Limitations to Empirical Extrapolation Studies: The Case of BMD ratios

Kevin P. Brand Paul J. Catalano James K. Hammitt
Lorenz Rhomberg John S. Evans

NRCSE

Technical Report Series

NRCSE-TRS No. 052

July 13, 2000

The **NRCSE** was established in 1996 through a cooperative agreement with the United States Environmental Protection Agency which provides the Center's primary funding.

EPA

Limitations to Empirical Extrapolation Studies: The Case of BMD ratios

By:

Kevin P. Brand[1,*], Paul J. Catalano[2,3],
James K. Hammitt[2,4], Lorenz Rhomberg[4,5]
and John S. Evans[2,4]

[1] Epidemiology & Community Medicine
U. of Ottawa, Faculty of Medicine
Ottawa, Canada

[*] **Correspondance to:** Institute of Population Health, 441 Smyth Road
Ottawa, Ontario K1H 8M8

[2] Harvard School of Public Health

[3] Dana Farber Cancer Institute

[4] Harvard Center for Risk Analysis
Boston, MA, USA, 02115

[5] Gradient Corporation, Cambridge MA, USA

Submitted to Risk Analysis (DO NOT CITE)

June, 2000

Abbreviated title: Are BMD Ratios Informative ?

Abstract

Extrapolation relationships are of keen interest to chemical risk assessment where they play a prominent role in translating experimentally derived (usually in animals) toxicity estimates into estimates more relevant to human populations. A standard approach for characterizing each extrapolation relies on ratios of pre-existing toxicity estimates. Applications of this "ratio approach" have overlooked several sources of error. We examine the case of ratios of benchmark doses (BMDs), and try to better characterize their informativeness by modeling the process by which they are generated in practice. Both closed form and simulation-based models of this "data-generating process" (DGP) are developed, paying special attention to the influence of experimental design. Our results show significant limits to informativeness, and revealing dependencies. The imprecision and bias of the ratio approach can no longer be ignored. We recommend bootstrap techniques, but they alone can only gauge imprecision. Proper characterizations of informativeness require more complicated techniques. Other recommendations are provided for better estimating and/or mitigating the errors. This analysis has implications that extend beyond the ratio approach to any empirical extrapolation study (involving quantal data). Such insights demonstrate the benefits of understanding the DGP and the advantages of using the notion of calibration in better characterizing informativeness.

Keywords: Noncancer risk assessment, relative potency, extrapolation, MLE, likelihood, sampling distribution, experimental design, uncertainty analysis, calibration, censoring, ignorability, measurement error, bootstrap, Monte-carlo simulation, dynamic optimization.

1 Introduction

Most determinations of chemical safety are based on animal experiments and must invoke extrapolation relationships to derive a toxicity estimate relevant to human populations. Consequently, approaches for characterizing each extrapolation relationship are of keen interest. A standard approach tries to discern each relationship from a dataset formed from ratios of pre-existing toxicity estimates. For example Baird et al. [1] found 51 chemicals for which ratios of chronic/subchronic toxicity could be formed, and used them to study subchronic to chronic extrapolation. Several others have applied the same empirical approach to look at this and other extrapolations (e.g., between species, routes of administration, and effect endpoints).

Although sensible[1], this "ratio approach" can be undermined by several sources of error. Owing to the relatively small number of chemicals for which ratios can be assembled, inferences are subject to finite sample error. Moreover, measurement error affects each of the estimates forming the ratios. Additional error can arise due to the patchwork of experimental conditions under which the various toxicity estimates were obtained. To date, applications of the approach have overlooked these errors.

This paper focuses on better characterizing informativeness. It builds upon a previous paper [2], which examined ratios of no-observed-adverse-effects-levels (NOAELs) — a toxicity estimate used in noncancer risk assessment. Here we examine ratios of a different toxicity estimate, the benchmark dose (BMD). This alternative estimator is widely thought to have better statistical properties than its NOAEL counterpart [3, 4, 5, 6, 7]. A motivating question is whether BMD ratios can transcend the substantial limitations found for NOAEL ratios [8, 2].

Our approach involves modeling the way BMD ratios are generated in practice and then using the model to study the impact of the aforementioned errors. For illustration, the analysis focuses on the problem of extrapolating between mouse and rat toxicity. The intent, however, is to be instructive for the more general use of BMD ratios.

We pay special attention to dose selection, a fundamental aspect of experimental design

[1]The approach presumes that the true underlying extrapolation relationship is well represented by a proportionality constant with some random variation (perhaps due to inter-chemical heterogeneity). Ratios are inappropriate for studying other types of relationships.

that can influence the quality of each BMD estimate. The practice of dose selection is an art, involving a certain degree of guesswork about the doses at which the agent in question will yield the most informative responses, and it is unlikely to be optimal across all experiments. To explore this complicated issue, we offer two simplified models of dose selection and examine their implications.

The paper begins with a background section which: i) summarizes the process by which BMD ratios are generated; ii) specifies the restricted form of extrapolation relationship considered in this analysis; and iii) introduces the framework we use to define informativeness. Next, the methods section introduces an idealized model of the data-generating process (DGP) for BMD ratios and then proceeds to describe assumptions needed for quantitative modeling. Our simulation approach is then described. The results section begins with the idealized case, comparing closed-form and simulation based results, and then presents results for a less idealized model of the DGP. Finally, results are interpreted, wider implications are identified, and recommendations are made for improving future extrapolation studies.

2 Background

Before examining the ability of measurements to reveal the underlying property, it is useful to clarify: i) the nature of the measurements; ii) the nature of the underlying property being measured; and iii) the notion of informativeness used in gauging correspondence between the two.

2.1 The Measurement and its Data-Generating Process

Fig 1 illustrates a hypothetical application of the ratio approach. It examines the case of mouse to rat extrapolation. After mouse and rat toxicity estimates are gathered for a set of chemicals, the paired (by chemical) data are checked for consistency with the assumption of a proportional relationship (see Fig 1A). Upon confirmation, ratios are calculated from the paired data to obtain the empirical distribution shown in Fig 1B. Its central tendency (solid vertical line) is used to infer the systematic difference in toxicity between mice and rats, and

its spread (variance) is used to infer the inter-chemical heterogeneity of that difference. In practice, geometric measures of central tendency (Geometric Mean, GM) and spread (Geometric Standard Deviation, GSD) are preferred, because of the consistency of observed ratios with lognormality and their restriction to non-negative values.

2.1.1 The Data Generating Process

In modeling how the empirical distribution (Fig 1B) arises in practice, we face several challenges. The Data-Generating Process (DGP [9, 10, 11]), entailing any factor (step) capable of influencing the individual BMD estimates (typically these are measured independently), can be complicated. We only consider three key steps:

1. The influence of experimental design. Dose selection is our primary concern, but we also consider other aspects such as the number of animals per study.

2. The BMD estimation procedure(s) and in particular the protocol(s) for dealing with atypical bioassay outcomes. For example, stipulating what estimate, if any, is to be recorded for a dataset showing no significant dose-response trend.

3. The "measurement error" affecting each BMD estimate. Specifically we model the impact of sampling error that occurs when only a finite number of animals are available to gauge the underlying *expected* response at each test dose.

These three steps are included in our model of the DGP, which is then used to study the informativeness of BMD ratios. Modeling the first two steps is particularly challenging because each is affected by traditions, precedents, or regulatory standards, which may go undocumented and may be inconsistent across BMD estimates. A more complete discussion of the DGP for BMD ratios is provided in Ref [8].

2.1.2 The BMD Estimate

A BMD can summarize either continuous (e.g., average body-weight change) or quantal (e.g., fraction of animals surviving) experimental outcomes, and can reflect either "best" or lower-bound estimates [4, 12, 13, 6]. We restrict attention to best estimates (Maximum Likelihood

Estimates, MLE) obtained from quantal data and denote them using the standard notation, EDq; where q denotes the quantal level response of interest (a value of 10%, 5% or 1%, is typical). Focusing on the mouse/rat example, we use the random variable θ to denote the pertinent ratio of BMDs,

$$\theta = \frac{\hat{EDq}_m}{\hat{EDq}_r} \tag{1}$$

where the subscripts m and r are used to index mouse and rat attributes and the $\hat{EDq}$'s, being estimates (as denoted by the hats), are random variables.

We are interested in both asymptotic and finite-sample properties of θ. The asymptotic, or large-sample, properties suggest the best that one can do as sample size (number of ratios) becomes effectively infinite. In reality samples are finite; thus finite sample error is of interest.

2.2 The Property

The property is an underlying biological relationship, presumed to exist between the two toxicity endpoints of interest (mice and rats in our case). A standard practice assumes a proportionality relationship,

$$\mathrm{ED}_{q_m} = \Theta \mathrm{ED}_{q_r} \tag{2}$$

where ED_q denotes the true effect-dose, and Θ, the "constant of proportionality," defines the property.

The property, Θ, is conceptualized as a random variable, representing the net result of a systematic difference in toxicity between mice and rats as well as inter-chemical heterogeneity in that difference. Systematic differences could accrue from the gross differences between species (e.g., rats are roughly 10-fold larger by mass), while inter-chemical heterogeneity could result if the chemicals invoke different toxicokinetic and or toxicodynamic mechanisms. In practice doses are often scaled, that is expressed in units which are intended to adjust for the systematic difference, so the observed systematic effect represents only that which is unaccounted for by the chosen scaling rule.

Because the interchemical heterogeneity is conceptualized as arising from multiplicative

deviations from the norm, a lognormal distribution is a sensible choice for representing Θ.

$$\Theta = \frac{\mathrm{ED}_{q_m}}{\mathrm{ED}_{q_r}} \sim \Lambda(\mathrm{GM}_\Theta, \mathrm{GSD}_\Theta) \tag{3}$$

where GM_Θ denotes the geometric mean, and GSD_Θ denotes the geometric standard deviation. These parameters correspond to the systematic effect and inter-chemical heterogeneity, respectively.

When the mouse and rat dose-response relationships are parallel, the property Θ applies to all effect levels, i.e., to any choice of q (See Ref [14, 8] for more detail). That is,

$$\Theta = \frac{\mathrm{ED}_{10_m}}{\mathrm{ED}_{10_r}} = \frac{\mathrm{ED}_{50_m}}{\mathrm{ED}_{50_r}} = \frac{\mathrm{ED}_{q_m}}{\mathrm{ED}_{q_r}}$$

We restrict our analysis to this special case of parallelism, noting that it is a standard assumption in research on comparative potency [15, 16, 17, 14, 18, 19], but further noting that the implications of non-parallelism deserve closer attention in future research.

2.3 Defining Informativeness by Analogy to Calibration

A basis for comparing the property Θ, with its' measurement, θ, is needed. Rather than comparing distributions directly, it is more convenient to compare their associated parameters (GM_Θ versus GM_θ and GSD_Θ versus GSD_θ).

