}\sum_{i=1}^{n} \log \sigma_i^2. \tag{13}$$

Note that μ_i and σ_i, hence R_i, depend on β and ϕ. Maximum likelihood estimates (MLEs) $\widehat{\beta}$, $\widehat{\phi}$ are computed by numerical maximization of (13) over (β, ϕ); see Appendix C. Fitted z-scores and variance estimates are then obtained by "plug-in", $\widehat{z}_i = \widehat{\mu}_i = \mu_i(\widehat{\beta}) = \gamma_i x_i \widehat{\beta}$, $\widehat{\sigma}_i^2 = \sigma_i^2(\widehat{\phi}) = 1 + \gamma_i^2 \delta_i^2\, x_i \widehat{\phi}\widehat{\phi}^T x_i^T$. The Fisher information matrix, whose inverse is the asymptotic covariance matrix of the MLE, is block-diagonal according to the dimensions of β and ϕ. Hence, β and ϕ are orthogonal parameters in the sense of Cox and Reid (1987), which is a desirable stability property. In particular, the estimates $\widehat{\beta}$, $\widehat{\phi}$ are asymptotically uncorrelated.

Residuals provide a standardized comparison of observed and fitted z-scores. For an RCR model with MLEs $\widehat{\beta}$, $\widehat{\phi}$ they are defined as

$$r_i \;=\; R_i(\widehat{\beta}, \widehat{\phi}) \;=\; (z_i - \gamma_i x_i \widehat{\beta})/\widehat{\sigma}_i \quad (i = 1, \ldots, n). \tag{14}$$

Under our assumptions $\widehat{\sigma}_i$ is a consistent estimate of σ_i (meaning that the ratio $\widehat{\sigma}_i/\sigma_i$ tends to 1 in probability), and r_i is standard normally distributed to a first approximation if the assumed model is valid. These facts form the basis for various, mostly graphical, techniques of model checking. We use two such techniques in the sequel. Finally, note that in the special case of an FCR model, where $\sigma_i = 1$, the second sum in (13) vanishes and Λ_r reduces to the FCR log-likelihood, $\Lambda_r(\beta, 0) = \Lambda_f(\beta)$. The residuals reduce to $r_i = z_i - \gamma_i x_i \widehat{\beta}$, and $\widehat{\beta}$ reduces to the ordinary least squares estimate in this case.

4. RCR-Analysis of the Data of Radin & Nelson

Setting up a regression model for the Radin & Nelson data requires specifying how the available additional information is encoded in terms of the covariates. The details of the construction are given in Appendix B, along with an account of the omission of a fraction of deficient data that are not suitable for our analyses. We also exclude a non-deficient study with a huge number N_i of bits and a z-score of 2, for reasons given in Appendix B.[8] Thus everything in the following refers to those $n = 516$ studies actually entering the analysis.

According to our general strategy the fit of a model has to be checked in the first place. The basic tools are residual plots and simulation. Inference pertaining to model parameters is dealt with afterwards.

4.1 Model Checking

4.1.1 FCR Model

Under an FCR model with "true" parameter β, the z-score z_i has the distribution $\mathcal{N}(\gamma_i x_i \beta, 1)$, and the corresponding residual r_i is approximately distributed as $\mathcal{N}(0, 1)$. The fit of the full FCR model (including all covariates) is summarized graphically in Figure 1.

The data underlying the three plots in the first *column* of Figure 1 (and similar figures to follow) are the original z-scores from the R&N data. Underlying the second and third columns are artificial z-scores z_i^* obtained by simulation from the estimated FCR model, i.e., by generating (two sets of) independent variables z_i^* distributed as $\mathcal{N}(\gamma_i x_i \widehat{\beta}, 1)$ $(1 \leq i \leq n)$. Generating a number of such simulated data sets – space limitations prohibit to present more than two – is helpful to get an idea of how the plots should look like if the assumed model (here FCR) was "true", particularly in regard to the sampling fluctuation to be expected under the model.

In the top *row*, the z-scores z_i and the fitted z-scores $\widehat{z}_i = \gamma_i x_i \widehat{\beta}$ are plotted against γ_i,[9] for the original and two simulated data sets. We note that if instead of the full FCR model the 'constant effect' model was fitted, the fitted values would lie on a straight line through the origin, a positive slope indicating a deviation in the envisaged direction. The scattering of the fitted z-scores in Figure 1 is due to the covariates included in the full model. In the second row scatter plots of the residuals r_i versus γ_i are shown. The two horizontal lines at height ± 2, corresponding to two standard deviations, are added for a rough gauging. The bottom row shows *(normal) probability plots* of the residuals. In such a plot

[8] In Appendix B, we shall demonstrate that the study omissions do not affect the main results of this paper.

[9] The scale on the abscissa is chosen so as to make the γ_is moderately sized (by division through $\max_i \gamma_i/8$).

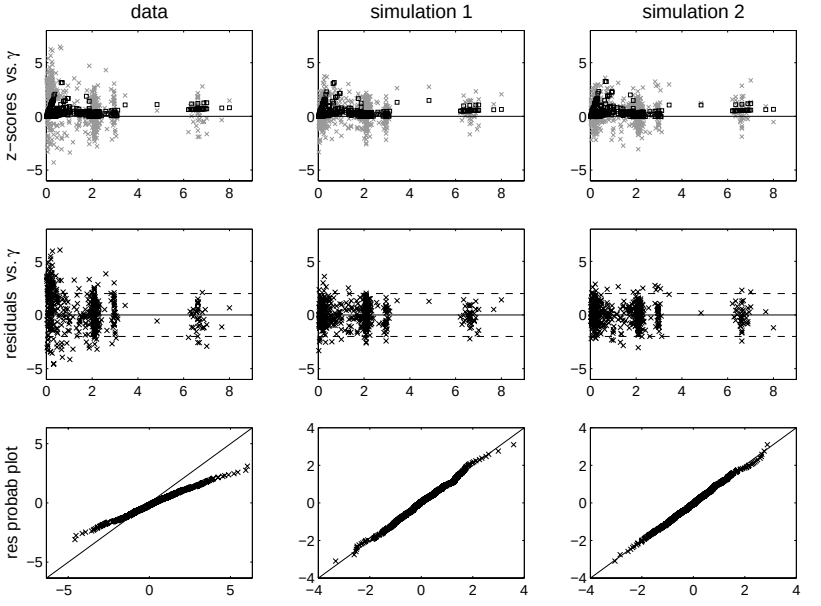

Figure 1: Fit of FCR model. Top row: observed (grey crosses) and fitted (black squares) z-scores vs. informativity quantities (γ_is). Middle row: residuals vs. γ_is. Bottom row: residual probability plots. First column: original data; second and third columns: two simulated data sets.

the residuals are ordered according to size and plotted (on the abscissa) against the respective expected values of a likewise ordered sample of n $\mathcal{N}(0,1)$-distributed random variables. If the assumed model happens to be "true", or is sufficiently close to the "true" one, the ordered residuals scatter along the diagonal $y = x$, which is also drawn for reference. Some deviation from the diagonal is always to be expected particularly in the tails due to sampling fluctuation. Probability plots with more pronounced deviations, exhibiting curvature or points lying consistently above or below the diagonal, indicate a bad fit and suggest that the assumed model is inappropriate. Again, comparison of the original with simulated data provides an important clue.

The plots in the top and middle row show that the original data are much more variable than the simulated data. This is confirmed by the residual probability plots (last row). In fact, the classical goodness-of-fit chi-squared test is highly significant, with p-value $p \doteq 0$ to computing accuracy, and we conclude that there is substantially more variability in the R&N data than can be explained by intrinsic (Bernoulli) randomness or by systematic variation induced by the covariates, or by both together.[10]

[10]This is *not* evident already from Figure 2 in R&N. For *if* there is a nonzero effect

This also rules out the 'constant effect' model which explains still less variability than the full FCR model. These findings suggest to look for other sources of variability.

4.1.2 RCR Model

According to the *pragmatic strategy* of Section 3.3 there are two options for choosing the scaling constants, $\delta_i = 1$ and $\delta_i^2 = N_i^{-1}$. The results for the respective RCR model are presented in Figures 2 and 3.[11] The fit is clearly better for the second choice, $\delta_i^2 = N_i^{-1}$. This suggests that if there is any kind of parameter variation, it is more likely to be connected with fluctuations of the Bernoulli probabilities at the bit level than with parameter variation at some higher level (across agents, say). Let us repeat that straightforward random effects modeling, disregarding the special data structure in meta-analysis, would lead one to choose $\delta_i = 1$. However, this would imply a higher extra variability than is actually observed. From now on we proceed with scaling constants $\delta_i^2 = N_i^{-1}$.

A simple chi-squared type goodness-of-fit test as for FCR models is not available for RCR models. Instead, internal consistency checks need to be carried out. As seen above, the residual plots along with simulation under the assumed model provide useful information in regard to goodness-of-fit, and we shall rely on these tools subsequently. A closer examination of the probability plot of the residuals of the original data in Figure 3 reveals a particular feature: most data points lie marginally but consistently below the reference line, even in the central region where the sampling fluctuation is smaller than in the tails. They seem to scatter around a slightly shifted line instead of the reference line. We shall return to this observation later on.

4.2 Experimental Effects

When speaking of experimental effects we usually refer to hypotheses concerning the parameter β. Specifically, we consider likelihood ratio tests between three hypotheses relating to β: the 'no effect' hypothesis ($\beta = 0$), the 'constant effect' (or 'no covariate effect') hypothesis ($\beta_k = 0$ for $k \geq 2$), and the full model (β unrestricted); cf. Section 3.4.1. The hypotheses give rise to three tests each, 'no effect' versus either 'constant effect' or the full model, and 'constant effect' versus the full model, with 1, 6, and 5 degrees of freedom, respectively. (The regression model (10) has $p = 6$ in the case of the R&N data; cf. Appendix B.)

---

then even the *expected* z-scores will show variability, depending on the variation of the scaled covariates $\gamma_i x_i$ (or of the γ_is in case of the 'constant effect' model) in design space. Detecting extra variation *beyond* that explained by the FCR model requires looking at the residuals rather than at the z-scores.

[11] Simulation under the RCR model is performed by generating (sets of) independent variables z_i^* distributed as $\mathcal{N}(\gamma_i x_i \widehat{\beta}, 1 + \gamma_i^2 \delta_i^2 \, x_i \widehat{\phi\phi}^T x_i^T)$ $(1 \leq i \leq n)$.

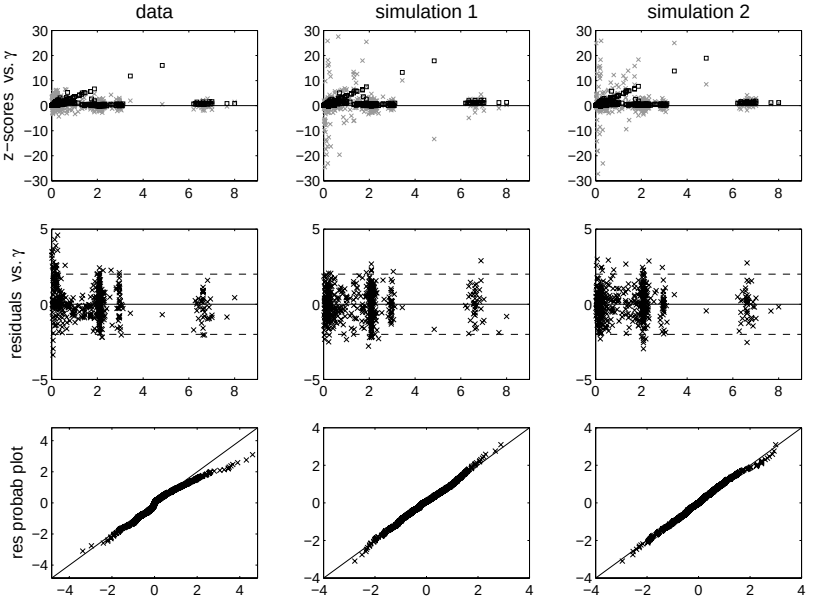

Figure 2: Fit of RCR model with $\delta_i^2 = 1$; otherwise as Figure 1.

