On Group Comparisons with Logistic Regression Models

Jouni Kuha* and Colin Mills†

September 1, 2017

Abstract

It is widely believed that regression models for binary responses are problematic if we want to compare estimated coefficients from models for different groups or with different explanatory variables. This concern has two forms. The first arises if the binary model is treated as an estimate of a model for an unobserved continuous response, and the second when models are compared between groups which have different distributions of other causes of the binary response. We argue that these concerns are usually misplaced. The first of them is only relevant if the unobserved continuous response is really the subject of substantive interest. If it is, the problem should be addressed through better measurement of this response. The second concern refers to a situation which is unavoidable but unproblematic, in that causal effects and descriptive associations are inherently group-dependent and can be compared as long as they are correctly estimated.

Keywords: Logit models; Probit models; Regression modeling; Latent variables; Average treatment effects

*Department of Statistics, London School of Economics and Political Science, UK; e-mail: j.kuha@lse.ac.uk

†Department of Sociology, University of Oxford, and Nuffield College, Oxford; e-mail: colin.mills@sociology.ox.ac.uk

1 Introduction

This paper is about the interpretation of binary response models when making group comparisons. It is widely believed that there is a problem in making such interpretations. This is expressed by Allison (1999, p. 187) in the following terms: "Differences in the degree of residual variation across groups can produce apparent differences in coefficients that are not indicative of true differences in causal effects." The implication of this is that when we apply standard methods to estimate the effects of treatments on binary outcomes, comparisons of the relative sizes of effects between different groups or between analyses with different explanatory variables can be misleading.

These sound like serious problems and Allison's paper and others which have taken up and built upon his message have been influential. Data from Google scholar shows that by 28th February 2017 Allison (1999), the earliest key contribution, had received 742 citations, Williams (2009) 262 and Mood (2010) a remarkable 1053. Scholars not only cite these papers, but also deploy their arguments to qualify the interpretation of their own data and to criticise the work of others. Platt (2009) for instance, in a review of Heath and Cheung (2007) takes the editors to task for the conclusions they allow to be drawn from intergroup comparisons using logit models saying: "...this would anyway fail to take account of the fact that such comparison involves assumptions about the equality of residual variances across models." Marks (2014, p. 175), taking his lead from Mood (2010), tells his readers that a particular technical point "...is important because it undermines conclusions from studies that have used logistic regressions that do not include relevant unobservables". Sikora (2015, p. 273) tells us that in her study: "To avoid problems inherent in comparing logit coefficients or odds ratios between groups ... the key findings are presented as predicted probabilities that are supplemented with tabulated relative risk ratios...", and Kleykamp (2013, p. 847) warns us before she reveals her results that: "... group comparisons expressed through interactions are problematic in non-linear models because such models cannot distinguish group coefficient differences from group differences in residual variation or unobserved heterogeneity".

Clearly it is now widely conceived that there is a problem with using binary response regressions to make group comparisons and that the parameters routinely estimated in such endeavours — for instance odds ratios, log odds ratios and other quantities related to them — are not to be trusted. If this was true, the problem would indeed be even more dramatic than is typically acknowledged. Consider, for example, the pair of hypothetical studies summarised in Table 1. Here we have a binary response variable Y, with values labelled as "Failure" and "Success"; these could for example represent the health outcome of a patient, exam performance of a student, success of a job application, or any of the multitude of binary outcomes that may be examined in empirical research. The studies have a single binary explanatory variable X, and the participants have been randomly assigned to one of its two levels ("Control" and "Treatment"). Two such randomized experiments have been carried out, with two groups of participants which could be for example groups of men and women, people from different countries, or any two types of individuals which it would be interesting to compare. Here the comparative conclusions are clear: the treatment has a positive effect on success, and this effect is larger in aggregate in group A than in group B. Suppose further that these were methodologically perfect studies, with large numbers of participants, error-free randomization, flawlessly and consistently operationalised and measured treatment and outcome, no interference between participants, perfect compliance and no missing data. Even then, however, there would be something incorrect or misleading in the obvious conclusions from these studies, if the group comparison problem was real.

===== Table 1 around here =====

In this paper we argue that this is not the case, and that the so called problem is not one that need concern most empirical researchers who wish to make group comparisons. Our view is that a lack of clarity about what the appropriate target quantities are that are to be estimated in particular empirical enquiries has led many researchers to draw the wrong conclusions from the literature we consider.

To avoid misunderstanding we should say that our argument is not about the technical correctness of various authors' exposition of how binary response models work. Their mathematics is correct. Our point is that some of the implications of what they then go on to conclude have, at the least, been misunderstood, and misunderstood in ways that seems to suggest that sociologists do not always think as hard as they should about what the estimation target is that is most relevant for the substantive question they are trying to answer. Thinking clearly about what it is that is estimated in a binary response model should lead one to conclude that the problem of group comparisons is largely chimerical and that any remaining difficulties arise from expecting these techniques to do things they were never designed to do in the first place.

There are two distinct versions of the supposed group comparison problem, which arise from two different meanings of "unobserved heterogeneity". The first version is that comparative conclusions about effects which are estimated for a binary outcome can be wrong if we want to treat them as estimates of effects for an unobserved continuous outcome which is supposed to have been measured only by the binary variable. The second is that, even for a binary outcome, estimated effects of a treatment are not comparable between groups because the individuals in different groups have different distributions of other predictors of the outcome. These two versions of the issue have not always been clearly separated in the literature. For example, Allison (1999) focuses on the first version and Mood (2010) to a larger extent also on the second, but both draw on both versions for motivation of their discussions.

We argue that the first version of the group comparison problem only exists if we are genuinely interested in the unobserved continuous variable and that, if we are, the problem should be resolved by more serious efforts to measure this variable. The second version arises because estimable causal (as well as descriptive) effects are unavoidably group-dependent — but this is not a problem or an error but an inherent part of what such effects mean. Some of these points have previously been made, although in somewhat different language, by Rohwer (2012) and Buis (2016).

In the rest of this article we discuss the first version of the group comparison issue in Section 3 and the second in Section 4. In preparation for these two main sections, it is first necessary to define clearly what we mean by regression coefficients and their interpretation. This is done in Section 2, and concluding comments are given in Section 5.

2 Interpretation of regression coefficients

2.1 Introduction

In this section we describe the interpretation of coefficients in some common regression models, to the extent that is needed to draw on in later sections. We begin in Section 2.2 with linear regression models, which serve as a point of reference for the binary models. Logit and other

models for binary outcomes are discussed in Section 2.3. The latent-variable motivation of binary response models, which is central to the version of group comparisons considered in Section 3, is then described separately in Section 2.4.

We will throughout focus on the simplest situations where the questions can be explained, omitting extraneous complications and variations. Firstly we will consider the interpretation of regression coefficients as causal effects. This is closest to the spirit in which the question of group comparisons has been discussed in the literature. An alternative interpretation of the coefficients would be as descriptive measures of associations, in a sample or in a finite population. With appropriate modifications, parallel versions of all of our conclusions apply also to such descriptive interpretations.

