Available online at www.sciencedirect.com

ScienceDirect

Journal of Memory and Language 59 (2008) 447–456

Journal of Memory and Language

www.elsevier.com/locate/jml

Models of accuracy in repeated-measures designs

Peter Dixon *

University of Alberta, Department of Psychology, Edmonton, AB, Canada T6G 2E9

Received 3 March 2007; revision received 9 November 2007
Available online 4 January 2008

Abstract

Accuracy is often analyzed using analysis of variance techniques in which the data are assumed to be normally distributed. However, accuracy data are discrete rather than continuous, and proportion correct are constrained to the range 0–1. Monte Carlo simulations are presented illustrating how this can lead to distortions in the pattern of means. An alternative is to analyze accuracy using logistic regression. In this technique, the log odds (or logit) of proportion correct is modeled as a linear function of the factors in the design. In effect, accuracy is rescaled in terms of a logit "response-strength" measure. Because the logit scale is unbounded, it is not susceptible to the same scaling artifacts as proportion correct. However, repeated-measures designs are not readily handled in standard logistic regression. I consider two approaches to analyzing such designs: conditional logistic regression, in which a Rasch model is assumed for the data, and generalized linear mixed-effect analysis, in which quasi-maximum likelihood techniques are used to estimate model parameters. Monte Carlo simulations demonstrate that the latter is superior when effect size varies over subjects.

Keywords: Accuracy; Logistic regression; Repeated-measures

Introduction

In research on language and many other areas in psychology, accuracy is often analyzed using analysis of variance techniques in which the data are assumed to be normally distributed. For example, in volume 26 of *Journal of Memory and Language*, 29 of 31 articles reported accuracy or similar categorical data. Of these, 23 or 79% used analysis of variance on untransformed response proportions. However, accuracy data are discrete rather than continuous, and proportion correct are constrained to the range 0–1. This can lead to averaging artifacts and distortions in the means, standard errors, and other statistics. An alternative is to analyze accuracy using logistic regression. In this technique, the log odds (or logit) of being correct is modeled as a linear function of the factors in the design. In effect, accuracy is rescaled in terms of a logit "response-strength" measure. In logistic regression, it is assumed that the underlying data are binomial, and because the logit scale is unbounded, it is not susceptible to the same scaling artifacts as proportion correct. However, special considerations apply to the use of logistic regression in repeated-measures designs. In the present paper, I first summarize some of the difficulties in using the normal model for analyzing accuracy and provide several arguments for using logistic regression instead. Possible distortions that arise from using the normal model are illustrated using Monte Carlo simulations. I then describe two ways of approaching

* Fax: +1 403 492 1768.
E-mail address: peter.dixon@ualberta.ca

0749-596X/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.jml.2007.11.004

repeated-measures designs: conditional logistic regression and generalized linear mixed-effects models. Monte Carlo simulations suggest that the latter approach is superior when the magnitude of the effects of interest varies over subjects.

Problems with the normal model of accuracy

It is readily apparent that accuracy data do not conform to the assumptions of the analysis of variance and related techniques (i.e., the "normal model"). To begin with, the response on each trial is dichotomous rather being continuous. Commonly, this is addressed by construing the data as the proportion correct in each condition. In some settings, though, there may be relatively few observations in each condition, and proportion correct will not come close to approximating a continuous distribution. Moreover, it is still the case that performance is constrained to be the range of 0–1. This leads to ceiling (or floor) effects when performance is very good (or very poor). Because of the constrained range, the variance across conditions will not be equal and instead will vary systematically with overall performance. For example, if observations in each condition are binomial, the variance in a condition with an accuracy rate of .80 will be proportional to $(.80)(.20) = .16$, while that in a condition with an accuracy of .60 will be proportional to $(.60)(.40) = .24$, an increase of 50%. This heteroscedasticity becomes even more pronounced as the accuracy level increases.

Because the normal model does not take into account the constrained range of proportions, one can sometimes derive statistical estimates that are nonsensical. For example, it is quite possible to find confidence interval limits that are larger than 1 (or smaller than 0). More generally, the conditions that make the least-squares solutions to linear models optimal under assumptions of normality do not apply to proportions. For example, the theoretical sampling distributions of estimates will not correspond to the actual distributions, and standard errors of estimates derived from the normal model will be incorrect. Averaging artifacts can also arise when the number of observations varies across conditions or subjects. In Simpson's paradox (Simpson, 1951), for example, the pattern of overall means may not reflect the pattern observed in individual conditions or with individual subjects. All of these issues suggest that accuracy would be better analyzed using tools that more closely match the nature of the data.