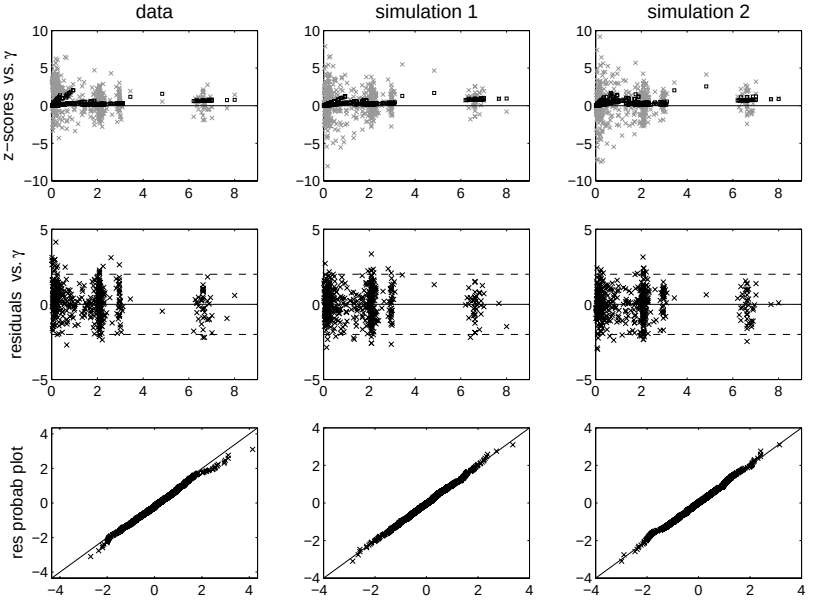

Figure 3: Fit of RCR model with $\delta_i^2 = N_i^{-1}$; otherwise as Figure 1.

The results may be sketched as follows; details are given in Sec. 6.3. Among the six tests (three each for both model types) all but one are highly significant: only the RCR test of the 'constant effect' hypothesis versus the full model is marginally significant, with a p-value a bit larger than 0.05. Altogether, this suggests the presence of a significant experimental main effect in the form of a nonzero mean shift constant across studies which is, perhaps, modified somewhat from case to case by the influence of the covariates. We will re-examine this conclusion from the perspective of corrections meta-analysis to be developed in the next section.

5. Allowing for Selection Bias: Corrections Meta-Analysis

5.1 The Selection Model

Biased samples can arise in many ways. In meta-analysis the focus is mainly on publication bias (cf. Sec. 5.3), meaning that (statistically) non-significant results are less likely to get published than significant ones. If so, the data collected for a meta-analysis is incomplete and fails to constitute a representative sample from the body of *all* studies conducted in the area under consideration. Let n_{tot} denote the number of all studies, published and unpublished, and let J_i be an indicator variable assuming the value 1 if the i-th study is included in the meta-analysis, and 0 if it is missing. A corresponding extension of the RCR model involves the following assumptions.

(S1) The pairs (Z_i, J_i), $i = 1, \ldots, n_{tot}$, are statistically independent.

(S2) Z_i is distributed as $\mathcal{N}(\mu_i, \sigma_i^2)$, where μ_i, σ_i^2 are given by (9).

(S3) Conditionally on $Z_i = z_i$, J_i is distributed as $Bernoulli(\kappa(z_i))$, i.e.,

$$P\,[J_i = 1 \,|\, Z_i = z_i] \;=\; 1 - P\,[J_i = 0 \,|\, Z_i = z_i] \;=\; \kappa(z_i). \qquad (15)$$

Intuitively, (S1)–(S3) say that the i-th study appears in the meta-analysis with a probability depending on the z-score Z_i, as if its (non-)inclusion was decided by flipping an accordingly biased coin. Let $\mathcal{J} = \{i \le n_{tot} : J_i = 1\}$ denote the subset of the studies included in the meta-analysis. The conditional likelihood of the available data $Z_i,\ i \in \mathcal{J}$ (conditionally on $\mathcal{J}$), is of the form

$$L \;=\; \prod_{i \in \mathcal{J}} \frac{f_{\theta_i}(z_i)\ \kappa(z_i)}{\int f_{\theta_i}(z)\ \kappa(z)\, dz}\ , \qquad (16)$$

with f_{θ_i}s as defined in (12). The inclusion probabilities $\kappa(z)$ may be modeled as a parametrized family of functions with a plausible shape. Here we leave the shape arbitrary and work with step functions, at the price of abandoning smoothness. Beforehand, note that L in (16) is invariant under multiplication of κ by a constant, so that κ needs to be specified up to a common factor only and may be seen as an inclusion *tendency,* which assumes arbitrary positive values.[12]

(S4) Let $I_k = (c_{k-1}, c_k]$ be an interval partition of the real line generated by the points $-\infty = c_0 < c_1 < \ldots < c_{K+1} = +\infty$. Then κ is a step function of the form

$$\kappa(z) \;=\; e^{\alpha_k}, \;\; z \in I_k \quad (k = 1, \ldots, K+1), \tag{17}$$

where $\alpha_{K+1} = 0$ and the other α_ks are arbitrary real parameters. The interval endpoints $c_1, \ldots, c_K$ are chosen as follows ($K = 9$),

$$-2.33 \;\; -1.96 \;\; -1.65 \;\; -1 \quad 0 \quad 1 \quad 1.65 \;\; 1.96 \;\; 2.33. \tag{18}$$

The c_ks are familiar percentiles of the normal distribution. Of course, the finer the partition the more flexible is the model. On the other hand, the number of unknown parameters has to be kept moderate, and the present choice with $K = 9$ additional parameters to be estimated appears as a reasonable compromise. The parametrization in (17) avoids positivity constraints such as $w_k\,(= e^{\alpha_k}) > 0$, and the condition $\alpha_{K+1} = 0$ removes the scale indeterminacy of κ.

The selection model defined by (S1)–(S4) is put on top of the RCR model, and we call it the *RCR selection,* or *RCRS* model. It is essentially the same as the one proposed by Hedges (1992), whose model did not include covariates, however. See also Dear and Begg (1992).

5.2 Corrections Meta-Analysis

The RCRS model covers departures from the null hypothesis that may be due to both experimental and selection effects. Is it possible to distinguish both types of effects on the basis of the model? In particular, can one make valid statistical inferences about the residual experimental effect when adjustment is made for possible selection effects? And, vice versa, about selection effects adjusted for possible experimental effects? In any case the idea is (to try) to correct the effect under consideration for biases possibly introduced by the other effect; hence the name *corrections meta-analysis* (Copas 1999).

[12] This corresponds to the obvious fact that it is impossible to estimate the absolute magnitude of the inclusion probabilities on the basis of the available data.

The statistical analyses are again likelihood-based. We may assume that $\mathcal{J} = \{1, \ldots, n\}$, where n is the number of studies included in the meta-analysis. By (16), (17), and with the interval partition defined by the cutpoints (18), the log-likelihood Λ_s of the RCRS model splits into the sum of the RCR log-likelihood

$$\Lambda_r(\beta, \phi) = -\frac{1}{2} \sum_{i=1}^{n} R_i(\beta, \phi)^2 - \frac{1}{2} \sum_{i=1}^{n} \log \sigma_i(\phi)^2$$

(cf. (13)) and a multinomial part,

$$\Lambda_s(\beta, \phi, \alpha) \;=\; \Lambda_r(\beta, \phi) + \sum_{k=1}^{K} \nu_k \, \alpha_k - \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K+1} e^{\alpha_k} \, \pi_k(\theta_i) \right), \quad (19)$$

where $\alpha = (\alpha_1, \ldots, \alpha_K)$ is the vector of the K log weights $\alpha_k = \log w_k$ that vary freely over the real numbers (recall that $\alpha_{K+1} = 0$), and

$$\begin{aligned} \nu_k &= \#\{i \leq n : z_i \in I_k\} \\ \pi_k(\theta_i) &= Pr_{\theta_i}(Z_i \in I_k) \;=\; \int_{I_k} f_{\theta_i}(z) \, dz \end{aligned}$$

are, respectively, the number of z-scores, and the probability (under the RCR model) of z-score Z_i falling into the interval I_k. Note that Λ_s reduces to the RCR log-likelihood Λ_r if all $\alpha_k = 0$, i.e., if there is no selection: $\Lambda_s(\beta, \phi, 0) = \Lambda_r(\beta, \phi)$. MLEs $\widehat{\alpha}$, $\widehat{\beta}$, $\widehat{\phi}$ are obtained by (numerical) maximization of the log-likelihood (19), with some parameter components set to zero if a submodel is assumed. In particular, the 'constant effect' and 'no effect' RCRS (sub-)models are defined by the restrictions $\beta_2 = \ldots = \beta_6 = 0$ and $\beta = 0$, respectively.

5.3 Important Note

Mathematically, the notion "selection effect" refers to the (multinomial) term that, when added to the RCR log-likelihood Λ_r, yields the log-likelihood Λ_s of the RCRS model; cf. (19). The label "selection effect" attached to this mathematical object derives from the toy model for publication bias by way of selective reporting described in Section 5.1. Thus, considered from inside the assumed (RCRS) model, this component refers to a data selection mechanism (one of many possible) that might underlie what is generally referred to as publication bias. It is parametrized by parameter α, just as parameters β and ϕ parametrize the "experimental" and "random" effects, respectively, within the RCR(S) model.

By contrast, the "selection effect" as estimated from empirical data is to be understood in a much wider sense. It comprises the whole bundle

of "effects" which yield (approximately) the same estimate of parameter α when "projected" onto the RCRS model. This bundle contains much more than what is covered by the RCRS model. Thus the selection effect detected by the RCRS model may have little or nothing to do with the selection mechanism that gave rise to its name; it may reflect any kind of deviation from the RCR model that induces similarly looking effects at the z-score level.[13] Nevertheless, all deviations from the RCR model detected by the RCRS model will be referred to as selection effects, no matter what their origin is.

In regard to this usage it has to be pointed out that *any* statistical analysis has to concede that an estimated "effect" might have an origin different from what it is ascribed to within the fitted statistical model. This is neither a special weakness of the RCRS model nor is it, within the RCRS model, restricted to the selection component: the estimated "experimental" and "random" effects, too, may be due to other "effects" than those they stand for within the RCR model. To some extent the risk of misattributing effects can be diminished by model checking. This is why model checking is important and is emphasized so much in this paper. However, that risk can never be entirely excluded. Again, despite all these necessary qualifications it is still possible that estimated effects are essentially due to what their denomination suggests.

These background considerations are decisive for a proper understanding of the paper and have always to be borne in mind.