We begin by considering models with only one explanatory variable, because the issues which are discussed in Section 3 can be described already in this context. Additional explanatory variables are relevant to the questions discussed in Section 4, so they will be introduced there. We take the explanatory variables to be binary, because this makes the interpretations particularly straightforward, but all of the conclusions apply also to models with continuous explanatory variables (effects of a continuous variable X can also be defined for pairs of values of X at a time, but a regression model then also includes parametric assumptions about how these effects vary smoothly across different values of X; the specification and adequacy of such assumptions would be extraneous to the questions considered here).

For the moment we thus consider a response variable Y and a single binary explanatory variable X, and assume an observed sample of data (X_i, Y_i) for n units $i = 1, \ldots, n$. The two values of X are $X = 0$ and $X = 1$; we refer to them as the "control" and "treatment" conditions respectively, but note that the discussion is general and not limited to experimental studies where this language is most natural.

Since the issues that we discuss are about the meaning and interpretation of model parameters we need not concern ourselves with details about how these parameters are estimated. For our purposes it is sufficient to note that regression coefficients with causal interpretations can be validly estimated if the observed data are appropriate for this purpose. In particular, this is the case if the values of X_i were randomly assigned to the units and certain other assumptions are satisfied. We will assume throughout that the estimators that we mention are to be applied to such data, but methods and assumptions of estimation and inference are not otherwise discussed.

2.2 Linear regression

Suppose that Y is a continuous variable, and consider a model where the observed values of Y for units with any given value of X are treated as a random sample from a distribution with mean $\alpha + \beta X$ and variance σ^2, where α, β and σ^2 are model parameters. The familiar linear regression model formulation of this is that

$$Y_i = \alpha + \beta X_i + \sigma \epsilon_i \tag{1}$$

for $i = 1, \ldots, n$, where ϵ_i are random variables which are independent of the X_i and which follow a distribution with mean 0 and a known variance (which is in this section taken to be 1); the residual standard deviation parameter σ is separated from ϵ_i here because doing so will be convenient later.

When X is binary, the least squares estimate of β is

$$\hat{\beta} = \bar{Y}_1 - \bar{Y}_0 \tag{2}$$

where $\bar{Y}_1$ and $\bar{Y}_0$ are the averages of Y_i among those units in the sample for whom $X_i = 1$ and $X_i = 0$ respectively. In other words, $\hat{\beta}$ is the difference of the sample means of Y between the treatment and control groups.

But what is it that $\hat{\beta}$ estimates — that is, how could we interpret β in (1) as a causal effect of X on Y? To define this effect carefully, it is useful to employ the concepts and notation of formal causal inference (for more on these topics, see e.g. Imbens and Rubin 2015). For each of the units $i = 1, \ldots, n$, let $Y_i(X)$ denote the *potential outcome* of Y for that unit if it had value X of the explanatory variable. Thus $Y_i(0)$ and $Y_i(1)$ are the potential outcomes for unit i under the control and treatment conditions respectively. We can also write them as $Y_i(X) = \alpha_i + \beta_i X$ where $\alpha_i = Y_i(0)$ and $\beta_i = Y_i(1) - Y_i(0)$. Here β_i is thus the difference between the values of Y for unit i if that unit had the value $X = 1$ and if it had $X = 0$. This is the *unit-level causal effect* of X on Y when an effect is quantified in terms of differences.

Suppose that we regard the n units in the observed data as the *population* for which we want to draw conclusions about causal effects. (Alternatively, we may think of them as a sample from a larger population, but that would of course require additional assumptions about generalizability from the sample to that population, which is not our concern here.) We will then consider the *distributions* of $Y_i(0)$ and of $Y_i(1)$ in this population, both of them over all of the n units. There is an *average causal effect* of X on Y in this population if these distributions are not the same. The most commonly used measure of an average causal effect is the difference

$$\beta = \overline{Y(1)} - \overline{Y(0)} \tag{3}$$

where $\overline{Y(1)}$ is the average of the $Y_i(1)$ and $\overline{Y(0)}$ the average of the $Y_i(0)$. This is thus the difference between the means of the distributions of the two distinct sets of potential outcomes for the same units in the population. Equivalently, β is also the average of the unit-level causal effects β_i.

Under sufficiently strong conditions for the observed data, model (1) can be treated as a representation of the distributions of the potential outcomes, and used to estimate the effect (3). Here α corresponds to $\overline{Y(0)}$ and $\alpha + \beta$ to $\overline{Y}(1)$. The residual variance σ^2 corresponds to the variances of the potential outcomes $Y_i(0)$ and $Y_i(1)$, under the assumption that these variances are equal (i.e. that X has no average causal effect on the between-individual *variability* of Y). The average causal effect β in (3) is then estimated by the least squares estimate $\hat{\beta}$, given by (2), of the regression coefficient β of the linear model (1). This estimate is unbiased for β even if the assumption that the two sets of potential outcomes have equal variances is not correct.

2.3 Logit models for binary response variables

Suppose now that the response variable Y is binary, with values coded as 0 and 1. Consider a model where the observed values of Y are treated as a random sample from a Bernoulli distribution with probability parameters $\pi_i = P(Y_i = 1)$. We focus on the binary logistic model where π_i depend on X_i through

$$\text{logit}(\pi_i) = \log \frac{\pi_i}{1 - \pi_i} = \alpha + \beta X_i \tag{4}$$

for $i = 1, \ldots, n$. If the observations are taken to be independent given X_i, the maximum likelihood estimate of β is

$$\hat{\beta} = \log \frac{\hat{p}(Y = 1|X = 1)/[1 - \hat{p}(Y = 1|X = 1)]}{\hat{p}(Y = 1|X = 0)/[1 - \hat{p}(Y = 1|X = 0)]} \tag{5}$$

where $\hat{p}(Y = 1|X = k)$ is the conditional proportion of $Y = 1$ given $X = k$ in the sample, for $k = 0, 1$ (and (5) is defined as long as neither of these proportions is 0 or 1). This $\hat{\beta}$ is the log odds ratio between X and Y in the sample. In the hypothetical example of Table 1 we have $\hat{\beta} = \log(6.0)$ in group A and $\hat{\beta} = \log(4.0)$ in group B.

The regression coefficient β in (4) may again be interpreted as a causal effect, using the same ideas as for linear models in Section 2.2. For a binary Y, the possible values of the potential outcomes $Y_i(X)$ are 0 and 1. Their averages over the n units are proportions: $\overline{Y(0)}$ is the proportion of the units for which $Y_i(0)$ is 1, and $\overline{Y(1)}$ is the proportion of the $Y_i(1)$ which are 1. Denoting these quantities by π_0 and π_1 respectively, an average causal effect of X on Y can be quantified by a comparison of π_0 and π_1. We could again consider the difference $\pi_1 - \pi_0$, but for a binary Y other measures are also commonly used. Here we focus on

$$\beta = \log \frac{\pi_1/(1 - \pi_1)}{\pi_0/(1 - \pi_0)}, \tag{6}$$

the log odds ratio between X and $Y(X)$ in the population of n units. It is estimated by (5) under appropriate conditions for the data, because $\hat{p}(Y = 1|X = k)$ is then an unbiased estimate of π_k for $k = 0, 1$. Thus an estimate of the regression coefficient β of model (4) can then be interpreted as an estimate of the causal log odds ratio β in (6).