We use the concept of calibration to evaluate informativeness. Calibration usually involves: i) engineering a series of standards (e.g., standard weights in gravimetric analysis) with a known value of the quantity being measured, ii) subjecting each standard to replicate measurements; and iii) plotting the replicate measurements, or some summary thereof, against their associated standards. This yields a calibration curve, well suited for conveying the concept of "informativeness," which we define as connoting not only accuracy and precision but also how each depends on the magnitude of the underlying standard (truth) [8, 2]. Our analysis uses a similar process to gauge the *informativeness* of an empirical distribution of $\hat{EDq}$ ratios, where the modeled "measurement device" encompasses the entire data-generating-process (DGP).

Figures 2a through 2c illustrate three potential calibration results. In Fig 2a, the device being calibrated is both accurate (unbiased) and precise (repeatable). In contrast, Fig 2c

exemplifies an uninformative device (despite good precision) — i.e., the measurements do not respond to changes in the standard. Fig 2b, illustrating a case where bias (and precision) depend on the truth, exemplifies the need for the more rigorous definition of informativeness.

An actual calibration is impossible for various reasons, not the least of which is the unavailability of standards (i.e., strains of mice and rats for which Θ is either known or can be engineered). Modeling the DGP can partially bridge the impasse. Ideally a statistical model can express the distribution of observations as a function of the truth. More generally, simulation can be used. Here Monte-Carlo sampling is applied to simulate replicate measurements — the fundamental element in a calibration exercise. A sufficiently large number of replicates (i.e., in our case a large number of simulated ratios) allows a large-sample approximation of $f(\theta)$, the distribution of observed $\hat{EDq}$ ratios. A calibration curve is built by re-computing a large-sample $f(\theta)$ for each of a sequence of standards (specifications of GM_Θ and GSD_Θ). Thus, the procedure for constructing a simulated calibration curve is identical to conventional procedures, the only difference being that replicate measurements are generated by simulation rather than actual experiment.

3 Methods

We use closed-form and simulation based approaches to model the DGP. The closed-form approach is examined first. Then, the assumptions necessary for quantitative modeling are introduced and finally the more generally applicable simulation approach is described.

3.1 The Idealized Data Generating Process

Under certain idealized conditions, a closed-form expression can be derived for the large sample GM or GSD estimate [8]. The modeled bioassays (mouse and rat) must be "identical" to each other up to a scalar transformation in dose (applied to their test doses and underlying dose-response curves). Under these conditions, Eq 1 can be re-expressed as,

$$\theta = \frac{u_m}{u_r}\Theta \tag{4}$$

where Θ is the property of interest and u_m and u_r represent $\hat{EDq}$s conditional upon a standardized set of conditions (i.e., experimental design and dose response attributes). The standardization implies that u_m and u_r are independent and identically distributed (*iid*) with one another. The large sample GM_θ can therefore be expressed as,

$$\begin{aligned} \mathrm{GM}_\theta &= \exp\left(\mathrm{E}[\log\left(\Theta\right) + \log\left(u_m\right) - \log\left(u_r\right)]\right) \\ &= \mathrm{GM}_\Theta \end{aligned} \tag{5}$$

where, $\mathrm{E}[\log\left(u_m\right) - \log\left(u_r\right)] = 0$, because $u_m \overset{iid}{\sim} u_r$. The expression reveals that under idealized conditions large sample estimates of the systematic effect are unbiased.

Similarly, applying the variance operator and appropriate transformations to Eq 4 obtains an expression for the large sample GSD_θ.

$$\begin{aligned} \mathrm{GSD}_\theta &= \exp\left(\sqrt{\mathrm{Var}[\log\left(\Theta\right) + \log\left(u_m\right) - \log\left(u_r\right)]}\right) \\ &= \exp\left(\sqrt{\log^2\left(\mathrm{GSD}_\Theta\right) + 2\left(\log^2\left(\mathrm{GSD}_u\right)\right)}\right) \end{aligned} \tag{6}$$

where $\mathrm{GSD}_u = \exp\left(\sqrt{\mathrm{Var}[\log\left(u_m\right)]}\right) = \exp\left(\sqrt{\mathrm{Var}[\log\left(u_r\right)]}\right)$. Equation 6 reveals that large sample estimates of the inter-chemical heterogeneity (GSD_θ) are upwardly biased. The magnitude of the bias depends on GSD_u, which essentially represents the measurement error associated with the standardized bioassay.

Both Eq's 5 and 6 provide important qualitative insight[2], however, only the expression for GM_θ (Eq 5) provides complete quantitative insight for large sample results (i.e., since additional analysis is required to quantify GSD_u). In addition, these expressions do not address finite sample imprecision. To get more complete insight and to examine less idealized DGPs, we turn to simulation.

3.2 A Simulation Approach

This section begins by outlining several categories of assumptions underlying our simulation analysis. Although logically unrelated to the property of interest, Θ, and therefore theoretically exogenous, these assumptions affect the 'lens through which we view' the property and

[2]It can be shown that these expressions also apply to NOAEL ratios (if the same idealized conditions hold).

therefore must be specified to fully characterize the DGP. Two factors including dose selection and the BMD estimation protocol, are given special attention. Several others, including dose-response and experimental design factors, are referred to collectively as 'contextual factors' and examined first.

3.2.1 Dose-Response Relationship

The response data used to estimate each $\hat{EDq}$ must be generated from the underlying dose-response relationship. A three-parameter Weibull model is used to represent this relationship for both mice and rats. The model expresses the expected response (fraction of animals with adverse effects), denoted R, as

$$R(d) = \Pr[\text{ adverse effect } |d] = 1 - \exp\left(-\alpha - \log(2)\left[\frac{d}{\mathrm{ED}_{50}}\right]^k\right) \tag{7}$$

where d is the dose, α corresponds to a background response rate, ED_{50} is a 'location' parameter, k determines dose-response shape, and $\log()$ is to the base e throughout.

The inclusion of α in Eq 7 allows for nonzero background response. We assume it to be equivalent in both species ($\alpha_m = \alpha_r$).

Adjusting the location parameter (ED_{50}) fixes the 50% response of the underlying dose-response curve, enabling the dose-response curve to be shifted. Adjusting the shape parameter k achieves varying degrees of non-linearity — as k is increased (for $k > 1$) the dose-response curve becomes increasingly sub-linear.

As described earlier (Section 2.2) parallel dose-response relationships are assumed throughout this analysis. The restriction, implying equivalent dose response shapes (i.e., $k_m = k_r$), should favor informativeness.

In this analysis, only the relative location of the mouse and rat dose-response relationships matters. Thus for the purposes of analysis, the rat ED_{50} is treated as the frame of reference and assigned an arbitrary value ($\mathrm{ED}_{50r} = 1$). With this definition, Eq 2 dictates that $\mathrm{ED}_{50m} = \Theta$.

3.2.2 Experimental Design Factors

Two aspects of experimental design have the potential to significantly influence $\hat{EDq}$'s and associated ratios:[3] the number of animals in each dose group and dose selection.

We assume that the number of animals (denoted n_a) is the same in each dose group for both mice and rats. For the base case, calculations are run with $n_a = 20$ animals.

A mouse (or rat) bioassay is assumed to consist of a control group (with dose $d_0 = 0$) and a set of n_d dosed groups arranged symmetrically on the logarithmic scale around their central test dose, denoted d_c. Two further assumptions are required to fully specify the test doses. These include i) the dose spacing, denoted s, which we define to be the geometric spacing between the highest and central test doses (d_{max}/d_c), and ii) the general strategy for centering the test doses in each of the bioassays. The latter amounts to specifying d_c for both mouse and rat bioassays.

Mouse and rat bioassays are assigned equivalent n_d and s. Plausible values for these parameters are readily based on general practice [21, 22, 23, 24, 25, 13].

The issue of dose centering is more complicated. Any model of this process must reflect not only the general strategy used by the experimenter to center test doses (for each chemical), but also its success. We offer two simplified models. The first, referred to as "enlightened dose centering," was implicitly assumed in Section 3.1's derivations for the idealized DGP. The second, referred to as "default dose centering," is intended to examine the implications of less optimal dose centering, more typical of practice.

Both models make use of a working definition for ideal dose centering wherein the experimenter is assumed to place the central test dose, d_c, exactly at the tested chemicals' underlying ED_{50} (i.e., $d_c = \text{ED}_{50}$)[4]. The enlightened model presumes ideal centering in both mouse and rat bioassays (across all chemicals).

The default model presumes ideal centering only for the rat bioassays. For mouse bioassays experimenters are assumed to set $d_{c_m} = d_{c_r}$ which would seem a sensible recourse when there

[3]Other experimental design decisions such as caging arrangement, feeding protocol, *etc* are well known to affect experimental outcomes [20], but are not considered, to simplify discussion.

[4]We do not claim that centering test doses exactly at the ED_{50} is optimal (this would depend on other factors such as dose response shape). We simply use it as a reasonable working strategy.

is no chemical specific evidence to prompt otherwise. Being centered around the **rat** ED_{50} (or equivalently d_{c_r}), the mouse doses will be 'misplaced' whenever ED_{50r} is a poor proxy for ED_{50m}. The misplacement will generally involve a random component (since $\mathrm{GSD}_\Theta > 1$), and will involve a systematic component when $\mathrm{GM}_\Theta \neq 1$.

3.3 Replicate datasets

To reflect the impact of measurement error on each $\hat{EDq}$ estimate, we use Monte-Carlo simulation [26]. It involves generating replicate measurements, subject to identical experimental-design and dose-response parameters. Each replicate's outcome is modeled as the number of animals showing a response (e.g., death) in each dose group and is denoted $\vec{x}$, which includes the set of $n_d + 1$ responses (i.e., one per dose group). It is the variation in $\vec{x}$ owing to chance that is responsible for measurement error. This is captured by modeling $\vec{x}$ as, $x_j \sim \mathrm{Bin}(n_a, R_j)$, for $j = 0, 1, \ldots, n_d + 1$. That is, the response in the jth dose group is drawn from a binomial distribution where the "success probability," R_j, is the expected response in the jth dose group determined by evaluating the true dose-response function (Eq 7) at dose d_j. A total of N_{rep} $\vec{x}$'s, is generated per experiment to simulate the impact of chance.

3.4 BMD Estimation

For each replicate, an $\hat{EDq}$ estimate is obtained using maximum likelihood estimation (MLE) to fit a model to the dose-response data. In our analysis, the model form fitted to each simulated dataset is chosen to be identical to the three parameter Weibull model defining the truth (Eq 7). In reality, less compatibility between true and fitted relationships is expected.

Since our analysis requires estimating parameters for a large number of datasets (approximately $50,000$ per large sample estimate) we developed our own algorithm for Maximum Likelihood estimation. The procedure needs to be automated, reliable, and efficient and also needs to allow parameters to be constrained (since standard practice precludes negative Weibull parameters). To do this, the Amoeba Downhill algorithm [26] was modified for non-negativity constraints, and an automated method for specifying starting values for the optimization 'search' was developed [8]. The automated method, which exploits the delta method [27],

yields rapid and reliable convergence [8]. Unlike some analysts, we do not constrain the shape parameter to be greater than 1 (some restrict $k \leq 1$ so as to preclude supra-linearity [4]).