6. RCRS-Analysis of the Data of Radin & Nelson

6.1 Model Checking

As above, checks are based on (corrected) residuals and simulation. The residuals corrected for selection are defined as $r_{i,s} = (z_i - \widehat{\mu}_{i,s})/\widehat{\sigma}_{i,s}$, where $\widehat{\mu}_{i,s}$, $\widehat{\sigma}_{i,s}$ denote the MLEs of the expectation $\mu_{i,s}$ and standard deviation $\sigma_{i,s}$ of Z_i under the RCRS model. Explicit expressions are given in Appendix C. Other than in the RCR case the corrected residuals are *not* approximately $\mathcal{N}(0,1)$-distributed under the RCRS model if selection is present, so in the residual probability plot some deviation from the diagonal is expected even if the RCRS model is "true" and sampling fluctuation is ignored. However, the residuals capture at least the most important model features encoded in the first two moments, and for proper gauging we rely on simulation anyway. Details about the procedure implementing simulation under the RCRS model are given in Appendix C.

[13] One might think here of cases where study termination does not follow a strict plan fixed in advance. Also, it may not always be clear what should count as a study. Splitting or combining studies might have effects that are not easily distinguishable from selection effects. Another, very different possibility is mentioned in footnote 17.

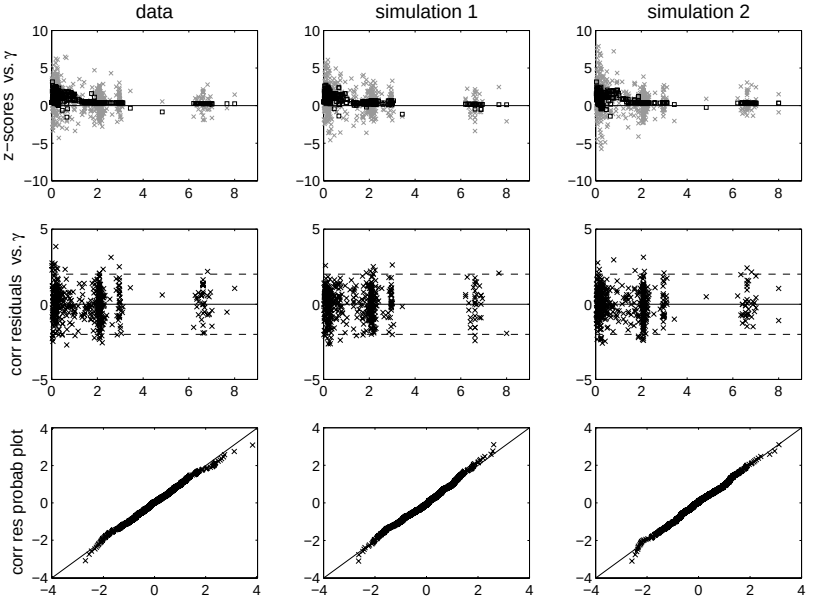

Figure 4: Fit of RCRS model; otherwise as Figure 1.

Figure 4 is organized as Figures 2 and 3. The residual probability plot for the observed data does not exhibit a shift from the diagonal as in case of the RCR model. The minor departures in the tails may be due to sampling fluctuation and the non-normality mentioned above. Comparison with the simulated data sets reveals no conspicuous features. Thus it seems that the RCRS model fits the data satisfactorily.

6.2 Inference Concerning Selection Effects

In the publication bias interpretation of the RCRS model, the parameters $w_k = e^{\alpha_k}$ represent the tendency of a z-score falling into interval I_k to receive publication. The w_ks are unique up to a common factor only, and for presentation it is convenient to scale them to average 1. Hence, the step function assuming the value

$$t_k \;=\; \frac{(K+1)\,\hat{w}_k}{\sum_l \hat{w}_l} \tag{20}$$

on the interval I_k is referred to as the *tendency-to-report profile.* It reduces to the constant 1 if there is no selection bias. Otherwise it indicates how the tendency-to-report depends on the value of the z-score. For the R&N

data, and interval cutpoints given by (18), the MLEs $\hat{t}_k$ of the tendencies-to-report t_k are:

$$0.31 \quad 0.52 \quad 0.44 \quad 0.43 \quad 0.61 \quad 0.87 \quad 0.97 \quad 2.21 \quad 1.28 \quad 2.35 \tag{21}$$

The $\hat{t}_k$s are roughly increasing, with a dip in the interval $[1.96, 2.33]$, and a report tendency for z-scores larger than 2.33 (smaller than -2.33) that is more than twice (less than one third) times as large as expected under 'no selection' conditions. In order to check whether these outcomes could be due to chance a formal test is required. The likelihood ratio test statistic for testing the null-hypothesis of 'no selection' within the (full) RCRS model – or: of the RCR model against the RCRS model – has the form

$$LR_{sel} \;=\; 2\left(\sup_{\beta,\phi,\alpha} \Lambda_s(\beta,\phi,\alpha) - \sup_{\beta,\phi} \Lambda_s(\beta,\phi,0)\right). \tag{22}$$

According to standard asymptotic theory, LR_{sel} is approximately χ^2_9-distributed (χ^2 with $K = 9$ degrees of freedom) under the null-hypothesis. Here we have $LR_{sel} \doteq 78.3$, corresponding to a (nominal) p-value of less than 10^{-12}. The p-values for the 'constant effect' and 'no effect' RCRS models are still smaller. All this speaks against the 'no selection' hypothesis. In Appendix B we show that these results cannot be explained by the omission of a certain fraction of studies (cf. first paragraph of Sec. 4).

6.3 Testing for Experimental Effects

As above we consider three tests concerning the β parameters: 'no effect' versus either 'constant effect' or the full model, and 'constant effect' versus the full model, now including the possibility of selection, however. The likelihood ratio test statistic for the first test, 'no effect' vs. 'constant effect', is[14]

$$LR_{nc,\,s} \;=\; 2\left(\sup_{C} \Lambda_s(\beta,\phi,\alpha) - \sup_{\phi,\alpha} \Lambda_s(0,\phi,\alpha)\right) \tag{23}$$

where $\sup_C$ means that the maximization extends over the set C of all triplets β, ϕ, α subject to the condition $\beta_2 = \ldots = \beta_6 = 0$, which characterizes the 'constant effect' model. The other two test statistics, $LR_{nf,\,s}$ and $LR_{cf,\,s}$, are defined analogously. According to standard asymptotic theory they are approximately χ^2-distributed with 1, 6, and 5 degrees of freedom under the respective null-hypothesis. Table 1 assembles the test results, including those for the FCR and RCR models indicated earlier in Section 4.

[14] The index s in $LR_{nc,\,s}$ indicates that the 'no vs. const. effect' test is carried out assuming the RCRS model. If the test is carried out assuming the RCR or FCR model, the log-likelihoods and maximizations in (23) have to be modified appropriately, and the index r or f is used instead of s.

Table 1: p-values of tests concerning experimental effects (β-parameters)

model	'no eff.' / 'full'	'no eff.' / 'const. eff.'	'const. eff.' / 'full'
FCR	~ 0	$\sim 10^{-10}$	~ 0
RCR	$8.860 \cdot 10^{-6}$	$1.975 \cdot 10^{-6}$	0.0563
RCRS	0.2484	0.6666	0.1750

The table suggests two conclusions. First, all tests pertaining to β become non-significant if adjustment for possible selection is made. Secondly, the outcomes are discordant across models. The apparent inconsistencies raise questions in regard to the validity and meaning of the test results. These are tackled in the next section.

7. Does Corrections Meta-Analysis Work?

The validity and efficiency properties of the tests are studied through simulation in diverse regimes. We generated 5000 sets of simulated z-scores obeying the RCRS model as described in Appendix C, under each of four parameter configurations. Let $\hat{\beta}, \hat{\phi}, \hat{\alpha}$ denote the MLEs of the full RCRS model parameters calculated for the original z-scores. The four configurations correspond to choosing the parameters (β, ϕ, α) underlying the simulations as follows:

(e1s1) $(\beta, \phi, \alpha) = (\hat{\beta}, \hat{\phi}, \hat{\alpha})$
(e0s1) $(\beta, \phi, \alpha) = (0, \hat{\phi}, \hat{\alpha})$
(e1s0) $(\beta, \phi, \alpha) = (\hat{\beta}, \hat{\phi}, 0)$
(e0s0) $(\beta, \phi, \alpha) = (0, \hat{\phi}, 0)$

The nomenclature s1/s0 (e1/e0) indicates that a selection (experimental) effect is included/not included. For example, the parameter configuration (e1s1) involves both selection and experimental effects; configuration (e0s0) involves neither of the two; and so on. The present non-null choice of the parameters α, β as MLEs implies that the configuration (e1s1) constitutes the best representation of the actual situation (i.e., the R&N data) in terms of the RCRS model, hence is primarily relevant here. For a broader picture alternative parameter choices have to be considered, too. Configurations with $\alpha = 0$, (e1s0) and (e0s0), involve no selection and imply simulation under the RCR model, effectively. The simulation procedure described in Appendix C then becomes unnecessarily expensive and is replaced by the obvious one; cf. footnote 11.

7.1 Assessment of Inferences Concerning Selection Effects

As seen in Sec. 6.2, the null-hypothesis of 'no selection' for the R&N data is rejected with high significance (i.e., small p-value) by the likelihood

ratio test. Is this because the test misinterprets a shift of the z-score means to higher values, connected with an experimental effect, as the result of biased selection? Technically, the question is whether the test is *valid:* does the test statistic (22) follow the nominal null distribution χ^2_9 if $\alpha = 0$ (no selection) *also if* β *is nonzero* (i.e., if an experimental effect is present)? Or equivalently, and more conveniently for graphical representation: are the associated p-values uniformly distributed over the unit interval under these conditions?

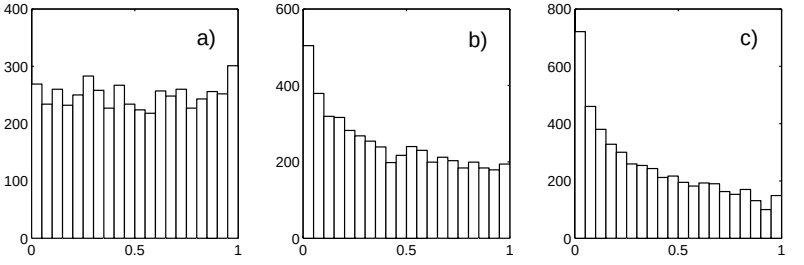

Figure 5: Testing the 'no selection' hypothesis. Histograms of the p-value distribution of the likelihood ratio test for three RCRS models, simulated under configuration (e1s0). a) full model; b) constant effect; c) no effect.

Figure 5a is a histogram of the 5000 p-values simulated under configuration (e1s0). Visually, it appears to be consistent with the nominal uniform distribution $\mathcal{U}$. The chi-squared goodness-of-fit test based on the 20 bin counts has a p-value of 0.022. Given the large sample size of 5000 this means that the p-value distribution fits the uniform distribution reasonably well, though not perfectly. The empirical frequency of simulated p-values less than 0.05 is 269, corresponding to an error of the first kind of $269/5000 = 0.0538$ in good agreement with the nominal 5% level. We conclude that for the R&N data the likelihood ratio test of the 'no selection' null-hypothesis is essentially unbiased. Figures 5b and 5c show corresponding p-value histograms for two other tests for selection, where the likelihood ratio statistic is calculated assuming a 'constant effect' or 'no effect' RCRS model instead of the full RCRS model. Apparently, there is some systematic deviation from $\mathcal{U}$ in both cases.