2.4 Latent-variable motivation of models for binary responses

The model formulation which will give rise to the first version of the group comparison problem is not a linear or a logit model on its own, but in a sense a combination of them. This is the latent-response formulation of the logit model, interpreted as a linear model.

Let Y^* be a continuous response variable which follows the linear model

$$Y_i^* = \alpha + \beta X_i + \sigma \epsilon_i \tag{7}$$

where the ϵ_i have a known distribution which is symmetric around 0 and has the cumulative distribution function $F(\epsilon)$. Suppose that Y_i^* is a latent variable which is not directly observed, but that in its stead we observe a binary variable Y_i which is determined by

$$Y_i = \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \tag{8}$$

for every $i = 1, \ldots, n$. If (7) and (8) hold, Y_i given X_i has a Bernoulli distribution with the probability

$$\pi_i = P(Y_i = 1) = F\left(\frac{\alpha}{\sigma} + \frac{\beta}{\sigma} X_i\right), \tag{9}$$

which also implies that $F^{-1}(\pi_i) = (\alpha/\sigma) + (\beta/\sigma)X_i$. Here the parameters α, β and σ are not all separately identifiable, so σ is typically taken to be equal to 1. The model for Y_i is then a binary regression model with the link function $F^{-1}(\pi)$ and parameters (α, β). This is the

logit model (4) discussed in Section 2.3 if ϵ_i follow a standard logistic distribution, for which $F^{-1}(\pi) = \log[\pi/(1-\pi)]$ and the variance of ϵ_i is $\pi^2/3$. In the rest of this paper we focus on this case for specificity, i.e. we assume that ϵ_i have a standard logistic distribution whenever we consider a latent variable formulation like (7). Other specifications of ϵ_i lead to different binary regression models; in particular, assuming that they have a standard normal distribution yields the probit model for Y_i. All of our general conclusions hold also for these other models which can be derived from (7) and (8).

What implications does this reasoning have for the interpretation of the regression coefficients? Very often, none at all. This is the case whenever the binary Y is the substantively interesting response variable. The latent Y^* is then nothing but a hypothetical mathematical device which may be introduced if it is helpful for motivating the binary response or convenient for examining the properties of the model. For example, the relationship between the coefficients of logit and probit models, and the usual definition of an intraclass correlation for clustered binary data, are both easily derived with the aid of a Y^* (see e.g. ch. 10 of Rabe-Hesketh and Skrondal 2012). The model of interest, however, remains the binary response model for Y, and its coefficients are interpreted as described in Section 2.3.

In some applications, however, the latent Y^* is regarded not as a convenient piece of mathematical fiction but as a real variable of interest. Some important examples include bioassays where Y^* is, say, an animals's tolerance to a toxicant and Y is death (this is the earliest application where the model formulations discussed in this section were worked out in full; see Finney 1947 for a summary and a review of the preceding literature), applications in genetics where Y^* is genetic liability and Y is a disease outcome (Falconer, 1965) and discrete choice models in economics where Y^* is the difference of the utilities of two products and Y is the choice between them (Thurstone 1927; Marschak 1960; Luce and Suppes 1965; McFadden 1974). In such cases, models for Y^* rather than Y can be of primary interest. The results discussed above then show that, assuming (7)–(8) with $\sigma = 1$, and a logistically distributed ϵ_i, the parameters of (7) can be estimated from a logit model for the binary Y_i. In particular, the estimate of β from this model can also be interpreted as an estimate of the coefficient in a linear model for Y^*, and thus also (under the appropriate assumptions for the data) as an average causal effect of X on Y^* — even though Y^* itself was never observed.

This is a useful and powerful result, when it holds, but it does come with some cost. First, the regression results can only be given on a standardized scale where the residual variance of Y^* is fixed at (say) $\pi^2/3$. Second, and more importantly, this model specification involves a set of unverifiable assumptions which can complicate the interpretation and cause the conclusions to be extremely sensitive to different choices for the assumptions. One situation where this happens is that of group comparisons, which we now turn to.

3 Group comparisons for a latent continuous response

3.1 Definition of the problem

Suppose now that we want to compare regression results between two groups of units (comparisons across three or more groups add no new issues). We consider the same type of regression model for both groups, and denote the coefficients of an explanatory variable X in them as β_A and β_B, where the subcripts A and B are labels for the groups. Estimates $\hat{\beta}_A$ and $\hat{\beta}_B$ of the coefficients are obtained using separate samples of observed data (X_i, Y_i) for the groups. The

question that we want to answer in this and the next section is whether the values of these estimates can be meaningfully compared to each other.

We assume that the data satisfy the necessary assumptions to allow for valid estimation of the coefficients *within* each group, and that X is defined in the same way in both groups. The remaining question is then whether the response variable of the regression models that we are estimating is defined in the same way and on the same scale in both groups. If it is, the estimated regression coefficients $\hat{\beta}_A$ and $\hat{\beta}_B$ are comparable between the groups (within some inherent limits of what such comparisons mean, which we will discuss in Section 4).

This condition is satisfied if the model of interest is for a response variable Y which is directly observed and correctly measured on the same scale (i.e. with the same unit of measurement) in both groups. This can usually be achieved for both continuous and binary responses. For a binary Y, its scale — i.e. the meaning of the values 0 and 1 — can be defined identically in the two groups. Causal interpretations of the coefficients of the logit model are based on proportions of potential outcomes of these values of Y, and the interpretations of these proportions are the same across the groups, so the estimated coefficients are directly comparable in this sense.

The situation is more difficult if we are in the setting of Section 2.4, that is if we want to interpret coefficients from a model for a binary Y as coefficients of a linear model for a latent continuous Y^*. Consider model (7), denoting its parameters in the two groups by $(\alpha_A, \beta_A, \sigma_A)$ and $(\alpha_B, \beta_B, \sigma_B)$. It will not be possible to identify all six of these parameters, so we will first fix one of the variance parameters, say σ_A at 1. The models for Y^* are then

$$Y_i^* = \alpha_A + \beta_A X_i + \epsilon_i \quad (10)$$

$$Y_i^* = \alpha_B + \beta_B X_i + \sigma_B \epsilon_i \quad (11)$$

in groups A and B respectively. The response Y^* is measured on the same scale in both of these models, namely the scale on which the residual variance of Y^* given X in group A is $\pi^2/3$. If Y_i^* were directly observed, they could be used to estimate (10) and (11), and the estimates of β_A and β_B would be comparable as estimated average causal effects of X on this Y^*.

Here, however, we observe not Y_i^* but only the binary Y_i determined by (8). The model that will then be fitted for Y_i will implicitly take the residual standard deviation to be 1 in both groups. In group B, this model is not (11) but

$$Y_i^{**} = (Y_i^*/\sigma_B) = (\alpha_B/\sigma_B) + (\beta_B/\sigma_B)X_i + \epsilon_i. \quad (12)$$

The coefficient of X from a logit model in group B will thus estimate β_B/σ_B rather than the correct β_B. This can clearly distort comparisons between the coefficients across the groups: for example, even if β_A is really smaller than β_B, it will be larger than β_B/σ_B if σ_B is large enough. The distortion occurs because the response variables Y_i^* and Y_i^{**} in the two implicit linear models (10) and (12) are measured on different scales. The same problem would arise even for models for an observed response if, say, we inadvertently measured it in inches in one group and in centimetres in another, but treated these as the same in the analysis. This is the first kind of group comparison problem for binary regression models which was identified by Allison (1999) and discussed in the subsequent literature.