Once the constrained MLE parameters are obtained, the fitted model is used to back-calculate the dose associated with an extra risk of q percent [4, 29]. Defining extra risk as, $(R(d) - \alpha)/(1 - \alpha)$, $\hat{ED}q$, is calculated as,

$$\hat{ED}_q = \left(\frac{\log(1-q)}{\log(1/2)} \right)^{(\frac{1}{\hat{k}})} \hat{ED}_{50}$$

where the hat notation is used to identify MLE parameter estimates. In this analysis, $\hat{ED}q$'s are calculated for an extra risk of 10 %, thereby obtaining an $\hat{ED}_{10}$.

Not all datasets are suitable for standard maximum likelihood estimation. Here, two classes of unsuitable datasets are considered: (a) those failing a test for positive trend (referred to as "non-significant"), and (b) those yielding unbounded MLE parameter estimates (referred to as "unfittable").[5] Any analysis must specify a protocol for dealing with such datasets. We assume that they are dropped from further analysis. This 'drop-protocol,' which has the potential for distorting the DGP, is just one of several possibilities. One might also defer to another related dataset, invoke alternative effect estimation techniques (e.g., NOAEL estimation as suggested in Ref [7]), assign "working values" to the dose-response data so as to avert the difficulty [30, p. 130], or even re-do the experiment, perhaps with different dose levels.

To identify "non-significant" datasets we use a Cochran-Armitage trend test [14]. Datasets either failing to exceed the significance criterion ($p \leq 0.05$) or demonstrating negative trend are dropped from further analysis. Though intended to mimic practice (see for e.g., Barnes et al. [7] who advocate significant trend as a prerequisite for BMD estimation), the screen is also partially motivated by pragmatic concerns associated with convergence of our MLE algorithm.

The retained datasets are then examined to screen out unfittable datasets which are, in the case of fitting Weibull models, characterized by an unbounded MLE for the shape parameter [30, 31, 18]. Unfittable datasets are identified (see Ref [8] for convenient criteria) and dropped from further analysis.

[5]Some guidance documentation for BMD estimation also recommends tests for goodness of fit as a prerequisite for obtaining the $\hat{ED}q$ estimates. We have not implemented this screen.

3.5 Methods for Simulating a Calibration

Given a particular set of specified "contextual factors," calibration curves are plotted to explore informativeness. The curves are constructed by simulating replicate bioassay experiments for each calibration standard (i.e., specification of the true distribution of Θ) and tracing out the DGP steps for each replicate. With a sufficient number of replicates, the outcomes provide a large-sample approximation of $f(\theta)$. These distributions, which correspond to the schematic distributions appearing in Fig 2, play a fundamental role in the construction of our calibration curves.

Because each $f(\theta)$ is a function of the distribution of $\hat{EDq}$s observed in mice and in rats, we begin by describing the generic procedure for obtaining a Monte-Carlo approximation of $f(\hat{EDq}|\text{ED}_{50})$; where the distribution is conditional on a fixed ED_{50}.

3.5.1 Sampling distribution for BMD ($\hat{EDq}$)

The procedure for computing $f(\hat{EDq}|\text{ED}_{50})$ involves: (i) generating N_{rep} replicate datasets, $\vec{x}$ as described in Section 3.3; (ii) discarding all replicates that are deemed unsuitable (see Section 3.4); and (iii) estimating the EDq (Section 3.4) for each of the N'_{rep} retained datasets. The resulting N'_{rep} estimates form the Monte-Carlo approximation of $f(\hat{EDq}|\text{ED}_{50})$.

The sample distribution $f(\hat{EDq}|\text{ED}_{50})$, corresponding to the case of $\text{ED}_{50} = 1$ is of particular interest. We refer to it as the "unit distribution," denoting it $f(u)$. It corresponds to the $\hat{EDq}$ sample distribution modeled for rats (since $\text{ED}_{50r} = 1$). In the case of enlightened dose centering, it also applies (with simple scalar transformations) to all mouse bioassays.

3.5.2 Assembling the Sampling Distribution for Ratios

The sampling distribution for $\hat{EDq}$ ratios, $f(\theta)$, can be obtained by combining the mouse and rat $\hat{EDq}$ sampling distributions. The former, $f(\hat{EDq}_r)$, is simply the aforementioned unit distribution, $f(u)$.

Obtaining the mouse distribution, $f(\hat{EDq}_m)$, is more complicated because there is no single ED_{50_m}. Instead the inter-chemical heterogeneity specified in our model implies a distribution for ED_{50_m} consistent with Θ. The mouse distribution, $f(\hat{EDq}_m)$, requires "averaging" condi-

tional distributions $f(\hat{ED}q_m|\Theta)$ across Θ, where the distribution of Θ was given in Eq 3. This is implemented using a statistical procedure called Monte-Carlo integration. The result is a numerical approximation of $f(\hat{ED}q_m)$ consisting of a set of $\sum_{i=1}^{N_{sim}} N'_{rep_i}$ realizations. Again the prime notation indicates that N_{rep} may be reduced after discarding unsuitable datasets.

Once the sampling distributions for the mice and rats have been approximated, their respective realizations are combined using *independent* resampling (N_{resamp} times with replacement) to approximate $f(\theta)$. Independent resampling is appropriate because of the way every rat bioassay is modeled; each is standardized (i.e., $\mathrm{ED}_{50r} = 1$) and consequently their outcomes are independent of one another as well as their counterpart mouse outcomes.

All "observed" summary statistics are computed from the approximation of $f(\theta)$. Large sample estimates of central tendency (GM) and spread (GSD) are readily computed using standard methods, while bootstrap techniques are applied to study the finite sample imprecision in either estimate (detail is available [8, 2]).

Throughout the analysis, choices of simulation size (N_{rep}, N_{sim}, and N_{resamp}) are made with the objective of getting a sufficiently stable approximation of the large-sample distribution; where stability is assessed informally by inspecting repeated analyses.

Simulation, computation, and graphics were programmed in S-PLUS V3.4.1 for SunOS 5.3 (StatSci Division, MathSoft, Inc., Seattle, WA, USA [32, 33]). The modified Amoeba Downhill algorithm was implemented in Fortran and linked to Splus. See Ref [8] for details.

4 Results

We examine the informativeness of GM and GSD estimates by presenting results in the form of calibration curves. The curves plot simulated observations versus their counterpart truths (standards). Here, each observation represents either a GM or a GSD estimate, while the specified value for its counterpart (GM_Θ or GSD_Θ respectively) acts as the standard.

In order to convey informativeness both large and finite sample results are plotted. Large sample estimates are plotted using +s, while the imprecision associated with a more realistically sized sample ($n_{ratio} = 50$) is examined using 95 % confidence intervals (plotted using

triangle symbols).

To help in gauging bias, a one-to-one reference line is plotted on each graph. It represents an ideal calibration curve, i.e., the case for which the estimate perfectly tracks the truth. A calibration curve falling on this line suggests an unbiased estimator; an implicit assumption when GM or GSD estimates are taken at face value. Curves falling off this line may still be informative provided they show a trend which is not obscured by the finite sample error.

Calibrations for the enlightened and default models are run separately to look at the influence of dose centering. All primary analyses assume a base-case set of contextual factors including, a background rate of $\alpha = 0.05$, a dose response shape of $k = 2$, a number of animals per dose group of $n_a = 20$, a dose spacing of $s = \sqrt{10}$, a number of dosed groups of $n_d = 3$, and (in the case of a GM calibration) a nominal value of $\text{GSD}_\Theta = 2.5$ (or $\text{GM}_\Theta = 1.0$ in the case of a GSD calibration).

4.1 Enlightened Dose Centering

The enlightened case is of interest for two reasons. First, it provides an upper-bound on the informativeness of $\hat{EDq}$ ratios. Second, being subject to the closed-form solutions derived in Section 4.1, it provides an opportunity to cross-check our simulation algorithm.

4.1.1 GM Calibration

The calibration shown in Fig 3 examines the informativeness of GM estimates. The large-sample results (+'s) closely track the one-to-one line and thereby reveal an unbiased estimator; as affirmed by the associated closed-form solution (Eq 5). Evidently, in the enlightened case, a sufficient sample size (number of ratios) can give a good estimate of the systematic effect.

The confidence intervals (triangles) shown in Fig 3, reveal non-negligible imprecision ($n_{ratio} = 50$). Implications of this imprecision can be explored by interpolating between the upper and lower bounds. For example, an observed GM of 3 is roughly compatible with a truth of anywhere between 2.2 and 4.0. Larger intervals would apply for GM > 3, since the intervals increase linearly with increasing GM_Θ. Similar logic can be used to gauge the power to differentiate rival hypotheses about the true GM.

4.1.2 GSD Calibration

Turning to the issue of inter-chemical heterogeneity, Fig 4 examines the informativeness of an observed GSD. The large-sample results (+) demonstrate upwards bias as affirmed by the closed form solution (Eq 6) presented in Section 4.1.

As in the case of GM calibrations, finite sample imprecision can be substantial. The triangle intervals plotted in Fig 4, show that imprecision increases exponentially with the true GSD. Judging from these 95% bounds, an observed GSD of 3.0 is compatible with a true underlying GSD of anywhere between 2.2 and 4.0 (a sizeable range for a GSD), with a corrected central estimate of 3.2.

The results presented in Figures 3 and 4 provide a reassuring cross-check of our simulation algorithm; the large sample results are consistent with closed form expressions presented in Section 4.1. The consistency is of particular interest because we are forced to rely on the simulation algorithm to evaluate the case of default centering (presented next).

4.2 Default Dose Centering

The default model explores the implications of imperfect dose centering. When severe, the imperfection can increase the number of unsuitable datasets, which in turn (by virtue of our drop-protocol) causes the censoring of $\hat{EDq}$'s and associated ratios. The censoring can cause a distinct pattern in the calibration curves; a plateau-effect (i.e., curve flattens out, relative to the one-to-one line), which affects both GM and GSD calibration curves (in the default case), and undermines informativeness.

To conserve space we present results only for GSD calibrations (results for GM calibrations, similar to those for GSDs, are available [8]).

Recognizing the prominent role of censoring we have added a diagnostic bar-plot to each calibration curve. Solid vertical bars are used to indicate the fraction of censoring associated with each observed GSD. It is a rather crude diagnostic because it lumps all censoring together [8]. More sophisticated censoring diagnostics will be explored in future research.

4.2.1 GSD Calibration: Default

Fig 5 shows a GSD calibration corresponding to the case of default centering. Although the large-sample results (+) show a trend with the truth, the trend flattens with higher values of the standard, displaying the aforementioned plateau effect. Notably, informativeness is undermined. For example, Fig 5 shows a limited capacity for resolving truths between 2.6 and 4 — the calibration suggests that truths within this range will yield roughly the same distribution of observed GSDs.

Finite sample error, as depicted by the triangular confidence intervals ($n_{ratio} = 50$), exacerbates the limitations. For example, judging by the confidence intervals, an observed GSD of 2.5 is roughly compatible with any true GSD larger than 1.8.

The bar-plots shown in Fig 5, are suggestive of an association between censoring and the plateau effect. The increase in censoring (ranging from 15% to 50%, left to right) with increasing values of the standard, parallels the onset and progression of the plateau effect.