For a further validity check let us examine the tendency-to-report profiles [TRP; cf. (20)] fitted to the simulated data sets. Since there is no selection under configuration (e1s0), these should equal the constant 1 up to sampling fluctuation. In the top row of Figure 6 ten profiles are presented for each of the three RCRS models (as polygonal lines rather than as step functions) that are randomly selected from the 5000 profiles simulated under (e1s0). This conveys an idea about the sampling variability of the fitted TRPs. For the full RCRS model the fitted TRPs recover

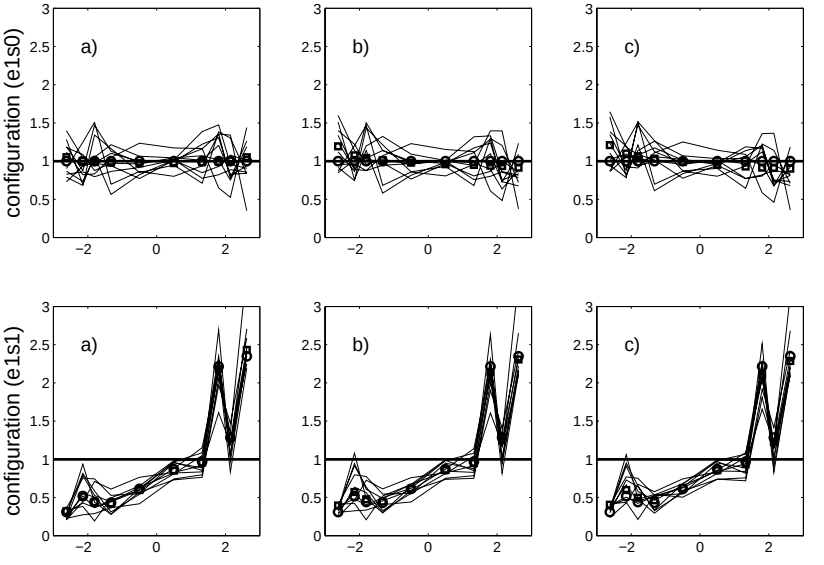

Figure 6: Tendency-to-report profiles simulated under configurations (e1s0) and (e1s1), for three RCRS models: a) full model; b) constant effect; c) no effect. Squares: average profile of all 5000 simulations. Circles: original tendency-to-report profiles underlying the simulations.

the underlying TRPs fairly well; for the restricted models there is a small bias. The picture is completed by the results of the simulations performed under configuration (e0s0). Then the p-value histograms are consistent with the uniform distribution, and the TRPs with the constant 1, for all three RCRS models (not shown).

Let us turn to the simulations involving selection. Not surprisingly in view of the high significance of the tests for selection, test power is very high under (e1s1): the simulated p-values are tightly concentrated near 0, the largest among them being about 10^{-4}. The corresponding histograms are degenerate and need not be shown. The bottom row in Figure 6 shows again ten TRPs each, selected randomly from the 5000 cases simulated under (e1s1). The average TRPs recover the underlying TRP (21) fairly well. The simulation results for configuration (e0s1) closely resemble those for (e1s1) (not shown).

7.2 Assessment of the Tests for Experimental Effects

For presentation we concentrate on one test, namely of the 'no effect' null-hypothesis within the full model. This test can be carried out assuming any one of our three basic models, i.e., on the basis of the likelihood ratio statistics $LR_{nf,\,f}$, $LR_{nf,\,r}$, $LR_{nf,\,s}$ associated with the FCR, RCR,

or RCRS model, respectively. For better comparison all three of them are studied together.

A test is valid in the RCRS model if the appropriate p-value is uniformly distributed under the null-hypothesis $\beta = 0$ no matter whether selection is present or not. The first case is the critical one, so the primarily relevant parameter configuration for checking validity is (e0s1). In the top row of Figure 7 the respective p-value histograms for the three models are shown. The difference is striking: while the RCRS test's p-value distribution fits the uniform distribution not too badly – the deviation is significant, though –, the distributions for the RCR and FCR test are concentrated on very small p-values. Hence, these tests are disastrously biased, rejecting the null-hypothesis $\beta = 0$ with high probability even though it is satisfied. Under configuration (e0s0) the RCR test's p-value distribution fits the nominal distribution $\mathcal{U}$ much better (although the chi-squared goodness-of-fit test is highly significant yet); see the bottom row of Figure 7. The FCR test remains invalid also under (e0s0).

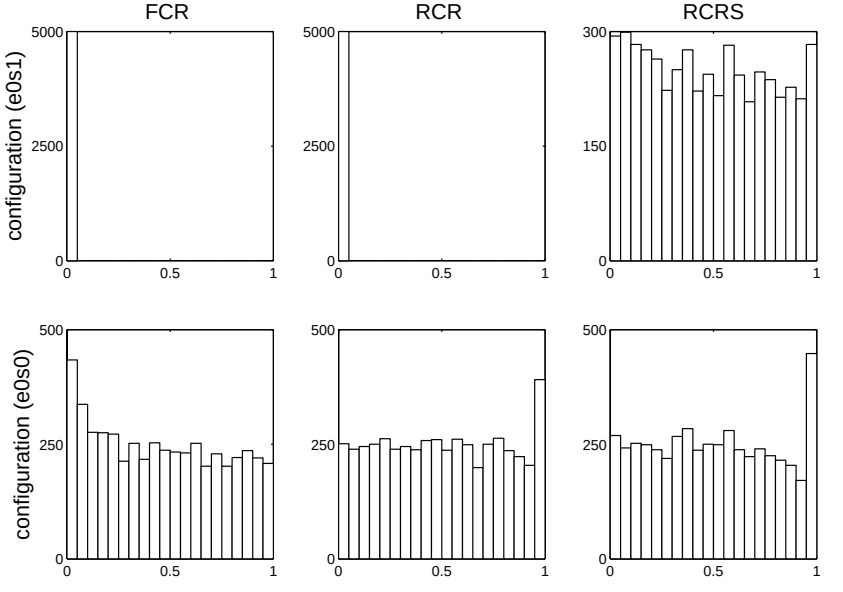

Figure 7: Testing the 'no effect' hypothesis. Histograms of the p-value distribution of the likelihood ratio test within the FCR, RCR, and RCRS model, simulated under configurations (e0s1) and (e0s0).

Relevant for the power properties of the tests are the p-value distributions obtained under the parameter configurations (e1s0) and (e1s1). For proper assessment it is important to distinguish between the nominal and the actual level, and between the nominal and the actual, or level-

adjusted, power of a test. Let us recapitulate our usage of these notions. We identify (actually *estimate*) the power of a test at nominal level γ under a particular parameter configuration (c) with the fraction among the 5000 p-values obtained by simulation under (c) that are $\leq \gamma$. If configuration (c) happens to belong to the null-hypothesis, that fraction represents the actual level of the test, which ideally should equal the nominal level, γ. If this is not the case, level-adjustment can provide more realistic information about the test power. Let (c0) denote an appropriate null-hypothesis configuration associated with the alternative configuration (c1). Let $\tilde{p}_\gamma$ be defined by the condition that $100\gamma\%$ of the p-values simulated under (c0) are $\leq \tilde{p}_\gamma$. The level-adjusted power of the test then is obtained as the fraction of the p-values simulated under (c1) that are $\leq \tilde{p}_\gamma$.

Table 2 summarizes, in lines 2–4, the related results for the FCR, RCR, and RCRS tests of the 'no effect' null-hypothesis $\beta = 0$. The entry in each configuration/test cell is the fraction of p-values ≤ 0.05, i.e., the actual level (configurations (e0s0), (e0s1)) or the nominal power (configurations (e1s0), (e1s1)) of the respective test.[15] The level-adjusted power is given, as a second entry, only in those cases where the test is not extremely biased. For completeness, we add in the last line the corresponding figures for the test of the 'no selection' hypothesis discussed in the previous subsection. The interpretation of the entries changes, of course: they now represent test level under configurations (e0s0), (e1s0), and test power under (e0s1), (e1s1).

Table 2: Level and power of tests of the 'no effect' and 'no selection' hypotheses

	(e0s0)	(e1s0)	(e0s1)	(e1s1)
$LR_{nf,\,f}$	0.7886	0.9914 / 0.5584	1	1
$LR_{nf,\,r}$	0.0686	0.6766 / 0.6294	1	0.9988
$LR_{nf,\,s}$	0.0688	0.5424 / 0.4782	0.0588	0.4234 / 0.3948
LR_{sel}	0.0514	0.0538	1 / 1	1 / 1

Let us make a few comments on the tests of the 'no effect' hypothesis. For the RCRS test the actual level is close to the nominal level no matter whether selection is absent or not (i.e., under both (e0s0) and (e0s1)), and level adjustment accordingly has a relatively minor effect on power. Under 'no selection' conditions a precaution such as adjusting for possible selection would be unnecessary, strictly speaking, and the power of the RCRS test under (e1s0) is accordingly less than that of the RCR test, which is optimal if there is no selection. Nevertheless, the loss in detection

[15] The four digits are not given to suggest accuracy – there is bigger sampling fluctuation! – but to allow reconstruction of absolute frequencies. For consistency therewith, the entries in Table 1 have been given to the same (otherwise absurd) accuracy.

probability incurred this way, from 0.6294 to 0.4782, is moderate. The results for the other two tests are less consistent. The power of the FCR test under (e1s0) reduces substantially after level adjustment, from 0.9914 to 0.5584. The high nominal power of the FCR and RCR test under (e1s1) must appear suspect considering that under (e0s1) the null-hypothesis is rejected with certainty. Quite surprisingly in view of the small p-value $p_{obs} = 8.860 \cdot 10^{-6}$ of the RCR test for the original R&N data (cf. Table 1), the power is only about 2/3 under configuration (e1s0), hence rather moderate, and not much larger than that of the RCRS test.

In conclusion, the pattern emerging from Table 2 is as follows. The power of the RCR and RCRS test depends on the presence of a selection effect, in opposite directions. For the RCRS test the power is slightly smaller under (e1s1) than under (e1s0). Conversely, the power of the RCR test is much higher under (e1s1) than under (e1s0). It would thus appear that the RCR test gains power if there is a selection effect besides the experimental effect. From this perspective, setting aside that the hypothesis tested is the 'no effect' hypothesis, one might reinterpret the entry 1 under (e0s1) as high power against (pure) selection rather than as bias.

The results so far suggest the presence of a significant selection effect in the R&N data, and reduction to non-significance of the experimental effect if adjustment for selection is made. Does this imply that there is no substantial experimental effect? There can be pitfalls in such a conclusion, and this possibility will be pursued in the next two subsections.

7.3 Adjustment for Selection, and Efficiency

A particular observation concerning the RCR test of the 'no effect' hypothesis remained unexplained so far, namely that its power under configuration (e1s0) is unexpectedly moderate, between 0.6 and 0.7. One explanation mentioned above would be that the test "borrows strength" from selection effects if such are present, and accordingly loses power upon their removal, as in (e1s0). Another, more hidden possibility connects to our choice of the configurations underlying the simulations, which are based on the MLE for the (full) RCRS model. It is conceivable that some part of the β-related effect is assigned to selection within the RCRS model. The MLE of β might then be shifted towards 0, and thus give rise to a configuration (e1s0) that is less easily separable from the null-hypothesis. Put differently, if there is no selection, the MLE $\hat{\beta}_r$ under the RCR model might give a more accurate, less damped estimate of the "true" β than the MLE $\hat{\beta}_s$ under the RCRS model.