In the hypothetical studies in Table 1 we concluded that the effect of X on the binary Y was stronger in group A than in group B. We now know that we could not take this as evidence for the same conclusion about a continuous Y^* measured by Y. The effects on Y^* could be in the opposite order, if the residual variance of Y^* was larger in group B than in group A.

Expressed in terms of the potential outcomes for Y^*, the problem arises because estimation with the binary response model only works if the variances of $Y_i^*(X)$ are the same in group A and in group B, but will be biased if this is not true. Here it is worth noting that we actually need a very similar assumption already when we are considering only one group, namely the assumption that the variances of $Y_i^*(X)$ are equal for $X = 0$ and for $X = 1$ within that group. This too can easily fail to be true. For example, suppose that the variances of $Y_i^*(0)$ and $Y_i^*(1)$ are $\pi^2/3$ and $(1 + \delta)^2\pi^2/3$ respectively, with $\delta \neq 0$. This corresponds to the linear model

$$Y_i^* = \alpha + \beta X_i + (1 + \delta X_i)\, \epsilon_i \tag{13}$$

for data on Y_i^*. In other words, this is a heteroscedastic model where the residual variance is $\pi^2/3$ in the control group ($X_i = 0$) but $(1+\delta)^2\, \pi^2/3$ in the treatment group ($X_i = 1$). This heteroscedasticity would not be a problem if Y_i^* were observed. Estimation with a binary response Y_i is, however, based on the implicit model (7) where the residual variance is homoscedastic. Model (13) can be transformed into this form by writing it as

$$Y_i^{**} = \alpha + \beta^* X_i + \epsilon_i^* \tag{14}$$

where $Y_i^{**} = Y_i^*/(1+\delta\, X_i)$ and $\beta^* = (\beta - \delta\alpha)/(1+\delta)$. The latent continuous response which is implied by the logistic model for Y is thus Y^{**} and not Y^*, and the model will be estimating β^*. This is clearly different from β, and can even have a different sign. In other words, when the assumption that the model for the unobserved Y^* has a constant residual variance is violated, the effects that we would estimate from a model between X and the binary Y can be in the opposite direction compared to the true effects between X and Y^*. This happens because the latent continuous response which is actually implied by the model for Y is measured on different scales in the control and treatment groups defined by X. Since this bias can arise already in the analysis of one study, it should arguably be more worrying than problems in the special case of group comparisons, if we are interested in models for a latent response variable.

3.2 Solutions to the group comparison problem

What, then, can be done to resolve this kind of group comparison problem? In this section we describe four types of solutions: (1) conclude that for your research question the problem does not exist, (2) reparametrize the model to change which parameters can and cannot be identified, (3) choose to report quantities which are comparable across groups, or (4) improve the measurement of the response variable. We argue that (1) and (4) should be the most important solutions, even though the literature on this topic has focused on (2) and (3).

The problem in Section 3.1 applies to comparisons of models for the continuous response Y^* in the latent-variable formulation of binary response models. It is thus of concern only if we actually care about Y^*, i.e. if it is regarded not as a convenient mathematical device but as a real and meaningful quantity which is assumed to underlie the observed binary Y, *and* if our substantive research questions are really about Y^*. If, however, the research questions are about the binary Y itself rather than any Y^*, the problem simply does not exist and can be safely ignored. This distinction was forcefully expressed by Joseph Berkson already in 1951:

> "If it is seriously believed that there is some physical property more or less stably characterizing each organism [Y^*], which determines whether or not it succumbs [Y], then it is justifiable to advance the hypothesis of a distribution of tolerances. In that case one should be prepared to suggest the nature of this characteristic so that

> the hypothesis may be capable of corroboration by independent experiments. If on the other hand the formulation is only that of a 'mathematical model,' to guide the method of calculation, then it would seem more objective and heuristically sounder not to create any hypothetical tolerances, but merely to postulate that the proportion of organisms affected follows the integrated normal function [a probit model]. I am interested in the slope of the dosage mortality line as a 'rate,' of the objectively observed increase of mortality with increase of dosage, not as a standard deviation of hypothetical tolerances of the animals. I should of course be very much interested in the last, if tolerance of the animals is what I was observing and studying. But we are not dealing with measured tolerances, we are dealing with a dosage mortality curve, and when my probitistic friends present a standard deviation of tolerances, they may be asserting a substantial quantity for the variability of something that in fact does not exist at all."

(Berkson 1951, quoted with permission from John Wiley & Sons). Here Berkson was discussing applications in bioassays, where it might not seem very implausible to grant an independent existence to such latent tolerances. In our view the situation is even clearer in the social sciences, where convincing and necessary examples of them are rarer still.

To substantiate this claim we examined a sample of 100 articles which cited the literature on the group comparison question, specifically the widely cited article by Mood (2010).[1] They are primarily from journals in the fields of sociology, political science, education, public health, social psychology, social policy, criminology and demography. Of the 100 papers, 84 had a response variable which at the analysis stage was treated as a dichotomy, and these form the sample we discuss below. In terms of the contexts in which the group comparison literature is referred to, the sample includes two coherent non-mutually exclusive sets of interest to us. The first (30%) contains discussions of group comparisons (in the spirit of this section, and/or Section 4.1 below) and the second (45%) discusses comparisons of models with different explanatory variables (also bringing in issues from Section 4.2 below). A third, extremely heterogeneous set (32%) concerns itself with neither of these but with other questions which are not the central focus of this paper.

None of the articles made an explicit conceptual distinction between the observed response Y and a latent continuous Y^*, let alone expressed their research questions explicitly in terms of a Y^*. In a large majority of them it was clear that the authors' interest was in the observed binary response and not some latent underlying variable with a real existence. Examples of such responses were death, entry into marriage, voter registration, making a school transition, the existence of a network tie, owning a house, being unemployed, adopting of a social media app, and detecting a software fault. However, for 17% of the articles it would in fact be possible to argue that there is an underlying Y^* variable that could be of interest even if the authors did not explicitly recognize it. These examples relate to phenomena, such as academic ability, degree of trust, intensity of pain, and degree of support for a policy position, where it is natural to talk about the intensity of the response. However, almost all of these cases where also ones where the response variable had originally been measured on a continuous or ordinal scale and only subsequently dichotomized. In other words, Y^* was not really latent because it had actually been measured, before much of the information on it had been discarded (and the Y^* problem introduced) at the analysis stage.

Suppose now that we conclude, nevertheless, that models for a latent Y^* are what we care about, so that the group comparison problem does need to be addressed. One way of doing

this involves changing the identifying assumptions of the model. To explain it, consider again the models for a latent Y^* in two groups A and B as given by equations (10) and (11). The comparison problem arises because from models for a binary Y it is not possible to separately estimate all five parameters in these models, but only the four quantities α_A, β_A, (α_B/σ_B) and (β_B/σ_B). These will yield estimates of β_A and β_B if we are willing to impose one additional parameter constraint. The discussion so far has considered the most commonly used constraint $\sigma_B = 1(= \sigma_A)$, but other choices are also possible. For example, we could leave σ_B estimable but assume instead that $\alpha_A = \alpha_B$, i.e. that the intercept parameters are equal across the groups. We could even assume that the regression coefficients are equal ($\beta_A = \beta_B$). This last possibility would often be pointless because it would assume away the very difference that we would want to estimate, but this is not always the case. For example, this assumption is sometimes used in econometric modelling of discrete choice data (Swait and Louviere 1993; Train 2003). The models are then used to allow for group differences in the residual variances of Y^* when using combined data for the groups to estimate regression coefficients which are simply assumed to be equal across the groups.