Logistic regression as an alternative

Many textbooks recommend logistic regression for dichotomous data with properties such as those of accuracy (e.g., Allison, 1999; Everitt, 2001). In logistic regression, the logit or log odds of being correct are assumed to be a linear function of the variables in the design:

$$\ln\left(\frac{P(C)}{1-P(C)}\right) = \text{logit}(P(C)) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

where the x_i are, for example, dummy variables coding the main effects and interactions in a factorial design. There is no closed form for estimating the regression coefficients, and these are instead estimated using incremental search procedures that maximize the likelihood of the data. Logistic regression is described as appropriate for the analysis of dichotomous data when there are two possible responses and several continuous or categorical predictors, and McCullagh (1980) suggests that logistic regression models are appropriate when the categorical responses can be construed as contiguous intervals on a continuous scale. It solves the problems of constrained range and heteroscedasticity described above, and it is immune to Simpson's paradox and related averaging artifacts. It is more mathematically tractable than some other alternatives and provides parameter estimates that are readily interpreted.

In addition to these convenience arguments favoring the use of logistic regression, one can also make fairly general theoretical arguments for the suitability of the approach for accuracy. One such argument is based on Luce (1963) choice theory (cf. McClelland, 1991).

Here, it is assumed that the correct and incorrect responses are associated with response strengths, s_C and s_E, and that the probability of selecting the correct response is given by the ratio of the correct response strength to the sum of the response strengths:

$$P(C) = \frac{s_C}{s_C + s_E}$$

Further, under some circumstances it may make sense to assume that processing variables have multiplicative effects on response strength (e.g., Luce, 1959; Townsend, 1971). Thus, one could write:

$$s_C = \prod_i \psi_{C,i}$$

where the ψ values indicate the various processes that affect response strength. With some rearrangement, the logit can then be written as:

$$\text{logit}(P(C)) = \ln\left[\frac{s_C}{s_E}\right] = \sum_i \psi'_i$$

where $\psi'_i = \ln(\psi_{C,i}/\psi_{E,i})$. Thus, the logit is a linear function of the processing components that determine the relative response strength. Logistic regression provides a tool for understanding these contributions. For example, experimental factors that affect components selectively will thus have additive effects in a logistic regression equation.

One may also make a related argument based on an analogy with signal detection theory. In this case, one might assume a distribution of response strengths for the correct response and the incorrect response. Each trial would then in effect be a two-alternative forced choice between the correct and the incorrect response, with subjects selecting the response with more potent strength. If the underlying strength distributions are normal, the probability of a correct response is:

$$P(C) = P(X_{\mathrm{C}} > X_{\mathrm{E}}) = F\left(\frac{\mu_{\mathrm{C}} - \mu_{\mathrm{E}}}{\sqrt{2}\sigma}\right)$$

where F is the cumulative standard normal distribution. A z-score corresponding to the probability correct thus provides an index of the separation of the two strength distributions. Given these assumptions, it would be appropriate to use probit regression to analyze the effects of experimental factors on this separation. Probit regression is similar to logistic regression except that the inverse cumulative normal, or probit, function replaces the logit function. Thus, the probit of the probability of correct is assumed to be a linear function of the predictor variables. However, the logistic and normal distributions have a very similar shape and differ significantly only in the tails. Consequently, logistic regression can be used to provide an (approximate) insight into the variables that determine relative response strength here as well.

There are also circumstances in which the logit transformation is closely related to substantive theories in a given domain. For example, Dixon and Twilley (1999) proposed a model of meaning resolution with ambiguous words in which meaning activation was a logistic function of perceptual and contextual input. Consequently, in this theory, logistic regression provides a suitable, theoretically guided analysis of meaning selection data (e.g., Twilley & Dixon, 2000). Similarly, McClelland (1991) described a related neural network model of speech perception in which the logistic was used as the activation function for individual units. Although his analysis was used to make somewhat different points, it does raise the possibility of applying logistic regression in that context as well.

Detecting interactions

In this section, I illustrate one of the problems in using the normal model in complex designs: distortions in the pattern of means. Because the accuracy scale is bounded, the pattern of means may not provide an informative reflection of the underlying processes if the levels of accuracy approach those bounds. In particular, as the level of accuracy increases, effects that are in principle additive may appear to exhibit an underadditive interaction, while data that derive from an overadditive interaction may appear to be additive.

Artifactual evidence for interactions

These distortions are illustrated with Monte Carlo simulations. A 2×2 design with 100 independent observations in each cell was simulated. Using a logistic regression model to generate the data, the probability of a correct response for factor levels A_i and B_j was:

$$p_{ij} = \mathrm{logit}^{-1}(\mu + \alpha_i + \beta_j + \gamma_{ij})$$

For all of the simulations, μ was set at 1.5; thus, in the absence of any effects, accuracy would be .818. In the first set of simulations, I was interested in evaluating the tendency for a normal model to provide evidence for an interaction where none exists. Consequently, γ was set to 0, and I considered a range of main effect magnitudes, m, from .2 to .8. In each case, $\alpha_1 = \beta_1 = m$ and $\alpha_2 = \beta_2 = -m$. To quantify the evidence for the interaction, I used the Akaike Information Criterion (AIC; Akaike, 1973). The AIC value provides a succinct description of the fit of a model relative to the number of degrees of freedom. Thus, it provides an index of the model's parsimony. For each set of simulated data, I computed the difference in the AIC value for a model with only main effects and a model that included an interaction. In this simple situation, the difference in AIC values is closely related to the obtained p value used in null hypothesis significance testing, and circumstances that lead to rejecting the null hypothesis would correspond to an AIC difference of about 2.