To examine this, let us consider hybrid configurations obtained by mixing the RCR- and RCRS-based MLEs, using the appropriate indices for distinction. The asterisks in the following list indicate that configu-

rations (e0s0), (e0s1) are not exactly identical to those introduced earlier because the random effects parameter ϕ here is estimated assuming the RCR rather than the RCRS model.

(e2s1) $(\beta, \phi, \alpha) = (\hat{\beta}_r, \hat{\phi}_r, \hat{\alpha}_s)$
(e0s1)* $(\beta, \phi, \alpha) = (0, \hat{\phi}_r, \hat{\alpha}_s)$
(e2s0) $(\beta, \phi, \alpha) = (\hat{\beta}_r, \hat{\phi}_r, 0)$
(e0s0)* $(\beta, \phi, \alpha) = (0, \hat{\phi}_r, 0)$

Table 3 presents the simulation results under these configurations. Entries are as in Table 2: the first entry is the fraction of simulated p-values that are ≤ 0.05; if present, the second entry gives the level-adjusted power (for moderately biased tests only). Qualitatively, the results resemble those in Table 2. The important new feature here is the high power of the RCR test under (e2s0). Does this confirm the presumption that the relatively poor power under (e1s0) has to do with a damping of the estimate $\hat{\beta}_s$ compared to $\hat{\beta}_r$? If so we encounter a dilemma: attempts to allow for selection would have to be paid by a severely diminished chance to detect experimental (β-related) effects. Notice, however, that the power of the RCRS test, too, is substantially increased under the hybrid configurations. In the next subsection we approach the puzzle by considering a slightly more simple case that is of interest in its own right.

Table 3: Level and power under hybrid configurations

	(e0s0)*	(e2s0)	(e0s1)*	(e2s1)
$LR_{nf,\,r}$	0.0714	0.9984 / 0.9964	1	1
$LR_{nf,\,s}$	0.0790	0.9136 / 0.8788	0.0752	0.8370 / 0.7856
LR_{sel}	0.0534	0.0548	1 / 1	1 / 1

7.4 Separate Analysis of the PEAR Data

A substantial part of the data collected in R&N's meta-analysis stems from the Princeton Engineering Anomalies Research (PEAR) laboratory alone. The PEAR data contribute 284 of the originally 597 single studies, and more than 75% of the total number of bits. They rank highly with respect to the quality criteria considered by R&N, and it is interesting to study this subsample by itself.

All proper covariates except for quality rating are constant across the PEAR data, hence may be omitted. The full regression model then involves only two β-parameters. The variability of the residuals is smaller than for the complete data. Nevertheless, the chi-squared goodness-of-fit test rejects the FCR model with a p-value of about $7 \cdot 10^{-4}$, so again, we focus on the RCR and RCRS models. Judged from Figures 8 and 9, both

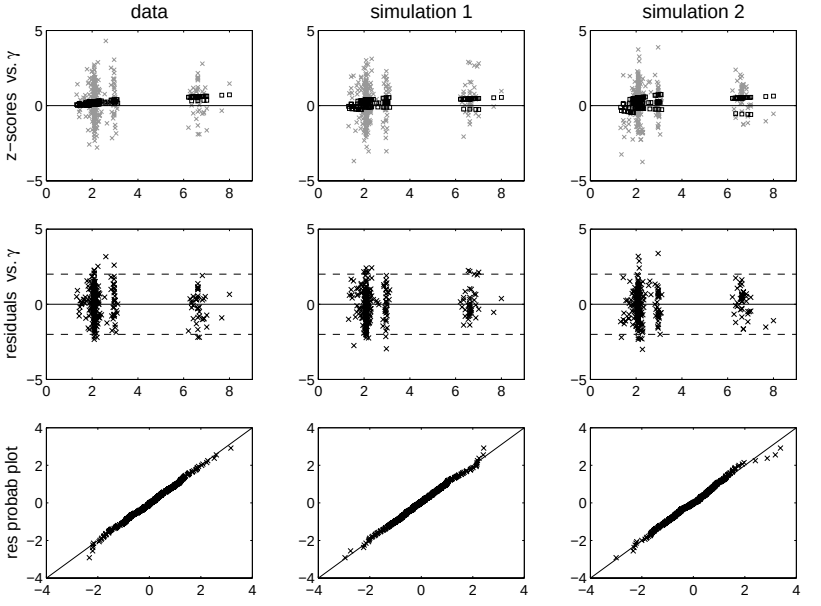

Figure 8: Fit of RCR model for PEAR data; otherwise as Figure 1.

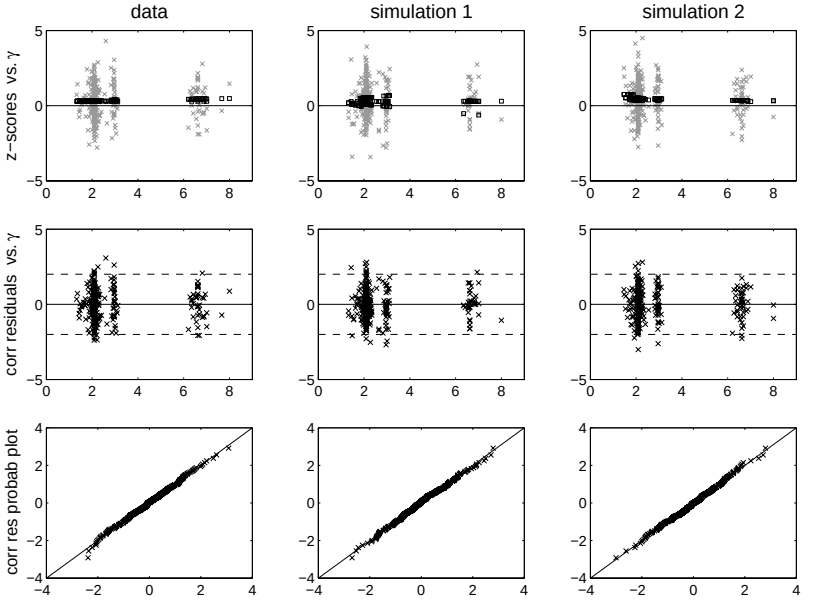

Figure 9: Fit of RCRS model for PEAR data; otherwise as Figure 1.

models fit satisfactorily. The RCR-fitted values scatter along a straight line through the origin, indicating that the quality rating is of minor importance. If there is no selection, the RCRS-fitted values should resemble the RCR-fitted values. However, the former seem to scatter along a straight line *parallel* to the abscissa. This might indicate the presence of a selection mechanism, as might the estimated TRP

$$0.62 \quad 0.57 \quad 0.79 \quad 0.70 \quad 0.85 \quad 1.03 \quad 1.18 \quad 1.91 \quad 0.83 \quad 1.53$$

which is similar in shape to, though less pronounced in numerical value than, that of the complete data; cf. (21). However, there is an important difference to the full R&N data set: the test for selection is *not significant* for the PEAR data, the (nominal) p-value being 0.5610 in this case. In regard to the 'no effect' hypothesis the situation is similar to the case of the complete data: the RCR test is significant, $p_{obs} = 2.101 \cdot 10^{-5}$, whereas the RCRS test is not, $p_{obs} = 0.5047$.

The validity and power properties of the tests can be checked by simulation under configurations determined from the RCRS-based MLE in the same manner as in Section 7.2. Results are shown in Table 4, and in Table 5 for the hybrid configurations constructed as in Section 7.4. The meaning of the entries is as in Tables 2 and 3, respectively.

Table 4: PEAR data level and power

	(e0s0)	(e1s0)	(e0s1)	(e1s1)
$LR_{nf,\,r}$	0.0450	0.3956 / 0.4162	0.7178	0.9928 / 0.5818
$LR_{nf,\,s}$	0.0450	0.1732 / 0.1846	0.0530	0.1608 / 0.1564
LR_{sel}	0.0544	0.0516	0.4216 / 0.4054	0.4532 / 0.4480

Table 5. PEAR data level and power under hybrid configurations

	(e0s0)*	(e2s0)	(e0s1)*	(e2s1)
$LR_{nf,\,r}$	0.0510	0.9912 / 0.9900	0.7138	1 / 0.9986
$LR_{nf,\,s}$	0.0534	0.6316 / 0.6238	0.0538	0.5930 / 0.5760
LR_{sel}	0.0564	0.0550	0.4142 / 0.3886	0.5130 / 0.4962

Qualitatively, the results are similar to those for the complete data, with two important differences. The most conspicuous new feature is the moderate power of the test for selection, which is in accordance with the non-significance of the test result for the PEAR data. Moreover, the power loss of the RCRS test of the 'no effect' hypothesis with respect to the RCR test is more severe than in Tables 2 and 3. These facts, and the reasonable fit of the RCR model, may suggest to discard the possibility of selection

and to conclude, on the basis of the RCR test of the 'no effect' hypothesis, that there is a highly significant experimental effect. The underlying assumption would be that a non-significant selection effect cannot possibly add enough strength to the RCR test of the 'no effect' hypothesis to make the experimental effect highly significant if it is actually negligible. However, this assumption is unjustified.

Figure 10 shows the dispersion of the RCR- and RCRS-based MLEs of β in the 2-dimensional parameter plane when simulated under four configurations of interest, with 50 repetitions each. The RCRS-MLE scatters around the β-parameter that underlies the simulations under the respective configuration, hence is correctly centered in all cases; in particular, there is no damping towards zero. On the other hand, the RCR-MLE is clearly shifted away from the "true" β in the positive β_1-direction under selection conditions (e0/1s1), which means that it estimates a larger experimental (constant) effect than is actually present. This occurs even though the added selection effect is far from being significant. Note furthermore that the scatter clouds of the RCR-MLEs under (e1s1) and (e2s0) look almost the same.

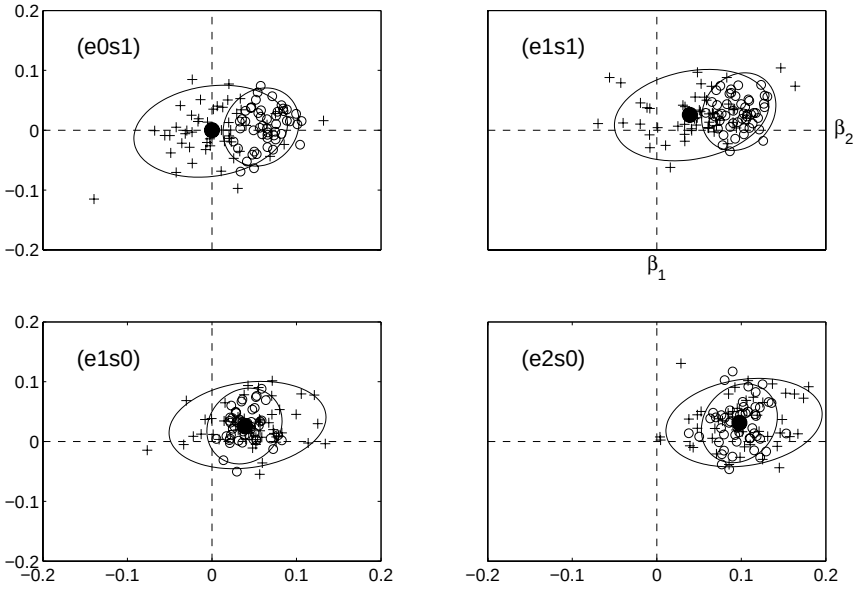

Figure 10: Scatter of RCR- (o) and RCRS-based (+) MLEs of parameter β simulated under four different configurations for the PEAR data. Filled circles represent parameter β underlying the simulation. Confidence ellipsoids with 90% coverage probability are calculated on the basis of all 5000 simulations.