All of these models with the same number of identification constraints are observationally equivalent, so they cannot be compared to each other in terms of goodness of fit. They can, however, give very different estimates of the parameters of interest. If model (10)–(11) holds, the estimate of the coefficient β_B will actually be estimating β_B/σ_B if we assume $\sigma_B = 1$, but $\beta_B/(\alpha_B/\alpha_A)$ if we assume that the intercepts are equal. These can be very different from each other, and neither needs to be close to the true β_B unless the identification restriction matches the true model for Y^*.

If the model includes several explanatory variables, it is also possible to impose constraints on the coefficients of one or more of them while leaving both the rest of the coefficients and the residual variances separately estimable. This can also be done for several explanatory variables at once, for example assuming that the coefficients of all control variables are equal across groups but the coefficient of main interest and the residual standard deviation need not be. In this case the model will include more than the minimum number of parameter constraints, and it will be possible to assess the appropriateness of some of the remaining constraints by testing them against less restrictive models. In the article where he introduced the group comparisons question, Allison (1999) proposed methods along these lines as solutions to the problem. However, any such comparison is still conditional on a specific set of assumed constraints, and we could always consider alternative ones which are equivalent in terms of fit but which can produce very different estimates for the parameters of interest (similar comments are made by Williams 2009 based on a simulation study). In other words, all that such specifications can really do is to shift the assumptions of an inherently poorly identified model from one part of the model to another. This will not solve the group comparisons problem, unless we are entirely convinced that a particular parameter constraint is substantively correct.

A different approach which has been proposed for obtaining meaningful group comparisons of models for a latent Y^* is to focus on comparable quantities which *can* be identified from the observed data. In particular, it is possible to estimate the ratio between the coefficient β in (7) and either the marginal standard deviation of Y^* (as implied by the model) or the conditional standard deviation of Y^* given X (or, more generally, given any subset of multiple explanatory variables), because these ratios can be estimated without estimating the parameter σ. The results may also be further standardized by marginal or conditional standard deviations of X. Breen et al. (2014) propose quantities of this form, and show that they include the marginal and partial correlations between X and Y^*, and all the commonly used versions of standardized

regression coefficients of X on Y^*. It is worth noting that the coefficient from the logistic regression of Y on X is itself also of such standardized form, since it estimates β/σ (when X is binary, this is in fact a simple multiple of Cohen's d measure of effect size).

This approach works well when describing and comparing standardized associations between X and Y^* is appropriate for the substantive research questions, but is less useful if this is not the case (Breen et al. (2014) illustrate these considerations with a range of sociological examples). It is not very natural when the aim is to interpret regression coefficients as average causal effects of X on Y^*, in the ways described in Section 2. In particular, the marginal variance of Y^* is a function of the distribution of X_i for the n units in the observed data. Standardized quantities which use this variance will then also depend on the distribution of X_i and not just on the distributions of $Y_i(0)$ and of $Y_i(1)$. This makes it difficult to interpret the standardized statistics as clear causal quantities.

The root cause of the group comparison problem as formulated in Section 3.1 — and of other difficulties with estimating models for Y^* — is one of weak measurement: we are trying to manage by measuring a continuous Y^* with just a single observed binary indicator Y. The measurement model for Y is assumed to be (8), which can also be written as $P(Y = 1|Y^*) = \text{logit}[\lambda(Y^* - \kappa)]$ with $\lambda = +\infty$ and $\kappa = 0$. Even if the assumed model (7) for Y^* is correct, this deterministic measurement model must also be correct. In other words, we must be willing to assume that for every unit with a given value of a real Y^*, Y is always observed to be 0 if Y^* is at or below a threshold value κ, and always observed to be 1 if Y^* is above κ, and that this threshold is at $\kappa = 0$ for everyone. These are likely to be implausibly strong assumptions for actual measurements in most applications.

Substantively motivated latent variables are of course common in many social science applications. A general and more flexible measurement strategy for them is also familiar: use multiple observed indicators which are all regarded as imperfect measures of the latent variable, and define and estimate latent variable measurement models which represent this situation (see e.g. Bartholomew et al. 2011 for an overview of such models). In particular, measurement by multiple binary indicators Y_j may be represented by logistic measurement models of the form $\text{logit}[P(Y = 1|Y^*)] = \tau_j + \lambda_j Y^*$ where λ_j are finite. Such measurement models are, for example, the workhorses of item response theory (IRT) modelling in psychological and educational testing. If there are at least three observed indicators (and they are assumed to be conditionally independent given Y^*), both these measurement models and the model (10) for Y^* are identified in one group (with one additional constraint on the intercept parameters, e.g. $\alpha_A = 0$ or $\tau_1 = 0$). Furthermore, in the two-group situation we can estimate models (10) and (11) for both groups — including all of β_A, β_B and σ_B — if we are also willing to assume that the measurement parameters τ_j and λ_j are the same across the groups for at least two of the indicators Y_j (this is the assumption of between-group equivalence of measurement; see e.g. Millsap 2011).

This is what we would recommend if models for a latent continuous Y^* are of genuine substantive interest: treat the task seriously as a measurement problem for a latent variable, collect data on appropriate multiple indicators of Y^*, and model the data using conventional latent variable models. The fact that this approach is not mentioned in the literature which discusses the group comparison problem again suggests that most applications considered in it are not really about Y^* but about the observed binary Y.

4 Group comparisons for a binary response

4.1 The value of an effect is group-dependent

As discussed in Section 1, the group comparison problem has typically been discussed in terms of "unobserved heterogeneity". In Section 3 we considered one version of this issue, where the heterogeneity refers to the residual variability of a latent Y^* and this is unobserved in the sense that it cannot be quantified on the same scale across groups when Y^* itself is unobserved. In this section we discuss the quite different issue of heterogeneity in unit-level causal effects, and how it affects group comparisons. This question applies directly to observed response variables Y so we focus on them. Latent Y^* are not needed here even for motivation, and they play no role in this section.

As explained in Sections 2.2 and 2.3, a unit-level causal effect is defined by a comparison of the potential outcomes $Y_i(X)$ given $X = 0$ and $X = 1$ for a unit i (e.g. an individual person), for example the difference $Y_i(1) - Y_i(0)$. These effects will vary across the units. For a continuous Y it is at least conceivable (but still quite implausible) that they could be constant — so that, say, $Y_i(1) - Y_i(0) = 7$ for everyone — but for a binary Y this is not in general even possible. The four possible values of the pair of potential outcomes $(Y_i(0), Y_i(1))$ are $(0, 0)$ and $(1, 1)$ — the cases where the treatment X has no effect on Y for unit i — and $(0, 1)$ and $(1, 0)$, the cases where the treatment does have an effect. The effect could only be the same for every unit if the treatment had no effect on anyone, or if it had the same effect on everyone (e.g. if every patient got healthy under the treatment condition, and none of them under the control). These situations are implausible and would in any case be trivially detectable from observed data. All real and interesting populations are thus mixtures of units with different unit-level effects.