To better explore the interplay between measurement error and censoring, Fig 5 results are viewed differently. The sample distributions, $f(\theta)$, underlying each of the five GSD plotting points are examined and compared with their underlying standards. Fig 6 plots each sample distribution (on a log-scale) as a histogram, and superimposes a Gaussian distribution corresponding to the lognormally distributed standard (Eq 3). Note the breadth of the Gaussian distributions increases from (A) to (E) reflecting the GSD_Θ increments requisite to a GSD calibration.

A comparison of the histograms with their Gaussian standards reveals the influence of measurement error and censoring. For example, in panel A ($GSD_\Theta = 1.2$), the wider breadth of the histogram (observed ratios) relative to its Gaussian counterpart is indicative of measurement error which has inflated the observed variance. As the GSD_Θ increases, however, the role of censoring increases, culminating in histograms whose tails are noticeably truncated relative to their Gaussian counterparts (see panels D and E). The truncation causes the plateau effect seen in Fig 5.

4.3 Dependence Upon Contextual factors

The results presented above demonstrate that informativeness strongly depends on dose centering. Next we explore the impact of varying the contextual factors.

4.3.1 Enlightened Case

The closed form expressions derived for the enlightened case prove that some estimation properties are independent of contextual factors. Namely, GM and GSD estimates remain asymptotically unbiased and upwardly biased, respectively, regardless of context. However, the extent of GSD bias varies with context as does finite sample imprecision. Both depend on measurement error (i.e., GSD_u) in an intuitive manner.

4.3.2 Default Case

The sensitivity to contextual factors is not as easy to gauge under default centering because the more prominent role of censoring, complicates matters. Predicting sensitivity for a particular contextual factor requires anticipating the net result of its impact on both "measurement error" and censoring; impacts which can be opposing. To illustrate the extent of sensitivity we present one set of results for the informativeness of GSD estimates (results for GM estimates are presented elsewhere [8]).

Fig 7 is obtained by repeating the calibration process; varying the numeric value of two contextual factors (others are maintained at base-case values). Dose response shape (k) is varied down the rows, while the number of animals per dose group (n_a) is varied across the columns. The calibration curves, ranging from reasonably informative (top-right panel) to essentially uninformative (bottom row of panels) and showing some disconcertingly strong plateau effects, demonstrate strong sensitivity to k and n_a. We find comparable sensitivity to the other contextual factors as well (results not shown).

In reality, informativeness would be even worse than appears in Fig 7. Finite sample imprecision, likely to be of similar in extent as in Fig 5, would tend to obscure the trend (if any) in the large-sample results.

The sensitivities displayed in Fig 7 can be explained. Increasing n_a across the columns

(left to right), increases the statistical power of each bioassay, and thereby improves informativeness. The improvement arises from a reduction in measurement error (as indicated by decreasing y-intercept) and from a tempering of the plateau effect (i.e., by reducing censoring).

The diminishing informativeness with increasing k is less intuitive. Given the presumed estimation protocol, increasing k (increasing step-like shape) increases censoring by increasing the fraction of unfittable datasets. The plateau effect ensues.

The bar-plots in Fig 7, affirm the association between censoring and poor informativeness. Less informative curves are characterized by higher censoring (i.e., see bottom row of panels, where associated barplots indicate censoring ranging from 54% to 95%), whereas the most informative curve shows the least (top-right panel indicates censoring ranging from 0.0% to 10%).

Within panels the association is not as pronounced (e.g., panel k5 n20). Further reflection suggests that the weak association may be attributed to the crude nature of our diagnostic. The impact of censoring strongly depends on where (in the uncensored distribution) that censoring occurs, a feature missed by a measure which lumps together all censoring.

The extreme censoring characterizing some panels in Fig 7 (e.g., the bottom row) raises a question. Can panels exhibiting such high censoring have a bearing on actual practice ? It seems implausible that testing programs would persist in the face of so much censoring. The relevance of each panel in Fig 7 to practice, may therefore vary, even though each is a function of plausible contextual factors.

5 Discussion and Conclusions

The ratio approach for characterizing extrapolation relationships overlooks several sources of error which are of interest in the evaluation and improvement of methods for chemical risk assessment. We found the informativeness of the approach to be sufficiently limited to undermine the current practice of taking results at face-value. Future applications of the ratio approach should attempt to quantify and adjust for bias and imprecision. We offer a standard approach for gauging imprecision, and suggest that more complicated techniques

will be required to gauge bias.

Interestingly, BMD ratios do not necessarily outperform NOAEL ratios, as one might first suspect. A preliminary comparison of results for NOAEL ratios [2], and their BMD counterparts, is equivocal [8]. Depending on the context, NOAEL ratios can be more or less informative.

We begin this discussion by summarizing our results. We then carefully scrutinize our model, and conclude that it is basically sound and provides a good starting point for quantifying bias and imprecision. The scrutiny is revealing. It suggests several ways to improve the ratio approach and reveals that our work has wider implications.

Our modeling results show substantial imprecision and bias in GM and GSD estimates. Although the extent of these errors is found to vary with several factors, imprecision is consistently non-negligible.

Distinct differences in calibration curve patterns are observed between the two (default and enlightened) dose centering models. Only default centering exhibits plateau-effects. The resulting potential for downward bias affects both GM and GSD estimates. In contrast, the enlightened case has unbiased GM estimates and only upward bias in GSD estimates.

Factors beyond dose centering also have substantial effects on calibration properties. Fig 7 and additional analyses, reveals sensitivity to all contextual factors. These factors need to be taken into account when interpreting BMD ratios.

Some of our calibration curves (e.g., Fig 7, upper right, i.e., k1 n50) suggest that the ratio approach can be informative. But our results hide an additional source of uncertainty. Each curve presumes that all contextual factors are identifiable (known with reliable precision), but in reality some, e.g., dose response shape (k), are poorly known. This translates into uncertainty about which calibration curve applies, clouding the estimation of bias.

Two calibration curve features, the plateau effect and the propensity for an upwards GSD bias, are explained by censoring and measurement error, respectively. Simplistically, their impacts can be described as subtracting and adding variation, respectively, and thus have a predictable influence on imprecision as well. Censoring plays a more influential role under default centering, while measurement error plays a roughly comparable role under both cen-

tering models. Owing to our drop-protocol, censoring is a direct result of unsuitable datasets, and thus increases with poor centering. Intuitively, measurement error varies inversely with statistical power.

The relevance of our results is contingent on how faithfully we have captured the true DGP. Efforts have been made to use only plausible assumptions, however, some are difficult to verify.

In most cases of ambiguity, we have made assumptions with the intention of favoring informativeness. For example, we assumed that mouse and rat dose response curves are parallel and that the true and fitted dose response relationships are compatible (both Weibull). Although, reality is likely to deviate from these assumptions, it is likely to do so in a way that worsens informativeness. Not all of our assumptions necessarily favor informativeness. Two, including the drop-protocol, and our implicit neglect of multi-comparisons, deserve closer consideration.

The drop-protocol, which automatically censors each unsuitable dataset, seems like a worse-case assumption. How can one do worse than outright censoring? Yet using an alternative protocol, such as reporting the NOAEL (where obtainable) for each unsuitable dataset, would not necessarily improve informativeness. A NOAEL imputation protocol has its own errors [2], and does not work for all datasets. Thus censoring and the attendant plateau effect is likely to exist even if wiser protocols are practiced.

Our analysis has overlooked the issue of multi-comparisons. It implicitly assumes that ratios are formed between specific effect endpoints in the mouse and rat bioassays, and not, as is common, between the most sensitive (or pivotal) endpoints. This added stipulation for "most sensitive" would require additional steps in our model of the DGP. Informativeness may or may not improve, depending on dose centering and how the dose response relationships are related across endpoints. This will be the subject of future research.

In Section 4.3.2 we questioned how much censoring could go unchecked in practice. Some clues can be obtained from past studies. For example, Faustman et al. [13] reported that among $2,000$ reproductive bioassay endpoints (not all independent), 43 % precluded the identification of a Lowest Observable Adverse Effect Level (LOAEL). Indicative of the prevalence

of the nonsignificant class of unsuitable datasets, this suggests that censoring on the order of 50% is conceivable (provided drop-protocol applies). Unfittable datasets, would only add to this percentage. Future work will explore this issue in more detail.

The ratio approach can be improved in several ways. Some actions can improve inferences from pre-existing ratio datasets, while others can assure that more complete information is recorded, or minimize the errors affecting the approach.

The bias and imprecision in GM and GSD estimates must be estimated. A simple solution exists for gauging imprecision: namely bootstrap techniques can be applied to the raw ratio data in much the same way as we did in our simulation. Bias, on the other-hand must be predicted.

Our results demonstrate that bias depends on several factors. Addressing these dependencies requires more information about each of the contributing bioassays. At the very least each BMD should be accompanied by its raw dose response data. This facilitates wiser and/or more consistent BMD estimation protocols. Additional information should be collected to account for influential co-factors. Information is: most readily available for experimental design factors (e.g., n_a, n_d, s); discernible, but with poor precision, for dose response factors; and, generally unavailable for dose centering and BMD estimation protocols.

Predicting bias is complicated. The problem simplifies under enlightened dose centering where only bias in the GSD is of concern (GM estimates are unbiased). In this case, the GSD bias can be determined with the aide of Eq 6, after estimating GSD_u (using either simulation, see Section 3.3, or, if certain asymptotic conditions hold, the delta-method [27]).

Accounting for bias will be more difficult in practice. Imperfect dose centering will complicate matters with the plateau effect. Moreover, the patchwork of contextual factors expected in practice will necessitate averaging calibration properties across the patchwork. New methods, perhaps building upon ours, will be required. Notably, some occasions, namely hypothesis testing, require more detail than simply the imprecision and bias of the estimate. The precision and bias under the null must also be known (or estimated). Detail of this kind requires modeling the DGP and producing a calibration curve or some simplifying assumptions.

Not all recommendations revolve around better characterizing imprecision and bias. Steps

aimed at a more strategic design of future bioassays can mitigate the errors. Measurement-error can be reduced by increasing the statistical power of the bioassays (e.g., increasing n_a, or n_d). This has the added benefit of reducing the number of non-significant datasets and thereby mitigating censoring. Relaxing the significance criteria (p-value) could have a similar effect. But, perhaps the most effective way to mitigate censoring is to aspire towards enlightened dose centering. The feasibility of these recommendations varies. In the case of increasing statistical power, opportunity costs can be prohibitive, while assuring enlightened dose centering, possibly requiring expensive dose ranging studies or sequential designs[6], has, in addition, its practical limits [8].