Programs that fit logistic regression models are widely available. In the present simulations, I used a generalized linear model approach to estimate the logistic regression parameters. In this technique, one identifies a link function that maps the parameters of any distribution in the exponential family to the parameters of the normal distribution. Iterative methods are then used to estimate the parameter values that maximize the likelihood of the data. The binomial distribution that forms the basis of logistic regression is a member of the exponential family, and as a consequence, logistic regression can be performed by using the logit function as the link function. It is worth noting that accuracy or similar data can be handled in a variety of ways in the context of generalized linear models. For example, if the identity is used as the link function, the approach provides another means of estimating the parameters of the normal model; if the probit function is used as the link function, the approach provides the fit of a probit regression model. However, the goal here was simply to use generalized linear models as a means to perform logistic regression; comparable results would be obtained with any of a wide range of other logistic regression tools.

In detail, the simulated data were fit with the R statistical language (R Development Core Team, 2006) using the generalized linear modeling program, glm. In glm and other R model-fitting programs, models are specified with a "formula" that describes the independent and dependent variables. For example, to fit a model in which the factors A and B are additive, one would use the command:

```
glm(Obs ~ A + B, family = binomial)
```

where Obs, A, and B are R vectors that hold the (binary) observations and the predictor variables for the A and B factors, and "family = binomial" indicates that the data are binomial and that the logistic link function should be used. To fit a model that includes the interaction, one would add the interaction to the formula, as in:

```
glm(Obs ~ A + B + A : B, family = binomial)
```

where the notation "A:B" indicates the interaction of the two factors. To assess the evidence for the interaction, one would compare the two model fits. In R, the results of two fits can be saved using the assignment operator, "←", and the AIC values can be subsequently extracted using the function "AIC":

```
additiveFit ← glm(Obs ~ A + B, family = binomial)
interactiveFit ← glm(Obs ~ A + B + A : B, family = binomial)
interactionEvidence ← AIC(additiveFit) − AIC(interactiveFit)
```

More detailed information about using the R language can be found in a variety of textbooks (e.g., Dalgaard, 2002; Everitt & Hothorn, 2006) and in sources listed on the R website (www.r-project.org).

The median AIC value and the interquartile range for the simulations is shown in Fig. 1. As can be seen, the normal model provides increasingly stronger evidence for the interaction as the magnitude of the main effects increases. This is because of ceiling effects that occur when the spread in the data is increased, which in turn produces an apparent underadditive interaction. No such effect is found with the logistic model since it matches the model used to produce the simulated data. Indeed, for the logistic model the difference in AIC values are negative, providing evidence for the (correct) additive interpretation.

Weak evidence for real interactions

Related to the tendency to produce evidence for an interaction where none exists is the tendency to obscure evidence for a real interaction. In particular, if the interaction is overadditive, the pattern of means will be distorted as observations approach the ceiling, and an analysis based on the normal model may suggest an additive interpretation. To demonstrate this tendency, m was fixed at .5, and an overadditive interaction of various magnitudes was added. In particular, $\gamma_{11} = \gamma_{22} = u$, $\gamma_{12} = \gamma_{21} = -u$, and u varied from .1 to .4. As before, 100 simulations were run at each value. In Fig. 2, the results are depicted in terms of the median and interquartile range of the difference in AIC values for the additive and interaction models. While the logistic model usually provided evidence for an interaction even with a magnitude of .3, the normal model failed to do so even with larger interactions. The culprit in these cases is that the normal model fails to account for the limited range of the accuracy data, and as consequence was more likely to suggest an additive interpretation when the overadditive interaction was constrained by the upper limit.

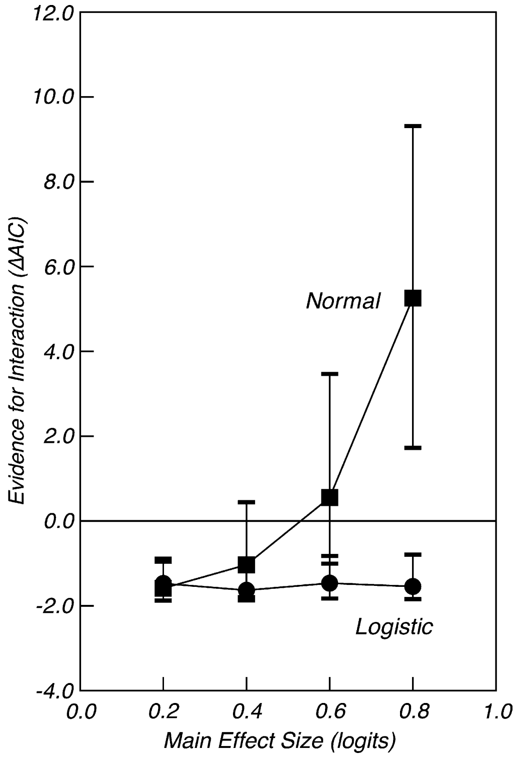

Fig. 1. Evidence provided for an interaction by normal and logistic regression models of additive data as a function of main effect size. Points indicate the median difference between AIC values for additive and interactive models in 100 Monte Carlo simulations, and the error bars indicate the interquartile range.