These observations may be summarized as follows: The individually weak (non-significant) experimental and selection effects as estimated by the RCRS-MLE appear, when combined under configuration (e1s1), as a strong (highly significant) experimental effect from the perspective of the RCR model. Or put differently: the RCR-test for experimental effects is very sensitive to small, even non-significant perturbations of the selection type. Taken together, the above findings explain the discrepancy in the results of the RCR and the RCRS test for the PEAR data, as well as the structure of the entries in Tables 4 and 5. The scatter plots also indicate how the RCRS test's efficiency loss comes about, namely by an increased dispersion of the RCRS-MLE compared to that of the RCR-MLE.

In the case of the full R&N data, matters are complicated by the higher dimensionality of the β-parameter (6 rather than 2), and by collinearity of the covariates precluding a straightforward interpretation of the single regression coefficients β_k. For a rough assessment one may consider the reduced RCR(S) model that includes only quality rating as a proper covariate, as in the case of the PEAR data. Then again, the RCRS-MLE is correctly centered under all configurations; the RCR-MLE exhibits a shift in the positive β_1-direction under selection conditions; and the scatter clouds of the RCR-MLEs simulated under the configurations (e1s1) and (e2s0) are similar to each other. Thus the above explanation of the test results for the PEAR data (Tabs. 4 and 5) essentially applies to the case of the full R&N data (Tabs. 2 and 3) as well.

7.5 Discussion

Two notions conceptually helpful in discussing the results of this section are *competing factors* and *(model) mis-specification.* The first refers to the fact that the RCRS model offers *two* competing explanations for large z-score values under experimental conditions. They could be due to a) a shift of the expected z-scores as incorporated in the FCR and RCR models, or to b) selective reporting (or to both a) and b), of course). Mis-specification means that the assumed model fails to incorporate important features of the data producing mechanism. It usually implies invalidity of related tests and a bad model fit. Both competing factors and mis-specification invite false explanations, i.e., may lead one to ascribe a phenomenon to a cause unconnected to its actually responsible origin. Of course, one hardly ever knows "the true cause" in such situations. However, through simulation under known regimes one can manipulate potential causes and study the impact on the outcome, as done here with various tests.

The problem of competing factors is serious because the standard normal density $\varphi(z)$ can be transformed into any other density by a multiplier $\kappa(z)$ as in (16). In particular, it can be transformed into a density

$\varphi((z-\theta)/\sigma)/\sigma$ differing from $\varphi(z)$ by an affine transformation of z, as in the RCR model. Evidently, it is impossible to decide whether the change to the density $\varphi((z-\theta)/\sigma)/\sigma$ "is due to" an affine transformation standing for an experimental effect or to a multiplicative modification interpretable as a selection effect. This problem has been pointed out clearly in recent work by Dobyns (2003), within a more restricted setting.

How, then, is it possible that, as found in Section 7.1, the likelihood ratio test for selection does not confound the two effects, yet has essentially the correct null distribution? The answer is that the z-score means (and variances) are not identical in the RCR model but inherit structure from the variation of study specific features across studies. The selection modification κ, on the other hand, is the same for all studies in the RCRS model, hence cannot adapt to all the different affine transformations $z \mapsto \mu_i + \sigma_i z$ simultaneously. It *could* do so if the z-scores were identically distributed. (The restriction of κ to a step function alters the situation only marginally.) Consequently, unbiasedness of the test for selection must not be supposed to apply generally. However, in the case of the R&N data the variation of the covariates and, most importantly perhaps, of the sample sizes as encoded in the informativity quantities γ_i, is apparently pronounced enough to ensure validity of the test.

Model mis-specification often is responsible for invalidity of tests and inconsistency of results.[16] In all cases of serious lack of validity found above, the test statistic is constructed on the basis of a mis-specified model. This applies, in particular, to the RCR test of the null-hypothesis $\beta = 0$ under configuration (e0s1). "Knowing" of no other explanation, the RCR model "is forced" to interpret the selection-induced shift of the z-scores occurring under (e0s1) in terms of features that are accounted for by the model. Considering the scatter of the RCR-MLE under (e0s1) (Figure 10) or the p-value distribution of the RCR test (Figure 7), "it does so" by ascribing the z-score shift to a shift of the parameter β, i.e., to an experimental effect. For comparison, under configurations (e0/1/2s0) no element unaccounted for by the RCR model is present, and the sampling distributions (approximately) exhibit the proper behavior. The important observation here is that a combination of experimental and selection effects can pretend a highly significant experimental effect when considered from the perspective of the RCR model, even if *both* underlying effects are non-significant.[17]

[16] The consequences tend to be particularly severe if effects are small and sample sizes are huge, as is typically the case in the experiments considered here.

[17] A referee argued that analyses based on the RCRS model can falsely detect selection when there is none (see also the paragraph after the tag for this footnote), and can reduce a truly significant experimental effect to non-significance after adjustment for selection. The conditions required to produce these failures imply model mis-specifications not considered in this paper. It is unclear whether they apply to the data under study.

For another, less dramatic example consider Figs. 5b and 5c where simulations do not involve selection but a nonzero β-parameter, namely the MLE $\hat{\beta}$ in the full RCRS model. This choice does not fit into either of the restricted RCRS models that are used for constructing the two tests of the 'no selection' hypothesis: one neither has $\hat{\beta} = 0$ ('no effect') nor $\hat{\beta}_2 = \ldots = \hat{\beta}_6 = 0$ ('constant effect'). The resulting discrepancy with the respective null-hypothesis, lacking a suitable model-internal explanation, is assigned to selection, hence increases the probability of a small p-value for the corresponding test. In turn, under configuration (e0s0) the p-value distribution is essentially uniform, as it ought to be. The FCR tests, finally, are systematically misled by high data variability due to the random effects incorporated in all simulation regimes.

Let us mention in passing that a test may lack validity also on grounds unrelated to mis-specification. In fact, the fit to the nominal uniform distribution $\mathcal{U}$ is rather poor especially in Figure 7, bottom row, although the simulations are carried out under the null-hypothesis. The reason is that $\mathcal{U}$ is only an approximation (hence the adjective "nominal") to the actual p-value distribution, derived within a framework where the sample of available data is supposed to be "large". The quality of such approximations depends not only on the sample size but also on the complexity of the statistical model and the related procedures. Complexity here alludes to statistical aspects of the model such as functional form and number of unknown parameters, as well as to the numerical algorithms used to compute the MLE. The RCR(S) models are fairly complex in this sense,[18] so some asymptotics bias has to be faced. However, as a rule such bias is quantitative in nature and less serious than bias due to model mis-specification, where errors may scale to orders of magnitude. Finally, the RCRS test's p-value distribution under both (e0s0) and (e0s1) is very similar to that of the RCR test under (s0e0). This indicates that the moderate (but significant) deviation from $\mathcal{U}$ is a matter of asymptotics bias rather than, in case of configuration (e0s1), bias due to the competing factors problem.

Experimental effects of the considered kind of mind-matter experiments have been identified with mean value shifts throughout this study. More generally one might argue that *any* deviation of the z-score distribution under experimental conditions from the one under control conditions (actual or theoretically expected) is indicative of an experimental effect. However, it is not clear *a priori* whether such a difference would be due to an effect of the envisaged kind, i.e., to interrelations between mental and physical states. At the phenomenological level, the manifest difference in

[18] E.g., the (full) RCRS model involves a total of $6+21+9 = 36$ parameters (β, ϕ, α-related), most of them in a nonlinear form. Compared to this, the sample size of 500-600 z-scores is not very large.

the z-score distributions of the R&N data under control and experimental conditions can be described by random and selection effects within the RCRS model. Since selection effects do not qualify as experimental effects, one may speculate whether the increased z-score variability under experimental conditions represents itself an experimental effect of the envisaged kind. This possibility was addressed by Dunne (1998).

The present study has a number of other limitations. We did not consider random or covariate effects in any detail (see *e.g.* Steinkamp *et al.* (2002) for work on the latter), nor sensitivity aspects other than those related to selection type perturbations of the RCR model, or more refined residual analyses. As for the fundamental competing factors problem, the performance (regarding validity and efficiency) of RCRS-based tests has been studied only via simulation. It would be desirable to achieve some theoretical understanding of that dependence. Although the general form of the problem is well understood (inference in the presence of nuisance parameters), more concrete (local asymptotic) analyses such as those of Copas and Eguchi (2001) in a related but different setting may be worthwhile.

8. Summary

In this paper we reconsider an influential meta-analysis by Radin and Nelson in 1989 (R&N) of experiments investigating possible interrelations between the intention of a human operator and a physical random event generator. In our approach, we shift the focus from the establishment of effects toward an explicit statistical modeling of "mechanisms" that could have produced data such as those reported. It thus becomes possible to rule out inappropriate models and to assess the relevance of the single model components for an explanation of the data. As usual in corresponding studies, the main experimental effect is modeled as a mean shift. Additional model components account for increased data variability, through unsystematic/random as well as systematic modifications of the main effect, and for possible selection mechanisms. Such RCRS models (Random Coefficients Regression involving Selection) allow us, *e.g.*, to estimate selection profiles or to adjust the experimental effect for possible selection (and *vice versa*).

Due to possible confounding of experimental and selection effects, corresponding statistical procedures may be seriously biased. We therefore study the consequences of those effects through simulation under various regimes of interest. Under conditions adapted to the data of R&N, procedures accounting for possible selection turn out to be surprisingly valid and efficient. On the other hand, procedures ignoring selection can exhibit serious bias, even if selection effects are not statistically significant. Such consequences of a mis-specification of the statistical model (in

this case due to disregarding possible selection) have not found adequate consideration in previous meta-analyses of mind-matter experiments.

It is as important to note, as it is a matter of course, that the actual data-producing "mechanism" need not follow the assumed model (here: RCRS model). Model checking is helpful, and used extensively in this paper, to detect possible deviations from the model, but it can never *confirm* the model. This means that the origin of the effects that are ascribed to experimental conditions, to random variation, or to selection by the RCRS model for the data under consideration must remain open.

With this *proviso*, the main conclusions in regard to the R&N data are as follows. First, there are indications for the presence of both random and selection effects. Secondly, the test for an experimental effect is significant if possible selection is ignored, and becomes non-significant after adjustment for selection. Let us emphasize again that it cannot be formally verified by our approach whether or not selection in fact explains the observed patterns in the data. Nevertheless, our results cast doubt on the notion that the experimental effect is primarily reflected in an increase of the basic hit probabilities. Thus, other explanations appear necessary to account for the obvious difference between the data obtained under experimental and control conditions. Corresponding proposals along lines different from those followed here were made by Jahn *et al.* (2000) and by Atmanspacher and Jahn (2003).

The results for the subset of studies conducted in the Princeton Engineering Anomalies Research laboratory (PEAR data) differ in one important respect from those of the full R&N data base: the test for selection is non-significant. This is in accordance with the complete reporting of the PEAR data pointed out in Dobyns (2003). However, when testing for an experimental effect one encounters the same problem found with the full data: the highly significant test result obtained if selection is ignored becomes non-significant after adjustment for possible selection, even though selection effects are non-significant in this particular case. This seemingly counterintuitive fact illustrates the delicate character of meta-analyses in this area and suggests a cautious view of the subject matter abstaining from strong assertions either way.