Although we examined the hypothetical case of mouse to rat extrapolation, our analysis has wider implications. The range of input values and assumptions explored is compatible with most chronic, sub-chronic, and (arguably) acute bioassays. Thus the general assertion, that imprecision and bias can not be overlooked, holds for essentially any application of the ratio approach to quantal data (e.g., to explore route to route, endpoint to endpoint, subchronic to chronic, and other extrapolations). Indeed, the roots of the problem (measurement error and censoring) are fundamental and likely to have broader implications. They are likely to affect most other empirical extrapolation approaches, including 'non-linear' analyses of ratios [38], regression studies extending beyond pair-wise comparisons [39, 40], and other studies exploring relationships between alternative toxicity estimates [41, 42, 43, 44, 45]. They also have implications for studies of optimal design and studies comparing the relative merits of alternative toxicity measures (e.g., BMD versus NOAEL). Such studies need to allow for the inherent unreliability of dose centering and specify what is to be done with unsuitable (e.g., insignificant) outcomes.

Generalize-ability to toxicity estimates drawn from continuous or categorical data is a topic of future research. Our analysis only dealt with quantal data, which may accentuate censoring.

Slob and co-workers [46, 47] have independently pointed out the potential for upwardly

[6]Sequential designs allow progressive re-testing, where each re-test's design doses are enlightened by the responses found in the last iteration [34, 35, 36, 37].

biased GSD estimates and finite sample error, while Faustman (reported in [21]) has expressed other qualitative concerns. Our analysis takes a more quantitative tact, demonstrating that the errors are non-negligible and revealing the key influence of dose centering, unsuitable datasets, and the estimation protocol.

We are not the first to examine the influence of dose selection on extrapolation studies. Several papers have debated whether the correlation observed between mouse and rat cancer potency's, is merely an artifact of a coordinated choice of test doses [48, 49, 50, 38, 51, 52, 53]. The artifact argument is explainable by the plateau effect and thus compatible with our default case. Our approach may help in better establishing the informativeness of the observed correlations.

Understanding the DGP and using a calibration construct to explore its implications has proven useful in this work. The same approach could be applied to other steps in regulatory toxicology, where an actual calibration is often impossible or prohibitive, and where an analyst may be prone to affirming the consequent. Considering the DGP suggests the distribution of observations to be expected under rival hypotheses — not just the one presupposed — and in this way supports more appropriate inference. We hope to apply this same general approach in future research.

ACKNOWLEDGMENTS: Many to come.

References

[1] S. J. Baird, J. C. Cohen, J. D. Graham, A. I. Shlyakhter, and J. S. Evans. Noncancer risk assessment: A probabilistic alternative to current practices. *HERA*, 2(1):79–102, 1996.

[2] K. P. Brand, L. Rhomberg, and J. S. Evans. Estimating noncancer uncertainty factors: Are ratios [of] NOAELs informative ? *Risk Analysis*, 19(2):295–308, 1999.

[3] D. Gaylor. The use of safety factors for controlling risk. *J. Toxic. Environ. Health*, 15:329–336, 1983.

[4] K. S. Crump. A new method for determining allowable daily intakes. *Fund. & Appl. Tox.*, 4:854–871, 1984.

[5] W. Leisenring and L. Ryan. Statistical properties of the NOAEL. *Regul. Tox. & Pharm.*, 15:161–171, 1992.

[6] B. C. Allen, R. J. Kavlock, C. A. Kimmel, and E. M. Faustman. Dose-response assessment for developmental toxicity: II. Comparison of generic BMD estimate with NOAELs. *Fund.& Appl. Tox*, 23:487–495, 1994.

[7] D. G. Barnes, G. P. Daston, J. S. Evans, A. M. Jarabek, R. J. Kavelock, C. A. Kimmel, and H. L. Spitzer. Benchmark dose workshop: Criteria for use of a benchmark dose to estimate a reference dose. *Regul. Tox. & Pharm.*, 21:296–306, 1995.

[8] K. P. Brand. *Interpreting Bioassays for Policy: Analysis of Extrapolation Uncertainties*. Sc.D Thesis, Harvard School of Public Health, May 1999.

[9] D. Freedman, R. Pisani, R. Purves, and A. Adhikari. *Statistics (Second Edition)*. W. W. Norton, 1991.

[10] A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian Data Analysis*. Chapman & Hall, 1995.

[11] A. W. F. Edwards. *Likelihood*. Johns Hopkins Univ. Press, Baltimore, Md, 1992.

[12] D. Krewski, C. Brown, and D. Murdoch. Determining "Safe" levels of exposure: Safety factors or mathematical models. *Fund. Appl. Tox.*, 4:s383–s394, 1984.

[13] E. M. Faustman, B. C. Allen, R. J. Kavlock, and C. A. Kimmel. Dose-response assessment for developmental toxicity: I. characterization of data base and determination of NOAELs. *Fund. & Appl. Tox*, 23:478–486, 1994.

[14] B. J. T. Morgan. *Analysis of Quantal Response Data*. Chapman & Hall, 1992.

[15] D.J. Finney. *Statistical Method in Biological Assay*. Charles Griffen & Co. LTD, 1978.

[16] M. S. Srivastava. Multivariate bioassay, combination of bioassays, and Fieller's theorem (corr: V44 p321). *Biometrics*, 42:131–141, 1986.

[17] C. Cox. Fieller's theorem, the likelihood, and the delta method. *Biometrics*, 28:709–718, 1990.

[18] E. Benton Cobb. Estimation of the Weibull shape parameter in small-sample bioassay. *Journal of Statistical Computation and Simulation*, 31:93–101, 1989.

[19] P. T. Kim, E. M. Carter, J. J. Hubert, and K. J. Hand. Shrinkage estimators of relative potency. *Journal of the American Statistical Association*, 88:615–621, 1993.

[20] J. K. Haseman. Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. *Environmental Health Perspectives*, 58:385–92, 1984.

[21] J. R. Bucher, C. J. Portier, J. I. Goodman, E. M. Faustman, and GW Lucier. Workshop overview. national toxicology program studies: principles of dose selection and applications to mechanistic based risk assessment. *Fundamental & Applied Toxicology*, 31(1):1–8, 1996.

[22] J. L. Counts and J. I. Goodman. Principles underlying dose selection for, and extrapolation from, the carcinogen bioassay: dose influences mechanism. *Regulatory Toxicology & Pharmacology*, 21(3):418–21, 1995.

[23] R. A. Griesemer. Dose selection for animal carcinogenicity studies: a practitioner's perspective. *Chemical Research in Toxicology*, 5(6):737–41, 1992.

[24] J. K. Haseman. Issues in carcinogenicity testing: dose selection. *Fundamental & Applied Toxicology*, 5(1):66–78, 1985.

[25] R. S. Chhabra, J. E. Huff, B. S. Schwetz, and J. Selkirk. An overview of pre-chronic and chronic toxicity/carcinogenicity experimental study designs and criteria used by the National Toxicology Program. *Env. Health Persp.*, 86:313–321, 1990.

[26] W. H. Press et al. *Fortran Numerical Recipes*, volume 1. Cambridge Univ. Press, Cambridge, England, 2nd edition, 1996.

[27] Christopher Cox. An elementary introduction to maximum likelihood estimation for multinomial models: Birch's theorem and the delta method. *The American Statistician*, 38:283–287, 1984.

[28] Gary W. Oehlert. A note on the delta method. *The American Statistician*, 46:27–29, 1992.

[29] US EPA. The use of the benchmark dose approach in health risk assessment. Technical Report EPA/630/R-94/007, US Environmental Protection Agency, February 1995.

[30] J. Berkson. Maximum likelihood and mimimum χ^2 estimates of the logistic function. *JASA*, pages 130–162, March 1955.

[31] S. J. Haberman. *The analysis of frequency data*, volume IV of *Statistical Research Monographs*. U. of Chicago Press, Chicago, 1974.

[32] StatSci Division, MathSoft, Inc., Seattle, WA, USA. *S-PLUS User's Manual: Version 3.3 for Windows*, September 1995.

[33] W. N. Venables and B. D. Ripley. *Modern Applied Statistics With S-Plus*. Springer-Verlag, 1994.

[34] SC Gad, AC Smith, AL Cramp, FA Gavigan, and MJ Derelanko. Innovative designs and practices for acute systemic toxicity studies. *Drug & Chemical Toxicology*, 7(5):423–34, 1984.

[35] A. Whitehead and R. N. Curnow. Statistical evaluation of the fixed dose procedure. *Fd. Chem. Toxic.*, 30(4):313–324, 1992.

[36] V. Diener, L. Siccha, U. Mischke, D. Kayser, and E. Schlede. The biometric evaluation of the acute-toxic-class method (oral). *Arch. Toxicol.*, 68:599–610, 1994.

[37] D. L. McLeish and D. Tosh. Sequential designs in bioassay. *Biometrics*, 46:103–116, 1990.

[38] B. Metzger, E. Crouch, and R. Wilson. On the relationship between carcinogenicity and acute toxicity. *Risk Analysis*, 9(2):169–177, 1989.

[39] C. C. Travis and R. K. White. Interspecific scaling of toxicity data. *Risk Analysis*, 8(1):119–125, 1988.

[40] E. J. Freireich et al. Quantitative comparison of toxicity of anti-cancer agents in mouse, rat hamster, dog, monkey, and man. *Cancer Chemotherapy Reports*, 50(4):219–244, 1966.

[41] Layton DW et al. Deriving allowable daily intakes for systemic toxicants lacking chronic toxicity data. *Regul. Tox. & Pharm.*, 7:96–112, 1996.

[42] H. J. Kramer et al. Conversion factors estimating indicative chronic No-observed-Adverse-Effect levels from short-term toxicity data. *Regul. Tox. & Pharm.*, 23:249–255, 1996.

[43] R. A. Woutersen, H. P. Till, and V. J. Feron. Sub-acute versus subchronic oral toxicity study in rats: Comparative study of 82 compounds. *J. Appl. Toxicol.*, 5(4):277–280, 1984.

[44] C. C. Travis, L. A. Wang, and M. J. Waelmer. Quantitative correlation for carcinogenic potency with four different classes of short-term test data. *Mutagenisis*, 6(5):353–360, 1991.

[45] C. C. Travis et al. Prediction of carcinogenic potency from toxicological data. *Mutation Research*, 241, 1990.

[46] W. Slob and M. N. Pieters. A probabilistic approach for deriving acceptable human intake and human risks from toxicological studies: General framework. *Risk Anal.*, 18(6):787–798, 1998.

[47] T. Vermeire, H. Stevenson, M. N. Pieters, M. Rennen, W. Slob, and B. C. Hakkert. Assessment factors for human health risk assessment: A discussion paper. *Crit. Rev. Toxicol.*, 29(5):439–490, 1999.

[48] L. Zeise, R. Wilson, and E. Crouch. Use of acute toxicity to estimate carcinogenic risk. *Risk Analysis*, 4(3):187– 199, 1984.

[49] L. Zeise, R. Wilson, and E.A.C. Crouch. A possible relationship between toxicity and carcinogenic potency. *J. Amer. Coll. Toxicol.*, 5:137–151, 1986.

[50] E. Crouch, R. Wilson, and L. Zeise. Tautology or not tautology? *J. Toxicology and Env. Health*, 20:1–10, 1987.

[51] L. Bernstein, L. S. Gold, B. N. Ames, M. C. Pike, and D. G. Hoel. Some tautologous aspects of the comparison of carcinogenic potency in rats and mice. *Fund. of Applied Toxicology*, 5:79–86, 1985.