Discussion

The simulation results demonstrate that at least under some circumstances, evidence for an interaction is distorted by the normal model. As performance approaches the upper bound, the effects of any given variable become smaller when compared to those effects when performance is more moderate. As a consequence, an artifactual underadditive interaction may be apparent. Similarly, an underlying overadditive interaction may display a pattern of means that seems additive

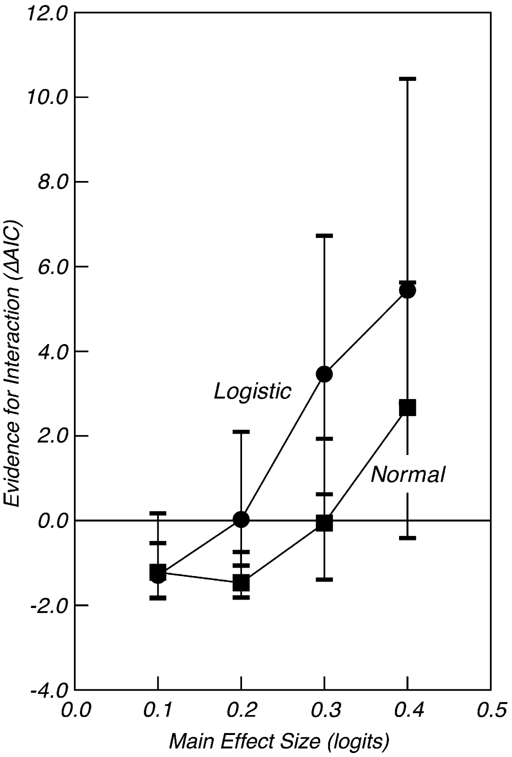

Fig. 2. Evidence provided for an interaction by normal and logistic regression models as a function of interaction magnitude. Points indicate the median difference between AIC values for additive and interactive models in 100 Monte Carlo simulations, and the error bars indicate the interquartile range.

because the highest levels of performance are constrained by the upper bound. Logistic regression eliminates these issues because, in effect, the data are recoded in terms of a response-strength measure that has no such constraints.

Of course, these scaling artifacts are well known, and a variety of techniques can be used to guard against them. For example, one may design experiments so that the accuracy does not approach the extremes and stays within a range of .25–.75 or so. This can be done by pilot testing materials and manipulations, and/or discarding subjects with performance outside of that range. However, this approach can be time consuming (because manipulations have to be carefully designed and tested) and inefficient (because data must be discarded). Moreover, this approach may not be feasible in some situations because the range of performance is determined by other considerations. For example, one may wish to analyze accuracy in a speeded choice task in which response time is the main dependent variable. Such tasks generally have to be designed so that accuracy is very high in order to make the response time analyses meaningful.

Another approach is to transform the accuracy data using an arcsine transformation to increase normality (e.g., Cohen & Cohen, 1983). While this reduces the tendency to provide artifactual evidence for interactions, it does not eliminate it. For example, in the simulations with main effects of .8 (i.e., the rightmost point in Fig. 1), the magnitude of apparent underadditive interaction was 42% of the size of the main effects. Applying an arcsine transformation to the means reduces this value to about 23%, but misleading interpretations would still be possible. Moreover, transformations of this general sort have other drawbacks. For example, unlike logits (or even proportion correct) the transformed means do not necessarily have a simple interpretation in terms of the underlying mechanisms. Moreover, because the choice of transformation is ad hoc, it may be difficult to defend that choice if it has a large effect on the pattern of means. Logistic regression is less susceptible to these concerns because it can be viewed as a theoretically justified transformation of proportion correct into a type of response-strength measure.

The present demonstrations used simulated data for which the logistic model was correct. Of course, it is possible to generate data using other models, and in these cases, logistic regression would not necessarily perform as well. In any given experimental context, there may be compelling theoretical reasons for assuming a model other than the logistic for the data. For example, multinomial modeling of various forms can be used when there are strong hypotheses concerning the component processes that determine performance and how they are combined (e.g., Batchelder & Riefer, 1999). Logistic regression would generally not provide precise information about those component processes. However, the normal model would be at least as inappropriate as an analysis tool in such situations. In the absence of strong claims about the determinants of accuracy, I argue that there are defensible theoretical analyses that would lead to the use of the logistic model in a wide range of circumstances. Thus, as a default assumption in the absence of more theoretically guided choices, the logistic model is superior to the normal model.

Distortion effects analogous to those illustrated here can arise even if condition means do not approach ceiling performance. For example, rather than independent observations in each condition, a design might include some number of observations from each of several subjects. In such a design, the performance of some subjects may be very high (or very low) and as a consequence, would be susceptible to the kinds of distortions investigated in Figs. 1 and 2. Thus, when averaged across subjects, the same distorted pattern might be apparent, even though the overall accuracy in a condition may be moderate. Techniques related to those described below would provide an approach that would be immune to these sorts of artifacts.