Appendices

A. Derivation of the RCR Model from the GBP: Mathematical Details

Let us conceive of the gBp model as being embedded into a sequence of similar models virtually indexed by $\nu = 1, 2, \ldots$. Thus every quantity should be thought of as carrying an additional index ν, and related statements should be accompanied by the phrase "as ν tends to infinity". E.g.,

a statement such as "ξ_i is small/large" should be read as "$\xi_{i,\nu}$ tends to zero/infinity as ν tends to infinity". The informally stated assumption (B4) has to be specified accordingly:

(B4') (i) N_i tends to infinity, $p_{i,0}$ is bounded away from 0 and 1, and s'' is uniformly bounded;
(ii) $N_i^{1/4} x_i\beta$ tends to zero;
(iii) $N_i^{-1/2} \sum_j E\,(x_i\rho_{ij})^2$ and $\mathrm{Var}\left(N_i^{-1}\sum_j x_i\rho_{ij}\right) = \delta_i^2\, x_i\,\Omega\, x_i^T$ tend to zero.

We shall make use of the stochastic big/little oh notation: *e.g.*, $X_i = O_p(1)$ or $X_i = o_p(1)$ means that the (virtual) sequence of random variables $X_{i,\nu}$ is stochastically bounded or tends to zero in probability, respectively, as ν tends to infinity.

Fixing the i-th study, we first derive stochastic expansions for the conditional expectation and variance of $S_i = \sum_j Y_{ij}$. These in turn imply a corresponding expansion for the z-score Z_i. The conditioning is on the bit-level hit probabilities $p_{ij} = s\big(\theta_{i,0} + x_i(\beta + \rho_{ij})\big)$ or, equivalently, on the σ-algebra $\mathcal{R}_i$ generated by the random effects ρ_{ij}, $j = 1, \ldots, N_i$. A first order Taylor expansion of s around $\theta_{i,0}$ using (B4'), (i) gives

$$\begin{aligned} p_{ij} &= p_{i,0} + s'(\theta_{i,0})\, x_i(\beta+\rho_{ij}) + O\left((x_i\beta)^2\right) + O\left((x_i\rho_{ij})^2\right), \\ p_{ij}(1-p_{ij}) &= p_{i,0}(1-p_{i,0}) + (1-2p_{i,0})\, s'(\theta_{i,0})\, x_i(\beta+\rho_{ij}) \\ &\quad + O\left((x_i\beta)^2\right) + O\left((x_i\rho_{ij})^2\right), \end{aligned}$$

with O-estimates holding uniformly in j (and i). Set $A_i = N_i^{-1}\sum_j \rho_{ij}$. Then by (B4')

$$\begin{aligned} E\left[S_i \,|\, \mathcal{R}_i\right] &= \sum\nolimits_j p_{ij} \qquad (24) \\ &= N_i\left(p_{i,0} \;+\; s'(\theta_{i,0})\left[x_i\beta + x_i A_i\right]\right) \;+\; o_p\left(N_i^{1/2}\right), \\ \mathrm{Var}\left(S_i \,|\, \mathcal{R}_i\right) &= \sum\nolimits_j p_{ij}(1-p_{ij}) \qquad (25) \\ &= N_i p_{i,0}(1-p_{i,0}) \;+\; N_i\,(1-2p_{i,0})\, s'(\theta_{i,0})\left[x_i\beta + x_i A_i\right] \\ &\quad + o_p\left(N_i^{1/2}\right) \\ &= N_i p_{i,0}(1-p_{i,0}) \;+\; o_p(N_i). \end{aligned}$$

For the last step note that one has $N_i x_i\beta = o(N_i^{3/4})$ and

$$\begin{aligned} N_i x_i A_i = \sum\nolimits_j x_i\rho_{ij} &= O_p\left(E\left(\sum\nolimits_j x_i\rho_{ij}\right)^2\right)^{1/2} \\ &= O_p\big(N_i^2\,\delta_i^2\, x_i\Omega x_i^T\big)^{1/2} = o_p(N_i). \end{aligned}$$

Conditionally on $\mathcal{R}_i$, S_i is a sum of independent Bernoulli variables whose (conditional) variance tends to infinity in probability; cf. (25) and (B4') (i). The central limit theorem and (25) imply that, with probability tending to 1, the re-centered z-score

$$Z_i^c \;=\; Z_i - E\left[Z_i \mid \mathcal{R}_i\right] \;=\; \frac{S_i - E\left[S_i \mid \mathcal{R}_i\right]}{\left(N_i p_{i,0}(1-p_{i,0})\right)^{1/2}}$$

is approximately $\mathcal{N}(0,1)$-distributed conditionally on $\mathcal{R}_i$. Using (24) and recalling the definition of γ_i in (6) one finds that the second term in the decomposition $Z_i \;=\; Z_i^c + E\left[Z_i \mid \mathcal{R}_i\right]$ admits the stochastic expansion

$$E\left[Z_i \mid \mathcal{R}_i\right] \;=\; \frac{N_i\, s'(\theta_{i,0})\, x_i(\beta + A_i)}{\left(N_i p_{i,0}(1-p_{i,0})\right)^{1/2}} \;+\; o_p(1) \;=\; \gamma_i\, x_i(\beta + A_i) \;+\; o_p(1)\,.$$

It follows at first that with probability tending to 1 the conditional distribution of Z_i given $\mathcal{R}_i$ is approximately the normal distribution $\mathcal{N}(W_i\,, 1)$, where $W_i = \gamma_i\, x_i(\beta + A_i)$. Hence, the unconditional distribution of Z_i can be approximated by the normal shift mixture $\mathcal{M}_i = \int_{-\infty}^{\infty} \mathcal{N}(w,1)\, \mathcal{G}_i(dw)$, where $\mathcal{G}_i$ denotes the distribution of W_i. Since $\mathcal{G}_i$ has mean $\gamma_i x_i \beta$ and variance $\gamma_i^2 \delta_i^2\, x_i \Omega x_i^T$, the corresponding moments of $\mathcal{M}_i$ are $\gamma_i x_i \beta$ and $1 + \gamma_i^2 \delta_i^2\, x_i \Omega x_i^T$, respectively. A mixture like $\mathcal{M}_i$ is (approximately) normal if the mixing distribution, here $\mathcal{G}_i$, is (approximately) normal. Therefore, under the further assumption

$$\text{(B6)} \qquad \mathcal{G}_i \;\simeq\; \mathcal{N}\left(\gamma_i x_i \beta\,,\, \gamma_i^2 \delta_i^2\, x_i \Omega x_i^T\right)$$

(in addition to (B1)–(B3), (B4') and (B5')) we may conclude that Z_i is approximately distributed as $\mathcal{N}\left(\gamma_i x_i \beta\,,\, 1 + \gamma_i^2 \delta_i^2\, x_i \Omega x_i^T\right)$. This completes the derivation of the RCR model.

Normality assumptions such as (B6) are common in random effects ANOVA, although a justification for single measurements data is often lacking. In meta-analysis, on the other hand, the mixing distribution $\mathcal{G}_i$ is the distribution of an *average* of random effects (namely of $\gamma_i x_i A_i$, up to a translation), which obeys the central limit theorem under broad conditions. Therefore, the assumption (B6) appears reasonable in the present context.

B. Definition of Covariates, Exclusion of Studies

B.1 Covariates

Our regression model includes supplementary information about four variables. Three of them are categorical in character: the type of REG used in a study, the kind of task posed, and the manner how agents were

recruited for study participation ("subject selection"). They attain 10, 3, and 2 different levels, respectively; see Radin and Nelson (1987). The fourth variable is a quality score Q_i assigned to each study which assumes discrete values 0, 1, 2, ..., 15, a high score signifying a high methodological quality. We treat it as a numerical regressor variable, i.e., we take Q_i itself as a covariate. An ANOVA model (main effects only) for the three categorical variables would add $9 + 2 + 1 = 12$ covariates (so-called dummy variables) and the constant $x_{i1} = 1$, so in total we had $p = 14$ β-parameters plus $p(p+1)/2 = 105$ ϕ-parameters for the covariance matrix Ω, which is too much. In order to reduce this to a statistically tractable number of unknowns to be estimated, we coarsen the classification of REG types by distinguishing only between truly physical and pseudo-random sources. Then the ANOVA model for the categorical variables has only 5 covariates $x_{i1}(= 1), x_{i2}, \ldots, x_{i5}$, and with $x_{i6} = Q_i$ we end up with $p = 6$ β- plus $p(p+1)/2 = 21$ ϕ-parameters, which is feasible. A crude comparison on the basis of the FCR models with and without dichotomization of the REG type suggests that nothing of importance gets lost by the coarsening.

In summary, our basic regression model has covariate vectors $x_i = (x_{i1}, \ldots, x_{i6})$, where $x_{i1} = 1$ and the 0/1 dummy variables $x_{i2}, \ldots, x_{i5}$ generate the ANOVA model without interactions for the three factors REG type (dichotomized), kind of task, and subject selection; the last covariate is $x_{i6} = Q_i$.

B.2 Exclusion of Studies

Among the primarily $n = 597$ studies there is a fraction that is excluded from our analyses because part of the required information is missing. Study $\#i$ is omitted if, according to the reported data, either $z_i = 0$, or $p_i = 0$, or $N_i = 1$. These special values were assigned by R&N if the actual values could not be retrieved. In particular, z-score 0 was assigned if the study was reported as "not significant" only. There are 80 (of 597) cases that satisfy one of the three conditions and hence are excluded here. While this makes for 13.4 % of the studies, the total number of bits removed makes only for 1% of the total number N_{tot} of bits in all studies. Therefore, omission of these studies should have little influence on the results, at least on those concerning experimental effects. The situation is less clear in regard to selection effects; but see below.

There is one further case that is excluded because it has *too much* influence. A study of Berger (1986), with z-score 2, collected $N = 9.23 \cdot 10^7$ bits, i.e., 20% of N_{tot}. In regression analysis such a case is called a *leverage point,* and there is a whole literature demonstrating, and warning about, the potentially overwhelming influence such a case can have on the results of a statistical analysis (Huber 1981). Since "robust" procedures damping

down the influence of leverage points do not seem to be available for the type of models considered here, we take the most simple measure of precaution and omit the study.

To summarize: after excluding the 80 defective studies plus Berger's study we are left with a total of $n = 516$ (of the originally 597) studies. These form the data base for our analyses.

B.3 Consequences of Study Exclusion

The omission of Berger's study is inconsequential: its re-inclusion affects neither the test results for selection and experimental effects presented in Sections 6.2, 6.3, nor the TRP given in display (21) in any essential way. This outcome may appear surprising in view of the *potential* influence of Berger's study. It becomes understandable when considering that a z-score of 2 is in fact rather modest for a study as large as Berger's, and too modest to swamp the results.

It is not obvious how the 80 studies with missing values should be made use of in our analyses. One may suspect that the selection effects detected by the RCRS model are partially due to the omitted studies. While there is no clean solution to this problem, a rough assessment can be achieved by multiple random imputation of missing values (Rubin 1996). Our implementation of this method is as follows. In studies reported as "not significant" (and assigned the z-score 0 by R&N), we assign a z-score drawn at random from the standard normal distribution subject to the condition that it is less than 1.65 (i.e., "non-significant"). In studies where the number of bits N_i is missing (and assigned the value 1 by R&N), we assign values drawn at random from the N_is belonging to some collection of studies for which this quantity is available. It is unclear which such collection is appropriate, so we consider both the collection of all non-PEAR studies (since the PEAR data contain no missings), and the collection of all studies (with known value of N_i). Values for missing $p_{i,0}$s are imputed by an entirely analogous procedure. This random assignment of missing values is repeated 1000 times, and the tests for experimental and selection effects are computed for each of the 1000 thus complemented data sets, yielding distributions of p-values that may be characterized by average p-values and percentiles of interest.