Recall that regression coefficients estimate average causal effects, aggregated over the unit-level effects in a specific group (population). For example, the coefficient of a logistic model estimates the population log odds ratio β given by (6), which is an average effect of this kind. This β is a group-level quantity, and its value will depend on the group, specifically on the mixture of units with different unit-level effects that make up the group.

For an illustration of this group-dependence of effects, consider the hypothetical situation in Table 2. Here we have two groups, each with 600 individuals. The upper part of the table shows the numbers of individuals with different potential outcomes $Y_i(X)$ under the two values of X. For instance, if X was set to 1 for everyone, the number of the 600 people who would have the value $Y = 1$ would be 300 in group A and 330 in group B. Here the population log odds ratio β is $\log(1.86)$ in group A and $\log(2.27)$ in group B. For added simplicity, suppose further that there is no one with $(Y_i(0), Y_i(1)) = (1, 0)$. The only individuals for whom X has an effect are then those for whom $(Y_i(0), Y_i(1)) = (0, 1)$. There are 120 of them in group B but only 90 in group A, and it is because of this difference that the population-averaged effect is larger in group B.

===== Table 2 around here =====

Coefficients in binary (and any other) regression models thus estimate effects which are group-specific quantities. In other words, the β in a regression equation like (4) does not represent a sort of universal constant — *the* effect of X on Y — which could be separable from the group context, but a quantity whose value depends not only on the nature of X and Y but

also on the set of units for whom the effect is evaluated. Furthermore, in most applications there is no reason to think that there should be such a universal effect which we should be seeking to determine, beyond the group-specific effects which are estimable from observable data. This is not a flaw or a bias, but an inherent characteristic of such effects. In particular, estimates of these effects can be compared between groups without conceptual problems, once we have clearly specified which groups we are interested in, and estimated the effects from sufficiently strong data from these groups. For example, the comparative conclusion from the hypothetical studies in Table 1, that the average causal effect of X on Y is stronger in group A than in group B, is justified and correct. In such comparisons we are typically interested in determining where the effects are weakest and strongest, and how much they vary between groups of interest. In addition, we may want to try to understand, even if imperfectly, *why* the effects vary as they do. Topics that are related to this question are discussed next.

4.2 Models with different covariates

Causal effects are thus group-dependent because individuals are heterogeneous in their responses to any treatment, and because groups are heterogeneous collections of individuals. Some of this heterogeneity is likely to be due to differences in other observable characteristics of the individuals which have causal effects of their own on the response. Such other characteristics, which we denote here by Z, can be taken into account in regression modeling by including them in a model as covariates (control variables), in addition to the treatment of interest X. This also bears upon the discussion of group comparisons of regression coefficients, in that models with covariates Z in effect refer to different (sub)groups defined by values of Z. The implications of this change of focus are the topic of this section.

What we consider here is how average causal effects of X on Y are affected by the covariates Z for the units among whom the effect is evaluated, for different choices and values of Z. We do not discuss situations where Z are mediator variables in the causal pathway from X to Y. In that context there are also differences in how linear and binary response models behave, but the reasons and interpretations of these differences go well beyond the topics of group comparisons discussed here (for the complex questions of mediation analysis, see e.g. VanderWeele 2015).

Our overall conclusions on this question follow from the discussion in Section 4.1. Models with different covariates Z do not ultimately raise any really separate questions or problems about group comparisons, because such models in essence estimate effects in different groups defined by values of Z. These are all true effects, and if we have a strong enough research design — such as randomization of X — we can estimate any of them, with or without Z. It is thus important to emphasise that we are here not talking about confounding by Z, i.e. situations where the data are such that we *must* control for Z in order to get valid estimates of the effects of X. Instead, covariates Z (whether observed or not) are inherently involved in what the average causal effect of X in a population means, in that the distribution of Z in that population sets the context in which the effect of X is realised. This also implies that effects estimated from models given some Z are not a priori more relevant or somehow purer than, say, ones from models without any Z; for example, the effect among men (i.e. given $Z = male$) does refer to a population which is homogeneous in gender, but this is not helpful if that is not the population in which we wanted to estimate an effect.

The literature on group comparisons is in large part fairly unclear on this topic. Much of the discussion in it raises as problems issues which only arise if we implicitly or explicitly think that estimates of effects for one group should also apply to other groups. To explain this, we

may ask the following questions: "What, if anything, can estimates from a model for Y given X only tell us about the effects of X on Y controlling also for Z — and, in reverse, what can estimates from a model given X and Z tell us about effects of X when not controlling for Z?". The rest of this section is about these questions.

It is sufficient to consider just one additional explanatory variable Z. We take it to be binary, and for concreteness refer to it as an individual's gender, with values $Z_i = 0$ for men and $Z_i = 1$ for women. We again take X to be binary as well, and consider its effect on a binary response Y. As defined in Section 2.3, a logit model for $\pi_i = P(Y_i = 1)$ given X only is

$$\text{logit}(\pi_i) = \alpha + \beta X_i. \tag{15}$$

We will contrast it with the following model which also includes Z:

$$\text{logit}(\pi_i) = \alpha + \beta X_i + \gamma Z_i. \tag{16}$$

Some of the discussion compares logit models with linear regression models. This can be done even with linear models for the same binary Y, so we will also consider the linear probability models

$$\pi_i = \alpha + \beta X_i \qquad \text{and} \tag{17}$$

$$\pi_i = \alpha + \beta X_i + \gamma Z_i, \tag{18}$$

corresponding to (15) and (16) respectively (the fact that the standard assumptions of a linear model are not fully appropriate for a binary response can be ignored for the questions considered here). We illustrate the discussion with reference to the example in Table 2, the lower part of which further separates the numbers of people in each of the two groups by gender (Z).

Suppose, first, that we fit the models (15) and (17) which include only X. We already know that, given appropriate data, the coefficients $\hat{\beta}$ from them are estimates of the population log odds ratio (6) and the difference of proportions (risk difference) (3) respectively. But are they also estimates of the effects given fixed levels of Z — i.e. effects of X on Y among just men and/or women in that population?

In general, the answer to this question is obviously no. This is because there is only one overall effect of X in a population, but two distinct effects for men and for women. One number can represent both of the latter only if the men's and women's effects are equal to each other, i.e. if there is no interaction between X and Z in their effect on Y. But the men and the women are different groups of individuals, and different again from the pooled group of both of them together. In light of the discussion in Section 4.1, we have every reason to expect that the effects in all of these groups are different in magnitude. In other words, an assumption of no interaction is unlikely to hold exactly, but can usually be only a convenient, parsimonious approximation at best (in which role it is of course extremely useful and routinely used). It is also worth noting that if an interaction is absent for one measure of an effect, it is in general present for others. For example, in Table 2 the log odds ratio is the same for men and women in group A but the risk difference is not the same, while the opposite is true in group B.