Repeated measures

Although logistic regression provides a reasonable approach to the analysis of accuracy data, it cannot be readily applied to data from repeated-measures designs. In particular, in standard logistic regression it is assumed that all of the observations in the design are independent. In repeated-measures designs, this assumption is violated because the performance of a given subject in one condition is typically correlated with that subject's performance in other conditions. Here, I consider two strategies for addressing this issue. In one, I use conditional logistic regression, inspired by the treatment of random person effects in item-response theory and Rasch models. In the second, I use generalized linear mixed-effects models. In this case, generalized linear models are extended to include an explicit specification of the random effects, and maximum likelihood methods are then used to estimate both the fixed and random effects.

Conditional logistic model

The Rasch model (1960/1980) is a development in item-response theory that provides a straightforward approach to analyzing random subject variation. As originally conceived, it provides a description of how different people perform with different items on (e.g.) an ability test. In particular, the probability of person i responding correctly to item j is assumed to be an inverse logistic function of a person parameter and an item difficulty parameter: $P(C) = \text{logit}^{-1}(\theta_i - \beta_j)$. To apply this development in the present context, one can construe the item parameter as a linear function of the experimental factors: $P(C) = \text{logit}^{-1} \ (\theta_i - (\mu + \alpha_j + \beta_k + \cdots))$. In that case, the Rasch model is identical to the logistic regression model but with the addition of a random subject term. Fischer and Molenaar (1995) discuss a variety of techniques for estimating the item and person parameters, including conditionalizing on the obtained performance of the subjects. A number of statistical packages perform the relevant conditional logistic regression; one example is the clogit program in the survival package in R (R Development Core Team, 2006). However, in many cases, a close approximation can obtained by using standard, unconditional logistic regression (with subjects included as a fixed effect) and then adjusting the parameter estimates by the factor $(n - 1)/n$, where n is the number of observations per subject (Fischer & Molenaar, 1995). The critical assumption in using conditional logistic regression model is that the random effect of subjects is limited to an overall variation in performance and does not interact with the effects of interest. Thus, after conditionalizing on the (random) contribution of subjects, each of the observations can be assumed to be independent, and the interpretation of the logistic regression parameters can proceed as before.

There is a close parallel between this approach and the sphericity assumption in traditional repeated-measures analysis of variance. If the sphericity assumption is correct, one may model the data from a single-factor design as:

$$X_{ij} = \mu + \alpha_i + S_j + \varepsilon_{ij}$$

where S_j is the random contribution of subject j and ε is the independent error. Observations from such a model are not all independent because the presence of the S_j term introduces a correlation between pairs of observations from the same subject. However, just as in the Rasch model, conditionalizing on the overall performance of subjects eliminates this correlation and leads to a set of independent observations. The sphericity assumption is reasonably accurate in many experimental settings, and, by extension, one might conjecture that applying the Rasch model to accuracy data would be similarly appropriate in many cases. Of course, there are also situations in which the sphericity assumption is clearly incorrect, and I consider the implications of comparable violations for the conditional logistic regression below.

Generalized linear mixed-effects models

Although the conditional logistic model incorporates the assumption that subjects are randomly sampled, the approach fails to address situations in which the magnitude of an effect varies over subjects. This shortcoming can be addressed by using mixed-effects, or multi-level, models. In a mixed-effects model, one specifies not only the fixed effects, but also effects that vary randomly. In a repeated-measures design, the level of performance for subjects varies randomly, and potentially the magnitude of the effects of interest may vary across subjects as well. The magnitude of both of these sources of variation would be estimated. Mixed-effects models can be used in the context of generalized linear models, allowing one to fit logistic regression models with random effects. I use the term "linear mixed-effects models" to refer to this approach in the present context.

Random subject effects

In this section, I compare the two approaches to logistic regression in repeated-measures designs. In the first set of simulations, I evaluate how they are affected by subject variability and compare their performance to that of the normal model. The simulated data consisted of 12 subjects, each of which made 50 responses in each of two conditions. The probability of a correct response was given by:

$$p_{ij} = \mathrm{logit}^{-1}(\mu + \alpha_i + S_j + A_{ij})$$

where S and A are, respectively, the random overall effect of subjects and the random variation in the effect magnitude over subjects. The random effects were normally distributed with zero mean and standard deviations σ_S and σ_A, respectively. The linear mixed-effects model was fit using the program lmer in the R package lme4 (Bates & Sarkar, 2006) using the default penalized quasi-likelihood method for estimating the parameters.