[52] D. A. Freedman, L. S. Gold, and T. H. Slone. How tautological are interspecies correlations of carcinogenic potencies ? *Risk Analysis*, 13(3):265–272, 1993.

[53] D. Krewski, D. W. Gaylor, A. P. Soms, and M. Szyszkowicz. An overview of the report: Correlation between carcinogenic potency and the maximun tolerated dose: Implications for risk assessment. *Risk Analysis*, 13(4):383–398, 1993.

Figure 1: A scatter plot of 50 hypothetical rat $\hat{EDq}$ measurements plotted against their mouse counterparts (on log-log scale) is shown in (A), the histogram of the log-transformed ratios taken from these same data is shown in (B). Under the ratio approach the central tendency (dashed line) of this distribution is thought to be indicative of the systematic difference between mice and rats, while the spread is interpreted as indicative of the interchemical heterogeneity in that difference.

Figure 2: A schematic representation of three calibration curves, illustrating different degrees of informativeness. In each panel 'bell-curves,' depicting large sample distributions, are plotted against their associated standards (truths). Dotted lines mark the one-to-one line, while solid lines track the central trend of the observations. **Panel A** exemplifies an informative measurement device, exhibiting tight centering around the one-to-one line (precise and unbiased). **Panel B** exemplifies the net result of a complicated DGP, leading to a dependency between the attributes of bias and precision and the truth; the plateau-effect displayed by the trend (solid) line undermines informativeness. **Panel C** exemplifies an entirely uninformative device (observations are unrelated to the truth of interest); note apparent precision can be misleading. An implicit assumption of panel A, often goes unchecked.

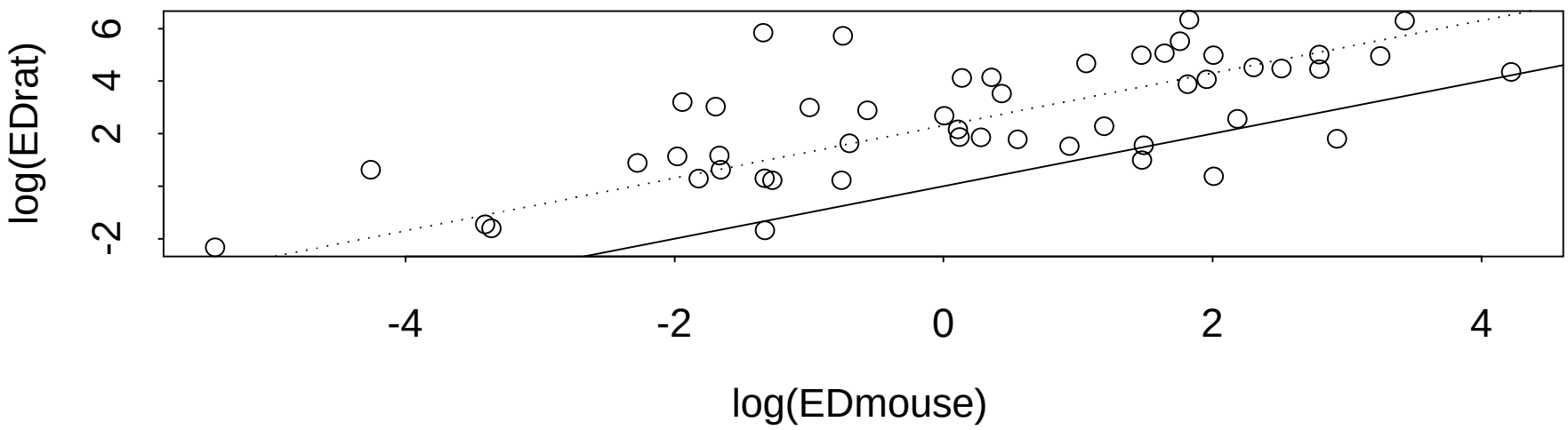
Figure 3: A GM calibration-curve is plotted for the case of enlightened dose centering. Large-sample GM_θ's (**Obs**), obtained via simulation, are plotted against their associated calibration standards, GM_Θ (crosses, +). Triangles denote 95% confidence intervals for GM_θ, calculated assuming 50 ratios. GM_θ refers to the geometric mean of $\hat{EDq}$ ratios taken between mice and rats. An unbiased calibration curve would fall on the one-to-one line (dotted line). The closed-form expression for the large sample GM_θ (Eq 5) corresponds to the one-to-one line. Plot based on the following contextual factors: $s = \sqrt{10}$, $n = 20$, $k = 2$, $\alpha = 0.05$ and $\text{GSD}_\Theta = 2.5$.

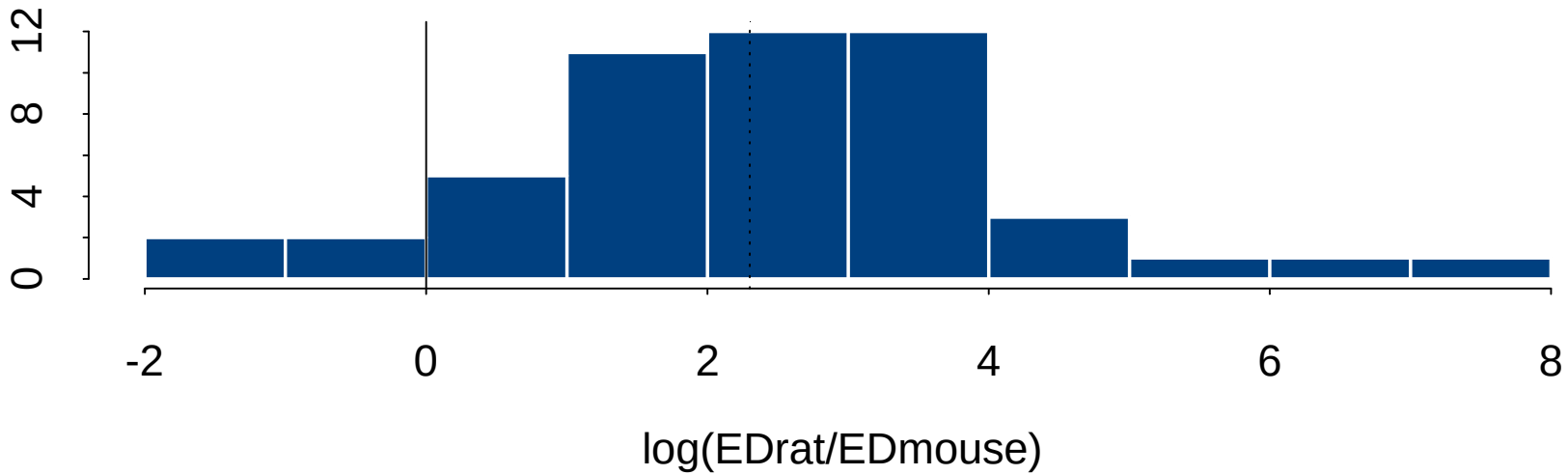
Figure 4: A GSD calibration-curve is plotted for the case of enlightened dose centering. Large-sample GSD_θ's (**Obs**), obtained via simulation, are plotted (crosses, +) against their associated calibration standards, GSD_Θ (**True**). GSD_θ refers to the geometric standard deviation of ratios taken between mice and rats. The closed-form expression for large sample GSD_θ (Eq 6), plotted as the solid line, affirms the simulation results (+s). See Fig 3's caption for an explanation of plotting symbols and contextual factors which are the same except $GM_\Theta = 1$ while GSD varies.

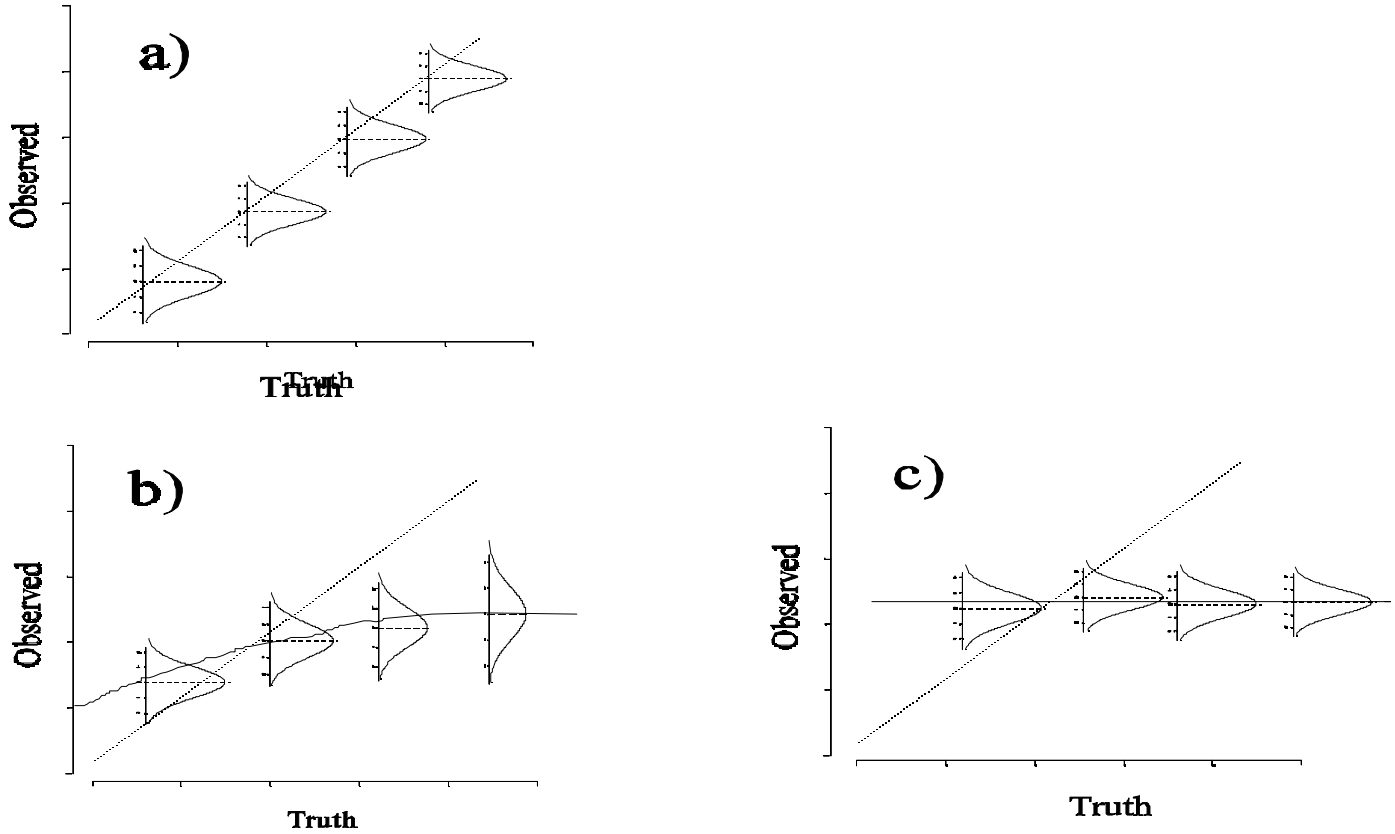
Figure 5: A GSD calibration-curve is plotted for the case of default dose centering. All results are based on simulation. Large-sample GSD_θ's (**Obs**) are plotted (crosses, +) against their associated calibration standards, GSD_Θ (**True**). Upper barplot shows the fraction of datasets censored (fraction of bar darkened) for each standard. See Fig 3's caption for an explanation of plotting symbols and contextual factors which are the same except $GM_\Theta = 1$ while GSD varies.