Suppose now that the interaction is, nevertheless, absent. Imagine first that this is the case for the risk difference, which thus has the same value β among both men and women, as in group B in Table 2. It will then have this same value also for the combined population of men and women together. In this case the estimates of β from the linear models (17) and (18), with and without controlling for Z, are estimating the same quantity, and either estimate can be interpreted as the estimated risk difference among men, women, or both together.

A similar conclusion does *not* hold for the log odds ratio. Suppose that the interaction is absent on this scale, so that the log odds ratio is β among both men and women. Now, however, the log odds ratio in the combined group is not equal to this β (unless gender has no effect on Y), but has a different (and smaller) value. For example, in group A in Table 2 we have $\beta = \log(2.00)$ among each of men and women separately, but $\beta = \log(1.86)$ among both of them combined.

This result is most often introduced in the context of models with a continuous Z, where the conclusion is the same (see Allison 1999); in the literature on models for longitudinal or clustered data it is known as the contrast between "population-averaged" and "cluster-specific" effects (see e.g. Section 13.2 of Agresti 2002). Mathematically, this property of the logit model (16) is a consequence of the fact that π_i is a nonlinear function of $\alpha + \beta X_i + \gamma Z_i$. Substantially and conceptually, it is a real and meaningful conclusion and not a problem or a paradox. Recall again that men, women, and both together are three different groups. What we now know is that even if an effect is the same among both of the genders separately, it still does not need to be the same in the combined group. This conclusion does not pose any new problems for the estimation of these effects. What it tells us is simply that even if there is no interaction, the log odds ratio which holds at each level of Z separately cannot be estimated from a logit model given X only, but must be estimated from a model which includes Z as well.

Suppose now that we want to move in the other direction, that is use estimates from a model which controls for Z to draw conclusions also about effects in the population pooled over Z. A particular question which has often been raised in this context starts with the following observation: If we have data where X_i and Z_i are uncorrelated (e.g. if X was randomized within the levels of Z, each with the same proportions of control and treatment conditions), (least squares) *estimates* of β for the linear models (17) and (18) are always identical, whereas the estimates of β for the logit models (15) and (16) are not the same. This is often presented as a problem or a limitation of the logit model, which is again thought to compromise group comparisons with such models. We argue that it is not, and does not.

Recall first that if there is in fact no interaction for the risk difference, the linear models (17) and (18) will be estimating the same true effect β. It is then to be expected that the estimates $\hat{\beta}$ from them should be similar — and they are indeed even identical if X_i and Z_i are exactly uncorrelated in the sample. This is not, and should not be, the case for the logit model, because the models with and without Z are estimating different true effects, as discussed above.

The special feature of this situation for the linear model becomes apparent when the assumption of no interaction does not in fact hold in the population. The no-interaction model (18) is then wrong, and the $\hat{\beta}$ from it cannot be estimating the true βs among both men and women, since those are not equal to each other. On the other hand, $\hat{\beta}$ from the X-only model (17) does estimate the true β for the combined population of men and women — and since the estimate from model (18) is equal to it if X_i and Z_i are uncorrelated, it too will then be an estimate of the combined-population effect. Denoting by $\bar{Y}_{jk}$ the mean of Y_i (i.e. the proportion of $Y_i = 1$) among the units i in the data for whom $X_i = j$ and $Z_i = k$, and by $n_{.k}$ the number of units for whom $Z_i = k$, for $j, k = 0, 1$, both of these estimates of β are equal to

$$\hat{\beta} = \frac{n_{.0}}{n}(\bar{Y}_{10} - \bar{Y}_{00}) + \frac{n_{.1}}{n}(\bar{Y}_{11} - \bar{Y}_{01}) \equiv p(Z_0)\hat{\beta}_0 + p(Z_1)\hat{\beta}_1 \tag{19}$$

where $p(Z_k) = n_{.k}/n$ is the proportion of observations in the data with $Z_i = k$. Here $\hat{\beta}_k$ is the estimate of β which would be obtained by fitting model (17) only to data with $Z_i = k$. In other words, the slightly peculiar conclusion from this situation is that when X_i and Z_i are uncorrelated and even when there actually is an interaction between X and Z, the estimated

coefficient of X from the model with Z but without the interaction is an unbiased estimate of the effect without Z, because it is in fact a weighted average of the two estimated coefficients from a model with Z and with the interaction.

There is no similar result for logit models, but neither do we need one. If we do want to aggregate up from quantities which are conditional on Z to ones which are not, this can easily be done, as long as it is done in a way which is appropriate for logit models (in essence, which averages first and takes logits second, rather than the other way round). For example, suppose that we have fitted the model (16), being willing to assume that there is no interaction between X and Z. Denoting its parameter estimates by $(\hat{\alpha}, \hat{\beta}, \hat{\gamma})$, an estimate of the proportion $\overline{Y(1)} = \pi_1$ in the population pooled over Z is

$$\tilde{\pi}_1 = p(Z_0)\,\frac{\exp(\hat{\alpha}+\hat{\beta})}{1+\exp(\hat{\alpha}+\hat{\beta})} + p(Z_1)\,\frac{\exp(\hat{\alpha}+\hat{\beta}+\hat{\gamma})}{1+\exp(\hat{\alpha}+\hat{\beta}+\hat{\gamma})},$$

and $\tilde{\pi}_0$ is calculated similarly but omitting $\hat{\beta}$. Any desired measure of an average causal effect can then be estimated as a function of $\tilde{\pi}_0$ and $\tilde{\pi}_1$, in particular the log odds ratio (6) by $\tilde{\beta} = \log[\tilde{\pi}_1/(1-\tilde{\pi}_1)]/[\tilde{\pi}_0/(1-\tilde{\pi}_0)]$. There are also various other statistics which employ this principle of aggregating estimates from a model given X and Z over the sample distribution of Z_i, for example the average partial effects discussed by Mood (2010). This approach can be particularly useful with observational (as opposed to experimental) data when controlling for confounders Z is necessary for valid estimation of causal effects. However, this motivation has little to do with the group comparisons which are our focus here.

5 Conclusions

When researchers use regression models for binary outcomes, they should first make sure to be clear about what the target quantities of their analysis are. In this paper we have argued that when this is done, they will in most cases be able to conclude that comparisons of estimates from such models between different groups or between different models pose no fundamental problems, or at least not the kinds of problems which have been raised in the literature on this question.

Of the two kinds of group comparison problems which have been discussed in the literature, one is expressed in terms of a hypothetical continuous latent variable Y^* underlying the observed binary response. This problem disappears if the substantive research questions of a study are not about such a Y^*. We believe that the vast majority of studies in the social sciences which analyse binary outcomes are of this kind. If the research question is really about a Y^*, the problem is both real and difficult. It is difficult because the study is then relying on an extremely weak and fragile measurement strategy, and all results from the analysis, not just group comparisons, will be correspondingly sensitive to a set of demanding and unverifiable assumptions.

The second form of the proposed group comparison problem does not involve latent outcomes. It arises instead when individuals are heterogeneous in their responses to the treatment of interest, and differently heterogeneous in different groups. We have argued that this type of heterogeneity is not a problem but an unavoidable fact, so that the kinds of average causal effects which we can hope to estimate are inherently group-dependent. Bearing this is mind,

such effects can be compared between groups (populations) as long as they are correctly estimated. The researchers' aims should then to be clear about what their target populations are, and to fit models which estimate effects in those populations.