In using the lmer program in R, one specifies a model for the data using a formula much like that for glm. However, the formula must also include a specification of the random effects. For example, if one assumed that subjects varied randomly in their overall performance but that the effect of the factor of interest was constant, one would use the following:

$$\mathrm{lmer(Obs \sim A + (1|S),\ family = binomial)}$$

The term "(1|S)" indicates that there is a random constant term that should be estimated given each value of S. These terms are constrained to sum to zero and in effect are estimates of the S_j terms in the model specified above. If one assumed that the magnitude of factor A could also vary randomly across subjects, the model would be fit as:

$$\mathrm{lmer(Obs \sim A + (1 + A|S),\ family = binomial)}$$

This adds an additional term (the random variation in A) for each subject, which again are constrained to sum to zero. Importantly, the choice of random effects structure need not be done a priori, and one could select one or the other based on how well they account for the data. Baayen (in press), for example, provides further details concerning the use of lmer for repeated-measures designs.

For the first set of simulations, I was interested in the effect of variability in overall performance across subjects. In this case, μ was set to zero, $\alpha_1 = -\alpha_2 = 1$, and σ_A was assumed to be 0 (i.e., A was identically zero). Two indices of the behavior of the model were assessed: bias in the estimate of the effect and bias in the estimated standard error. The first reflects how accurate the model estimation procedures are, and the second whether the estimation procedures provide an accurate measure of the variability. To compute bias in the estimate, 100 simulations were performed, and a ratio was formed between the mean estimate and the actual value used to generate the data. To compute bias in the estimated standard error, the mean (across the 100 simulations) of the standard error provided by the model fitting procedure was compared to the actual standard deviation of those estimates in the sample. The estimates and standard errors produced by conditional logistic regression and linear mixed-effects are in logits. To provide a fair comparison, the estimates and standard errors for the normal model were converted to logits as well.

Figs. 3 and 4 depict the simulation results for values of σ_S ranging from .0 to 2.0. The results demonstrate that both conditional logistic regression and linear mixed-effects analysis provide an improvement over the normal model. As subject variability increases, the normal model tends to underestimate the magnitude of the effect. This is because with increasing variability, the likelihood that some subjects run up against the ceiling or floor increases, and the effect size for those subjects will be smaller because of the constrained range. The normal model fails to account for this scaling artifact, and as a consequence the estimate of the effect (averaged across subjects) will also be smaller. For the same reason, the normal model overestimates the standard error of the estimate: for the normal model, the effect size will vary not only because of the inherent variability in the data, but also because of the scaling artifacts that vary with subjects' overall performance. These results suggest that the normal model will lack power unless the variability in subjects' performance is small. Neither of the two alternative models are affected in this way because the data of subjects high or low in overall accuracy are rescaled appropriately in terms of logits. However, the conditional logistic seems to provide somewhat more precise measures of the effect when

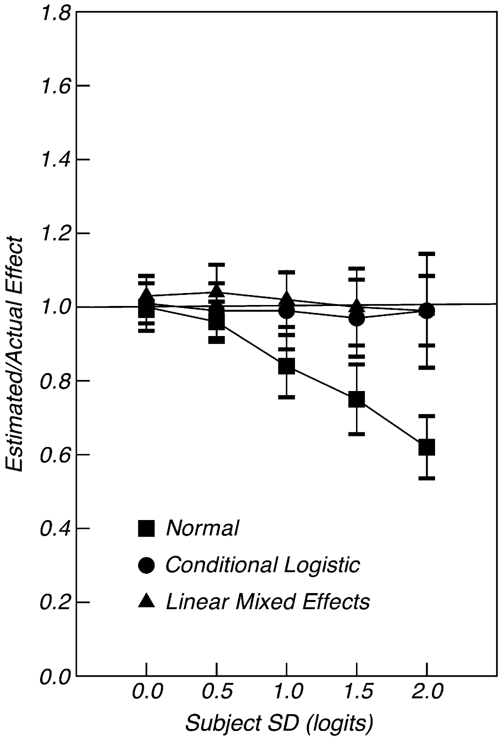

Fig. 3. Bias in the estimate of the effect size as a function of the variability in subjects' performance. Error bars depict the standard deviation in 100 Monte Carlo simulations.

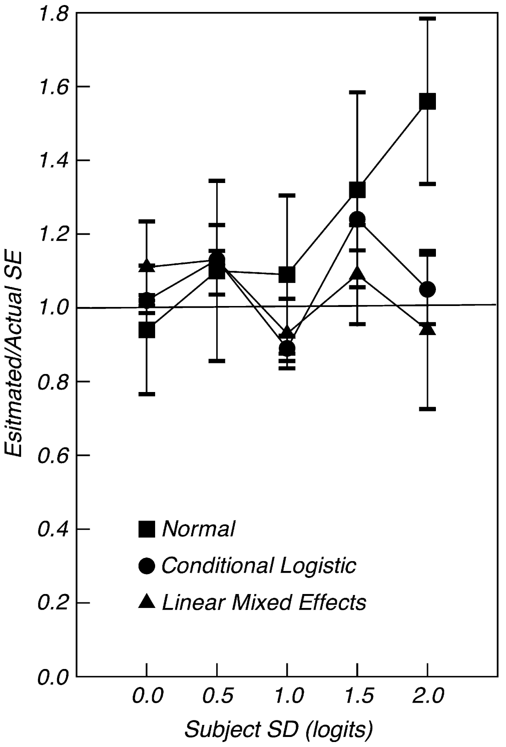

Fig. 4. Bias in the estimate of the effect size standard error as a function of the variability in subjects' performance. Error bars depict the standard deviation in 100 Monte Carlo simulations.