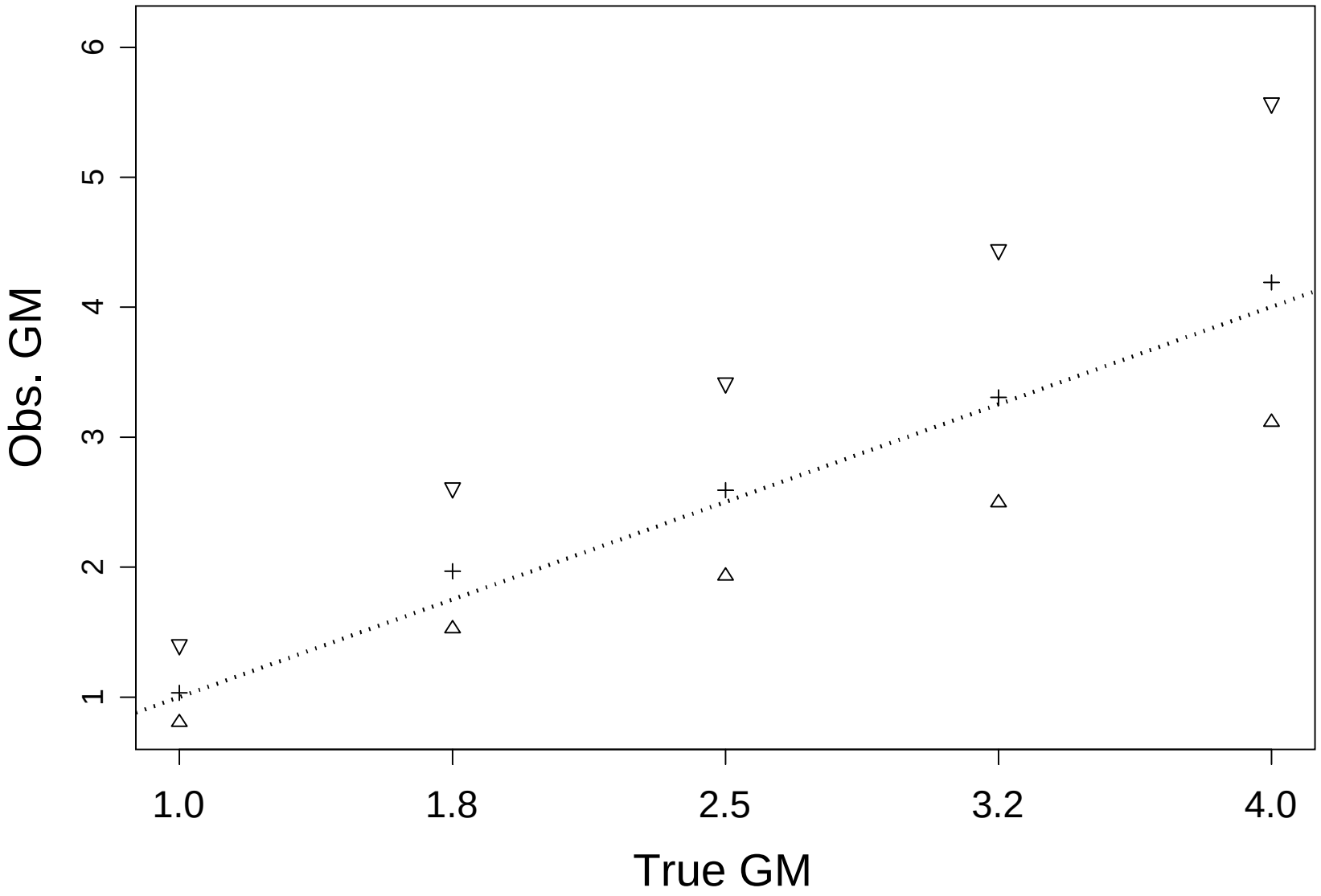
Figure 6: Re-examines the results shown in Fig 5. Using histograms it plots the sample distribution of observed ratios (on log-scale) corresponding to each of the (+) plotting points shown in Fig 5. A Gaussian distribution, representing the relevant standard, is superimposed on each histogram. The true GSD increases from (A) to (E), as evident in the increasing spread. Discrepancies between the histogram and its counterpart truth (Gaussian distribution) indicate censoring, and account for the plateau-effect seen in Fig 5.

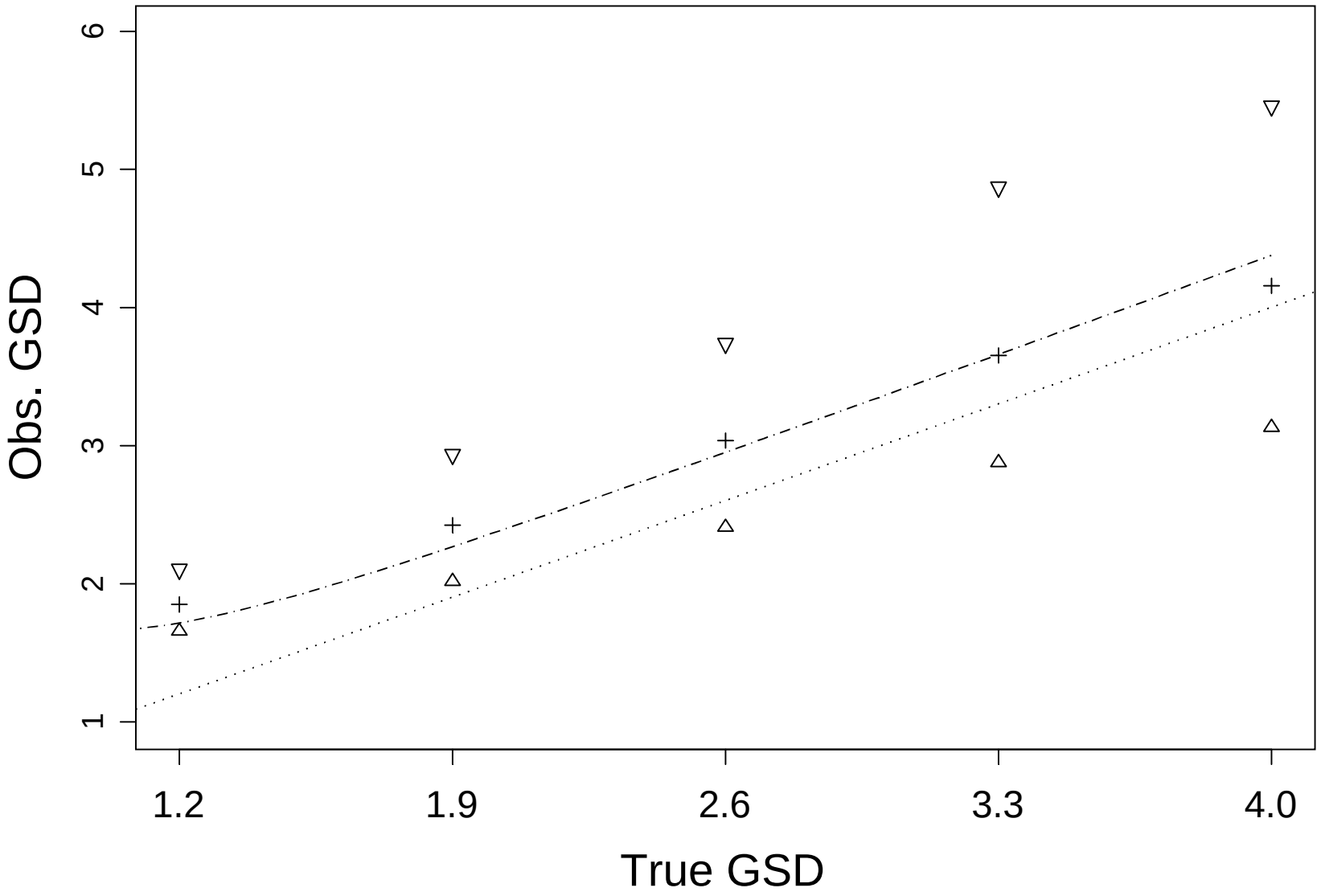
Figure 7: Nine simulated calibrations illustrate the dependence of curves on contextual factors. Dose-response shape k is varied down the rows ($k = 1, 2, 5$), while n_a is varied across the columns ($n_a = 10, 20, 50$). Each calibration curve plots **Observed** GSD_θ's versus their (**True**) calibration standards, GSD_Θ (shown with crosses +'s). All 9 plots based on the following contextual factors: $s = \sqrt{10}$, $\alpha = 0.01$ and $GM_\Theta = 1$. See Fig 3's caption for an explanation of plotting symbols.