While these should be reassuring conclusions, they of course do not mean that it will be easy to estimate causal effects (or population associations either), in one group or several. The real problems with doing this are the ones which were not discussed in this article and which we assumed away at the start of it, namely ensuring that the research design, measurement and analysis are sufficiently powerful to allow valid conclusions to be drawn from a study. The true challenges of methodology lie there.

Notes

[1]The sample was drawn from citations listed on Google Scholar (GS) on 21st January 2017. To be eligible for inclusion in the sample the publication had to be a journal article about a substantive research question. We then selected the first 100 articles that satisfied these criteria, in the 'relevance' ordering used by GS. This amounts to sbout one tenth of the citations to Mood (2010) on GS at the time. In the 84 articles with a binary response, as discussed in the main text, we include cases where the response variable was time until an event handled in discrete time.

References

Agresti, A. (2002). *Categorical Data Analysis* (Second ed.). New York: Wiley.

Allison, P. D. (1999). Comparing logit and probit coefficients across groups. *Sociological Methods & Research 28*, 186–208.

Bartholomew, D., M. Knott, and I. Moustaki (2011). *Latent Variable Models and Factor Analysis: A Unified Approach* (Third ed.). Chichester: Wiley.

Berkson, J. (1951). Why i prefer logits to probits. *Biometrics 7*, 327–339.

Breen, R., A. Holm, and K. B. Karlson (2014). Correlations and non-linear probability models. *Sociological Methods & Research 43*, 571–605.

Buis, M. L. (2016). Logistic regression: When can we do what we think we can do? Unpublished note, V. 2.3.

Falconer, D. S. (1965). The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Annals of Human Genetics 29*.

Finney, D. J. (1947). *Probit Analysis: A Statistical Treatment of the Sigmoid Response Curve*. Cambridge: Cambridge University Press.

Heath, A. F. and S. Y. Cheung (Eds.) (2007). *Unequal Chances: Ethnic Minorities in Western Labour Markets*. Oxford: Oxford University Press.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. New York: Cambridge University Press.

Kleykamp, M. (2013). Unemployment, earnings and enrollment among post 9/11 veterans. *Social Science Research 42*, 836–851.

Luce, R. D. and P. Suppes (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. H. Galanter (Eds.), *Handbook of Mathematical Psychology*, Volume 3, pp. 249–410. New York: Wiley.

Marks, G. N. (2014). *Education, Social Background and Cognitive Ability: The Decline of the Social*. Abingdon: Routledge.

Marschak, J. (1960). Binary-choice constraints and random utility indicators. In K. J. Arrow, S. Karlin, and P. Suppes (Eds.), *Mathematical Methods in the Social Sciences, 1959: Proceedings of the First Stanford Symposium*, pp. 312–329. Stanford, CA: Stanford University Press.

McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics*, pp. 105–142. New York: Academic Press.

Millsap, R. E. (2011). *Statistical Approaches to Measurement Invariance*. New York: Routledge.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review 26*, 67–82.

Platt, L. (2009). Review of '*Unequal Chances: Ethnic Minorities in Western Labour Markets*' by Heath and Cheung (eds.). *European Sociological Review 25*, 265–267.

Rabe-Hesketh, S. and A. Skrondal (2012). *Multilevel and Longitudinal Modeling using Stata* (Third ed.), Volume II. College Station, TX: Stata Press.

Rohwer, G. (2012). Estimating effects with logit models. NEPS Working Paper 10, German National Educational Panel Study, University of Bamberg.

Sikora, J. (2015). Gender segregation in Australian science education: Contrasting post-secondary VET with university. In C. Imdorf, K. Hegna, and L. Reisel (Eds.), *Gender Segregation in Vocational Education*, Number 31 in Comparative Social Research, pp. 263–289. Bingley: Emerald Group Publishing.

Swait, J. and J. Louviere (1993). The role of the scale parameter in the estimation and comparison of multinomial logit models. *Journal of Marketing Research 30*, 305–314.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review 34*, 273–286.

Train, K. E. (2003). *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

VanderWeele, T. J. (2015). *Explanation in Causal Inference: Methods for Mediation and Interaction*. New York: Oxford University Press.

Williams, R. (2009). Using heterogeneous choice models to compare logit and probit coefficients across groups. *Sociological Methods & Research 37*, 144–148.

Table 1: Hypothetical data from two randomised experiments, with two distinct groups of participants (labelled groups A and B). The participants have been assigned to one of two experimental conditions, indicated by a binary variable X with values 0 ("Control") and 1 ("Treatment"). The response variable Y is also binary, with values 0 ("Failure") and 1 ("Success"). The tables show the observed conditional proportions $\hat{p}(Y = j|X = k)$, $j, k = 0, 1$, of the responses by the experimental condition, separately for the two groups. Standard measures of association between X and Y in these data are shown below the table; here the logarithms of the odds ratios are also equal to the estimated coefficients of X in binary logistic regression models fitted for Y given X.

	Group A:		*Group B:*	
	$Y = 0$ (Failure)	$Y = 1$ (Success)	$Y = 0$ (Failure)	$Y = 1$ (Success)
$X = 0$ (Control)	0.80	0.20	0.50	0.50
$X = 1$ (Treatment)	0.40	0.60	0.20	0.80
Odds ratio $[\hat{\pi}_1/(1-\hat{\pi}_0)]/[\hat{\pi}_0/(1-\hat{\pi}_1)]$	6.0		4.0	
Risk ratio $(\hat{\pi}_1/\hat{\pi}_0)$	3.0		1.6	
Risk difference $(\hat{\pi}_1 - \hat{\pi}_0)$	0.4		0.3	

Note: Here $\hat{\pi}_k = \hat{p}(Y = 1|X = k)$ for $k = 0, 1$.

Table 2: A hypothetical example of two groups, each with 600 people. The upper part of the table shows the numbers of these people with different values of the potential outcomes $Y_i(X)$ for a binary outcome Y, given the values $X = 0$ and $X = 1$ of a binary treatment X. The lower part of the table splits these counts between men and women. As measures of the effect of X on Y, the population-averaged odds ratios (OR) and risk differences (RD) are shown below each table.

	Group A:		*Group B:*	
	Men and women together:		*Men and women together*:	
	$Y_i(X) = 0$	$Y_i(X) = 1$	$Y_i(X) = 0$	$Y_i(X) = 1$
$X = 0$	390	210	390	210
$X = 1$	300	300	270	330
OR:		*1.86*		*2.27*
RD:		*0.15*		*0.20*

	Men:		*Women*:		*Men*:		*Women*:	
	$Y_i(X)=0$	$Y_i(X)=1$	$Y_i(X)=0$	$Y_i(X)=1$	$Y_i(X)=0$	$Y_i(X)=1$	$Y_i(X)=0$	$Y_i(X)=1$
$X = 0$	150	150	240	60	150	150	240	60
$X = 1$	100	200	200	100	90	210	180	120
OR:		*2.00*		*2.00*		*2.33*		*2.67*
RD:		*0.17*		*0.13*		*0.20*		*0.20*