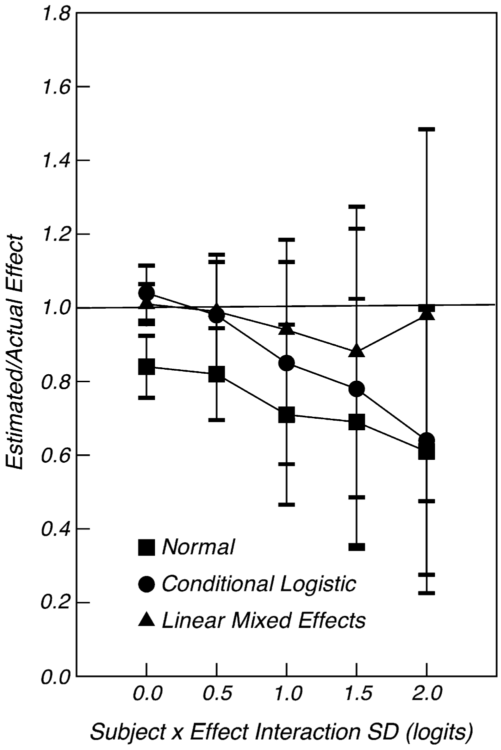

Fig. 5. Bias in the estimate of the effect size as a function of the variability in effect size over subjects. Error bars depict the standard deviation in 100 Monte Carlo simulations.

variability is high: with $\sigma_S = 2.0$, the (actual) standard error of the estimate was .08 for conditional logistic regression compared to .11 for the linear mixed-effects model.

Random subject interactions

The most serious test of the conditional logistic model is the assessment of its behavior when the assumption of no random variation in effect size is relaxed. To evaluate the models under such circumstances, σ_S was fixed at 1.0, and σ_A was varied from .0 to 2.0. The results for bias in the estimate of effect size is shown in Fig. 5. Both conditional logistic regression and the normal model tend to underestimate the effect size in the face of large variations in effect size. The behavior of the linear mixed-effects model also becomes more variable, but there is less evidence of bias.

Bias in the standard error of the estimate is shown in Fig. 6. In this case, the conditional logistic model drastically underestimates the standard error of the estimate as the effect size becomes more variable. This is an intuitive result: As the magnitude of the subject × effect interaction increases, the size of the effect in any given sample becomes more variable, and as a consequence, the estimate of the population effect size becomes less stable. However, there is no subject × effect term in the conditional logistic model, and consequently, there is no means for incorporating this source of variability into its estimates. Thus, with increasing values of σ_A, the estimated standard error remains the same even though the actual variability of the estimate increases substantially. Although the magnitude of the estimate becomes somewhat smaller (as shown in Fig. 5), this does not compensate for the large underestimation of the standard error. It is clear that conditional logistic regression is not an appropriate approach when the effect can be assumed to vary across subjects.

Generalized estimating equations (GEEs) provide another approach to fitting models with random effects (Liang & Zeger, 1986). In the GEE approach, one directly estimates the marginal means, averaged over subjects, without considering the likelihood of the original scores. One advantage of the approach is that one need not specify the structure of the random effects, and only minimal assumptions are needed to produce consistent parameter estimates. In the simple situation used in the simulations presented here, the GEE approach performs similarly to the linear mixed-effects approach. However, GEEs have some disadvantages.

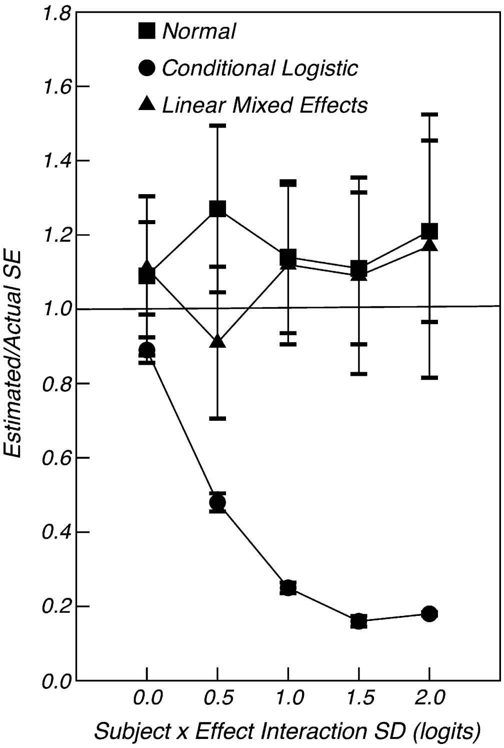

Fig. 6. Bias in the estimate of the effect size standard error as a function of the variability in effect size over subjects. Error bars depict the standard deviation in 100 Monte Carlo simulations.