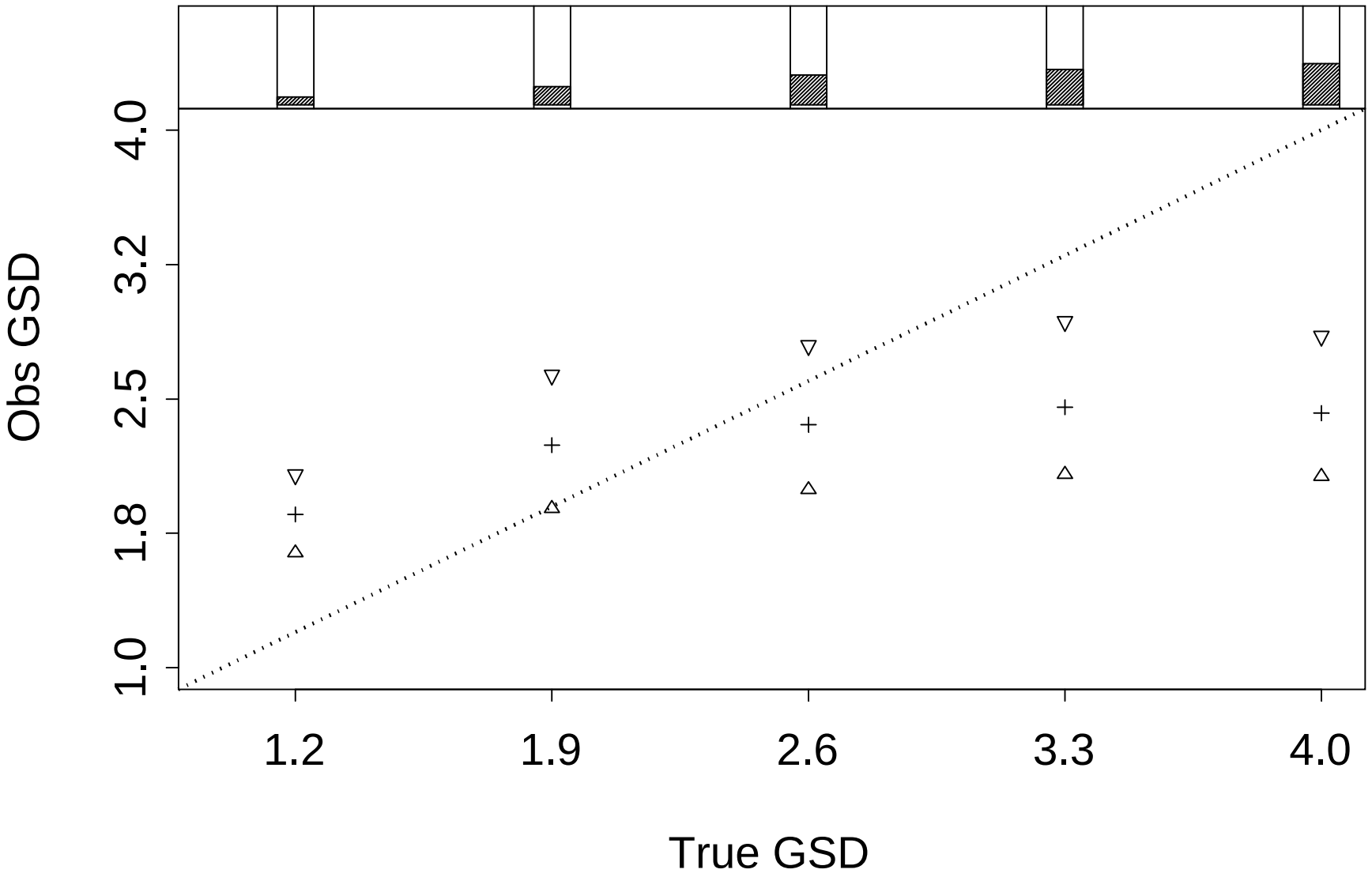

log(EDrat)
-2
2
4
6
-4
-2
0
2
4
log(EDmouse)

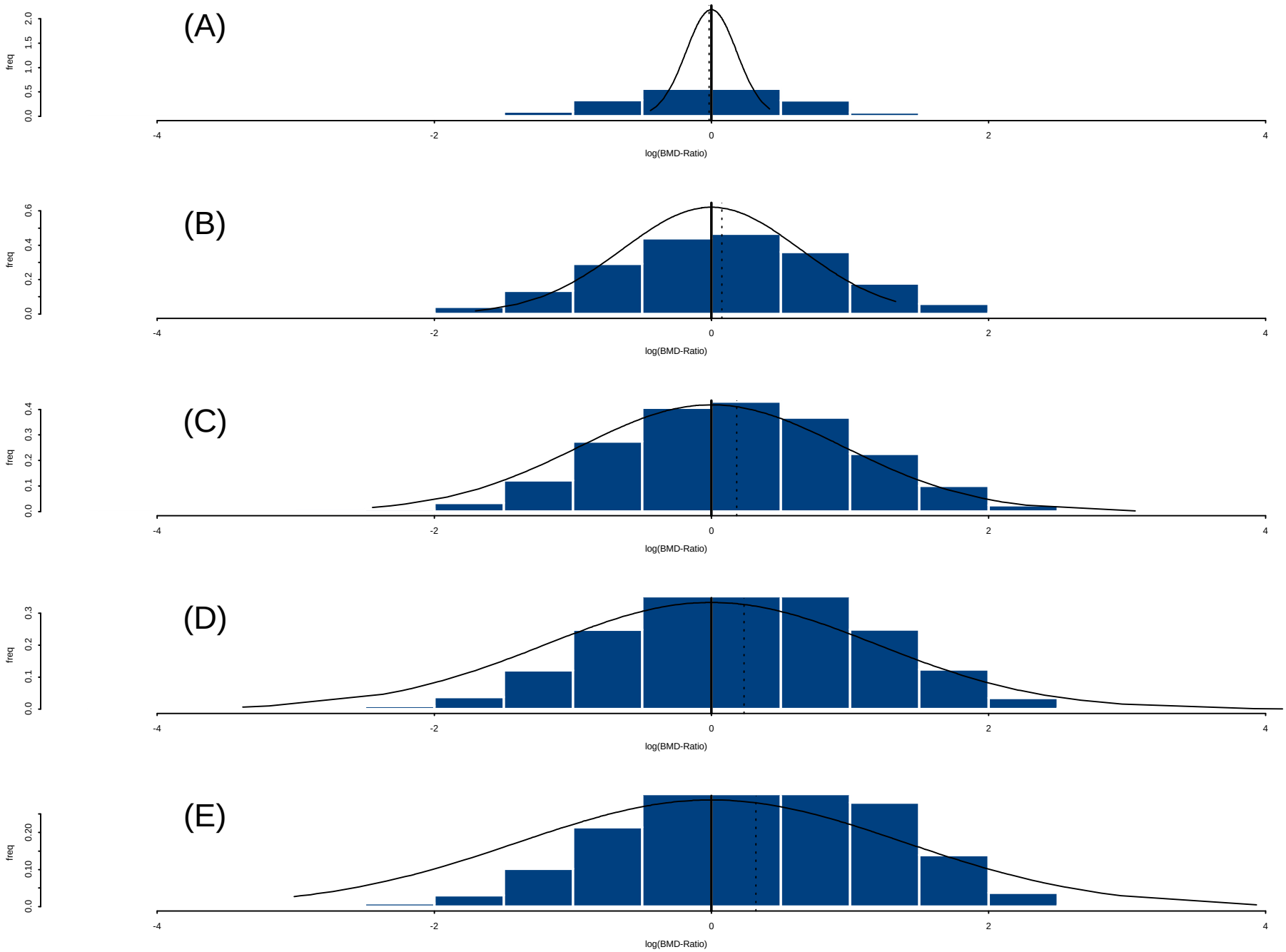

0
4
8
12
-2
0
2
4
6
8
log(EDrat/EDmouse)

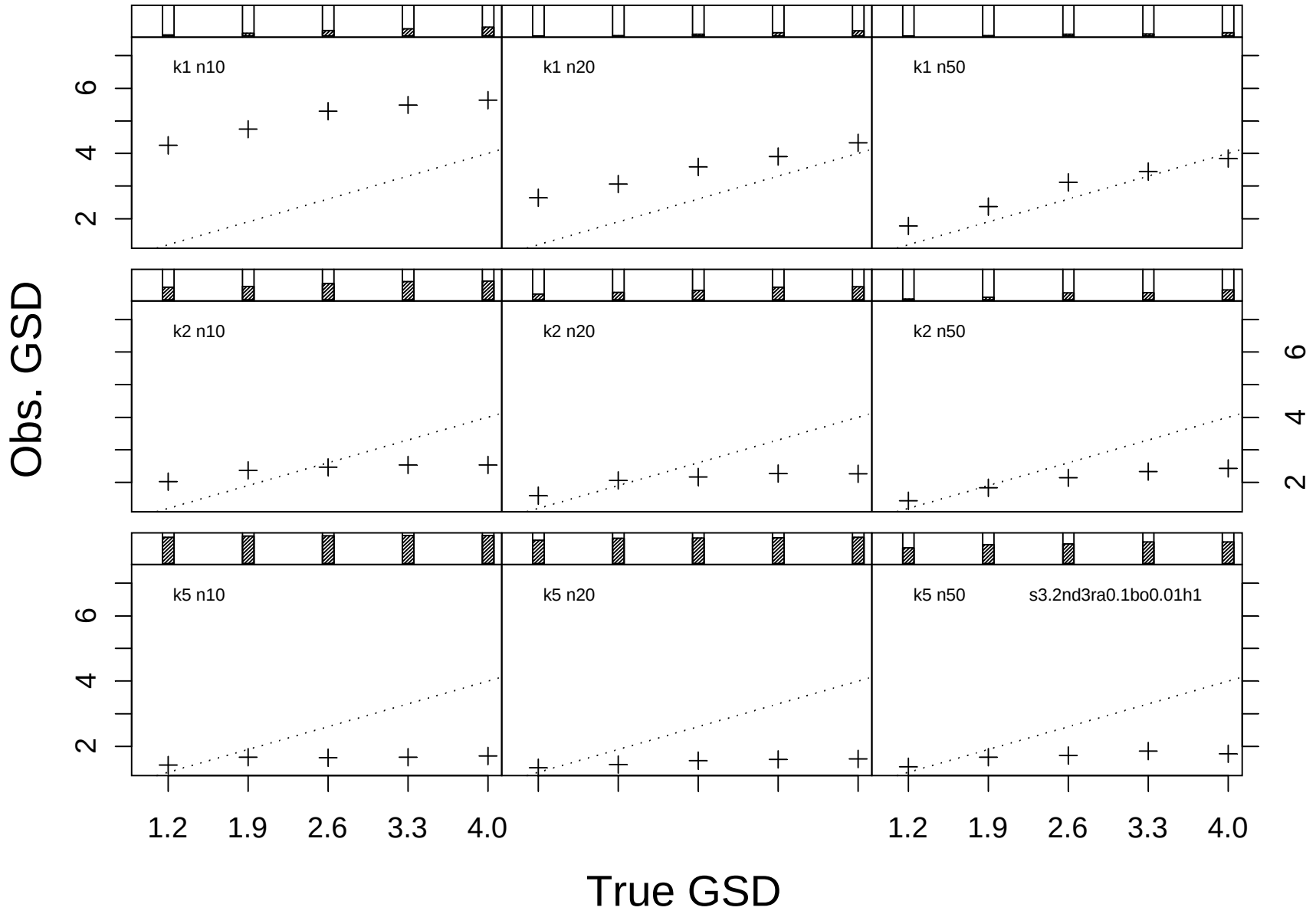

a)
Observed
Truth
b)
Observed
Truth
c)
Observed
Truth


Obs. GM
1
2
3
4
5
6
1.0
1.8
2.5
3.2
4.0
True GM


Obs. GSD
1
2
3
4
5
6
1.2
1.9
2.6
3.3
4.0
True GSD


Obs GSD
1.0
1.8
2.5
3.2
4.0
1.2
1.9
2.6
3.3
4.0
True GSD


(A)
freq
0.0
0.5
1.0
1.5
2.0
-4
-2
0
2
4
log(BMD-Ratio)
(B)
freq
0.0
0.2
0.4
0.6
-4
-2
0
2
4
log(BMD-Ratio)
(C)
freq
0.0
0.1
0.2
0.3
0.4
-4
-2
0
2
4
log(BMD-Ratio)
(D)
freq
0.0
0.1
0.2
0.3
-4
-2
0
2
4
log(BMD-Ratio)
(E)
freq
0.0
0.10
0.20
-4
-2
0
2
4
log(BMD-Ratio)


k1 n10
k1 n20
k1 n50
k2 n10
k2 n20
k2 n50
k5 n10
k5 n20
k5 n50
s3.2nd3ra0.1bo0.01h1
Obs. GSD
True GSD
2
4
6
1.2
1.9
2.6
3.3
4.0