The test for selection (applied to the full, complemented R&N data) turns out to remain highly significant also under random imputation: the average p-values are $< 10^{-10}$ for both collections, with none of the (altogether 2000) p-values exceeding 10^{-9}. The average TRP for the non-PEAR collection,

$$0.29 \quad 0.50 \quad 0.51 \quad 0.52 \quad 0.79 \quad 1.04 \quad 1.11 \quad 2.15 \quad 1.24 \quad 1.86$$

is very close to the one for the "all studies collection". It resembles the profile given in (21) fairly well, with a slight shift towards the value 1 at

some places. The tests for an experimental effect exhibit patterns similar to that of Table 1. For example, the average p-values for the RCR test of the 'no eff.' vs. 'full' hypotheses are about 10^{-5} for both collections. For the RCRS test, the corresponding figures are 0.27 and 0.41 for the non-PEAR and all studies collections, respectively. These p-values show some variability across the 1000 data sets, with qualitatively similar results. Thus the lower 1% quantile of the RCRS test p-values is still larger than 0.05 for both collections.

We conclude that the omission of the 81 studies has no essential impact on the results of this paper. In particular, it does not account for the selection effects found by the RCRS model in the reduced data base that we consider in the main body of this paper.

C. Implementation Details

C.1 Selection-Corrected Residuals

In the presence of a selection mechanism the usual residuals are not even approximately $\mathcal{N}(0,1)$-distributed. To exclude the most serious biases we standardize the z-scores using means and variances corrected for selection, i.e., computed under the RCRS model. Denoting the corrected moments by $\mu_{i,s}$, $\sigma^2_{i,s}$ we get by straightforward calculation

$$\begin{aligned}
\mu_{i,s} &= \mu_i - \sigma_i \, \frac{\sum_k e^{\alpha_k} \, (\Delta\varphi)^i_k}{\sum_k e^{\alpha_k} \, (\Delta\Phi)^i_k} \,, \\
\sigma^2_{i,s} &= \sigma^2_i \left[1 + \frac{\sum_k e^{\alpha_k} \, (\Delta\varphi')^i_k}{\sum_k e^{\alpha_k} \, (\Delta\Phi)^i_k} - \left(\frac{\sum_k e^{\alpha_k} \, (\Delta\varphi)^i_k}{\sum_k e^{\alpha_k} \, (\Delta\Phi)^i_k} \right)^2 \right],
\end{aligned}$$

respectively, where for any function g we put

$$(\Delta g)^i_k = g\left((c_k - \mu_i)/\sigma_i\right) - g\left((c_{k-1} - \mu_i)/\sigma_i\right).$$

The corrected residuals then are given by

$$r_{i,s} = (z_i - \hat{\mu}_{i,s})/\hat{\sigma}_{i,s} \,,$$

where the hats indicate that MLEs are substituted for the unknown values of the model parameters β, ϕ, and α.

C.2 Computation of Maximum Likelihood Estimates

This requires numerical optimization of the respective log-likelihood over the appropriate parameter region. The computations are carried out using the MATLAB routine `fminunc.m`, which implements a sophisticated quasi-Newton method, see Coleman *et al.* (1999). Submodels are fitted by rewriting the problem as an unconstrained optimization problem (and

then using the same routine). Starting values are set as follows. For both the RCR- and RCRS-MLE the starting value for parameter β is chosen as the least-squares estimate on the basis of the FCR model. The starting value for parameter ϕ is chosen so that $\Omega = \phi\phi^T$ equals the $p \times p$ unit matrix; starting values for αs are $\alpha_k = 0$ for all k (RCRS-MLE only). Different starting values can give different figures. However, the general pattern of the results and the ensuing conclusions remained unaffected in all cases considered. Default convergence criteria are met in all cases with a maximum number of iterations set to 60.

C.3 Simulation under the RCRS Model

Suppose given parameters β and ϕ for the RCR model, and $K + 1$ parameters $w_k = e^{\alpha_k}$ representing the tendency-to-report profile. In order to mimic the selection process we generate a larger source data set following an RCR model with parameters β, ϕ and afterwards thin it out by coin tossing. Let z-score-dependent retain probabilities $\kappa(z)$ be piecewise defined as $\kappa(z) = 0.9\, w_k \,/\, \max_j w_j$ if z falls into the interval I_k.[19] The procedure then goes as follows. In the first step, generate a source data set by making five copies each of the $n = 516$ available cases (studies). From the latter, keep the covariate vectors x_i and informativity quantities γ_i, and replace the original z-scores by (independent) simulated z-scores distributed as $\mathcal{N}(\gamma_i x_i \beta\,,\, 1 + x_i\, \phi\phi^T x_i^T)$. In the second step these 5×516 cases are randomly thinned out: a case with (simulated) z-score z is deleted with probability $1 - \kappa(z)$, independently of the others. If the number n_r of remaining cases is $\geq n = 516$, then in the third step n out of the n_r cases are selected by random sampling without replacement, yielding a simulated data set of the same size as the original. If $n_r < n$, steps 1 to 3 are repeated until the first time when $n_r \geq n$. (Five copies suffice to keep the number of repetitions reasonably small.)

A subtle point has to be mentioned here. The (fivefold) duplication of the pairs $x_i\,, \gamma_i$ keeps the external conditions (design characteristics) in the simulation as close as possible to those in the actually available data. These conditions need not be representative for the external conditions prevailing in the original, unknown source data base from which the observed data is drawn by biased selection according to the RCRS model. For since the z-scores generally are correlated with $x_i\,, \gamma_i$ under the RCR model, biased selection of z-scores may well produce a biased sample of $x_i\,, \gamma_i$ pairs. On the whole this means that the above simulation procedure mimics the RCRS model in an internally correct manner (regarding the

[19] Recall that κ is determined only up to a common proportionality factor in the RCRS model. This factor, chosen here as $0.9/\max_j w_j$, should be as large as possible in order to avoid an unnecessarily high rejection rate. On the other hand, the maximal retain probability should be strictly less than 1 – here it is 0.9 – so that for *all* cases there is some chance of getting rejected. Hence the above rule.

relation between covariates, experimental, and selection effects). However, it does so under the external conditions of the available sample and not necessarily those of the source data.

Acknowledgments

The author wishes to thank: Roger Nelson for making available the Technical Report by Radin and Nelson (1987), which contains the data used in this paper; York Dobyns for providing the preprint by Dobyns (2003) prior to publication; the referees and Roger Nelson for their detailed criticism and valuable comments; and Harald Atmanspacher for editorial assistance.

Note

Readers interested in the numerical algorithms used in this work are welcome to contact the author at `ehm@igpp.de` for more detailed information and copies of MATLAB m-files.

References

Atmanspacher H. and Jahn R.G. (2003): Problems of reproducibility in complex mind-matter systems. *Journal of Scientific Exploration* **17**, 243–270.

Berger R.E. (1986): Psi effects without real-time feedback using a PsiLab video game experiment. In *Proceedings of the 29th Annual Convention of the Parapsychological Association.* The Parapsychological Association, Rohnert Park, 109–128.

Coleman T., Branch M.-A., and Grace A. (1999): *Optimization Toolbox for Use with MATLAB. User's Guide, Version 2,* The MathWorks, Natick.

Copas J. (1999): What works? Selectivity models and meta-analysis. *Journal of the Royal Statistical Society, Ser. A* **162**, 95–109.

Copas J. and Eguchi S. (2001): Local sensitivity approximations for selectivity bias. *Journal of the Royal Statistical Society, Ser. B* **63**, 871–895.

Copas J. and Li H.G. (1997): Inference for non-random samples (with discussion). *Journal of the Royal Statistical Society, Ser. B* **59**, 55–95.

Cox D.R. and Reid N. (1987): Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, Ser. B* **49**, 1–39.

Dear K.B.G. and Begg C.B. (1992): An approach for assessing publication bias prior to performing a meta-analysis. *Statistical Science* **7**, 237–245.

DerSimonian R. and Laird N. (1986): Meta-analysis in clinical trials. *Controlled Clinical Trials* **7**, 177–188.

Dobyns Y.H. (2003): Statistical consequences of data selection. Unpublished manuscript.

Dunne B.J. (1998): Gender differences in human/machine anomalies. *Journal of Scientific Exploration* **12**, 3–55.

Egger M., Smith G.D., Schneider M., and Minder C. (1997): Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal* **315**, 629–634.

Hedges L.V. (1992): Modeling publication selection effects in meta-analysis. *Statistical Science* **7**, 246–255.

Hedges L.V. and Olkin I. (1985): *Statistical Methods for Meta-Analysis*, Academic Press, San Diego.

Huber P.J. (1981): *Robust Statistics,* Wiley, New York.

Iyengar S. and Greenhouse J.B. (1988): Selection models and the file drawer problem (with discussion). *Statistical Science* **3**, 109–135.

Jahn R.G. (1982): The persistent paradox of psychic phenomena: An engineering perspective. *Proceedings of the IEEE* **70**, 136–170.

Jahn R.G., Dunne B.J., Nelson R.D., Dobyns Y.H., and Bradish G.J. (1997): Correlations of binary sequences with pre-stated operator intention: A review of a 12-year program. *Journal of Scientific Exploration* **11**, 345–367.

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

McCullagh P. and Nelder J.A. (1989): *Generalized Linear Models,* Chapman & Hall, London.

Radin D.I. (1997): *The Conscious Universe: The Scientific Truth of Psychic Phenomena,* Harper Edges, San Francisco.

Radin D.I. and Nelson R.D. (1987): Possible interaction between intention and random physical systems, Technical Report Nr. 87001, Princeton Engineering Anomalies Research.

Radin D.I. and Nelson R.D. (1989): Evidence for consciousness-related anomalies in random physical systems. *Foundations of Physics* **19**, 1499–1514.

Radin D.I. and Nelson R.D. (2002): A meta-analysis of mind-matter interaction experiments from 1959 to 2000. In *Healing, Intention and Energy Medicine*, ed. by W.B. Jonas and C.C. Crawford, Churchill Livingstone, Edinburgh, pp. 40–48. Also available at `www.boundaryinstitute.org/articles/mgma.pdf`

Rosenthal R. (1979): The "file drawer problem" and tolerance for null results. *Psychological Bulletin* **86**, 638–641.

Rubin D.B. (1996): Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association* **91**, 473–489.

Schmidt H. (1970): A PK test with electronic equipment. *Journal of Parapsychology* **34**, 175–181.

Schmidt H. (1975): Toward a mathematical theory of psi. *Journal of the American Society for Psychical Research* **69**, 301–319.

Steinkamp F., Boller E., and Bösch, H. (2002): Experiments examining the possibility of human intention interacting with random number generators: A preliminary meta-analysis. In *Proceedings of 45th Annual Convention of the Parapsychological Association*, The Parapsychological Association, Paris, pp. 256–272.

Thompson S.G. and Pocock S.J. (1991): Can meta-analysis be trusted? *Lancet* **338**, 1127–1130.

Utts J. (1991): Replication and meta-analysis in parapsychology (with discussion). *Statistical Science* **6**, 363–403.

Received: 23 December 2003
Revised: 23 September 2004
Accepted: 04 January 2005

Reviewed by York Dobyns/Robert Jahn, Martin Schumacher and another, anonymous, referee.