Because they are not based on the likelihood of the data, one cannot derive measures of model parsimony such as the AIC. This means that it may be difficult to compare different models, particularly if those models are not nested. Because they are based on the marginal means, GEE model fits are also susceptible to averaging artifacts such as Simpson's paradox when the data are unbalanced. Lindsey and Lambert (1998) argue that the apparently minimal assumptions in the approach hides important constraints on the form of the underlying probability model. Thus, GEEs may not form a general solution to the analysis of accuracy or similar data.

General discussion

In this paper, I have argued that using the normal model to analyze accuracy data (or similar dichotomous data) is inappropriate. Many authors have noted a variety of defects in this approach (e.g., Allison, 1999; Everitt, 2001). Here, I illustrated how the normal model can distort the pattern of means, so that (for example) evidence for additive and interactive effects may be artifactual. I argued that logistic regression provides an effective alternative to the use of the normal model. Logistic regression can be defended on ad hoc grounds since it provides an appropriate analysis of dichotomous data. Moreover, a plausible theoretical argument can be made that logistic regression represents the data in terms of a more meaningful response-strength measure.

One difficulty with using logistic regression in many contexts is that standard logistic regression cannot be applied to repeated-measures designs. I described two solutions to this problem: applying the Rasch model commonly used in item-response theory (resulting in a conditional logistic model), and using a (generalized) linear mixed-effects model. In the simulations reported here, both provide reasonable results when the effect size can be assumed to be constant across subjects, and the conditional logistic model may have a small advantage in precision under some circumstances. If the effects of interest interact with subjects, though, conditional logistic regression can severely underestimate the standard error of the estimate. Under such circumstances, the linear mixed-effects approach is preferred.

Using logistic regression in the analysis of real-world designs requires a distinct analysis strategy from that commonly used in analysis of variance. Generally, the estimates of the different effects and interactions in which one might be interested are not independent, and they will typically vary depending on what other variables are included in the analysis. For example, in an additive, two-factor model, the estimates for the main effects would be different than they would be in the fit of a model that included the interaction. Thus, one cannot simply fit a "full" model that includes all possible effects and interactions and expect that the estimates for any given subset of effects will be appropriate. Because a full model usually includes a number of variables that have negligible effects, the results are overfitted, and, as a consequence, estimates of even important effects may be distorted. This is not the case in the normal model with a balanced design because the estimates of the effects are all independent of one another. Thus, an appropriate strategy with logistic regression models is to proceed incrementally by adding effects one at a time until the most parsimonious fit is obtained. In this sense, using logistic regression is more akin to the techniques and methods of inference used in hierarchical linear regression.

I have framed the issues here in terms of accuracy since accuracy is likely the single most common form of dichotomous variable encountered by experimental psychologists. However, my comments apply to a range of other comparable variables, such as preference or choice responses, strategy selection, and other categorically coded behaviors. In many of these situations, it would be inappropriate to use the normal model as an analysis technique, and logistic regression provides a useful alternative. Because logistic regression tools are readily available, there would seem to be no compelling reason for the use of the normal model for accuracy or similar data.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.jml.2007.11.004.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov & F. Csaki (Eds.), *Second international symposium on information theory*. Budapest: Academiai Kiado.

Allison, P. D. (1999). *Logisitic regression using the SAS system: Theory and application*. Cary, NC: SAS Institute, Inc..

Baayen, R. H. (in press). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University.

Batchelder, W. H., & Riefer, D. M. (1999). Theoretical and empirical review of multinomial process tree modeling. *Psychonomic Bulletin & Review, 6*, 57–86.

Bates, D., & Sarkar, D. (2006). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.9975-6.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/ correlation analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Dalgaard, P. (2002). *Introductory statistics with R*. New York: Springer.

Dixon, P., & Twilley, L. C. (1999). Context and homograph meaning resolution. *Canadian Journal of Experimental Psychology, 53*, 335–346.

Everitt, B. S. (2001). *Statistics for psychologists: An intermediate course*. Hillsdale, NJ: Erlbaum.

Everitt, B. S., & Hothorn, T. (2006). *A handbook of statistical analyses using R*. Boca Raton, FL: Chapman & Hall.

Fischer, G. H., & Molenaar, I. W. (1995). *Rasch models: Foundations, recent developments, and applications*. New York: Springer-Verlag.

Liang, K.-Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika, 73*, 13–22.

Lindsey, J. K., & Lambert, P. (1998). On the appropriateness of marginal models for repeated measurements in clinical trials. *Statistics in medicine, 17*, 447–469.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Luce, R. D. (1963). Detection and recognition. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.). *Handbook of mathematical psychology* (Vol. I). New York: Wiley.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of the Royal Statistical Society, Series B, 42*, 109–142.

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23*, 1–44.

R Development Core Team (2006). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Available from http://www.R-project.org.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests (Expanded edition)*. Copenhagen: Danish Institute of Educational Research, Chicago: University of Chicago Press (Original work published 1960).

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13*, 238–241.

Townsend, J. J. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics, 9*, 40–50.

Twilley, L. C., & Dixon, P. (2000). Meaning resolution processes for words: A parallel independent model. *Psychonomic Bulletin & Review, 7*, 49–82.