Running head: Categorical Data Analysis

Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models

T. Florian Jaeger

Brain and Cognitive Sciences, University of Rochester

Contact address:

Assistant Professor

Brain and Cognitive Sciences,

University of Rochester,

Meliora Hall, Box 270268,

Rochester, NY 14627-0268

P: (585) 276 3611

E: fjaeger@bcs.rochester.edu

U: http://www.bcs.rochester.edu/people/fjaeger/

Abstract

This paper identifies several serious problems with the widespread use of ANOVAs for the analysis of categorical outcome variables such as forced-choice variables, question-answer accuracy, choice in production (e.g. in syntactic priming research), et cetera. I show that even after applying the arcsine-square-root transformation to proportional data, ANOVA can yield spurious results. I discuss conceptual issues underlying these problems and alternatives provided by modern statistics. Specifically, I introduce ordinary logit models (i.e. logistic regression), which are well-suited to analyze categorical data and offer many advantages over ANOVA. Unfortunately, ordinary logit models do not include random effect modeling. To address this issue, I describe mixed logit models (Generalized Linear Mixed Models for binomially distributed outcomes, Breslow & Clayton, 1993), which combine the advantages of ordinary logit models with the ability to account for random subject and item effects in one step of analysis. Throughout the paper, I use a psycholinguistic data set to compare the different statistical methods.

Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models

In the psychological sciences, training in the statistical analysis of continuous *outcomes* (i.e. *responses* or *independent variables*) is a fundamental part of our education. The same cannot be said about *categorical data analysis* (Agresti, 2002; henceforth CDA), the analysis of outcomes that are either inherently categorical (e.g. the response to a *yes*/*no* question) or measured in a way that results in categorical grouping (e.g. grouping neurons into different bins based on their firing rates). CDA is common in all behavioral sciences. For example, much research on language production has investigated influences on speakers' choice between two or more possible structures (see e.g. research on syntactic persistence, Bock, 1986; Pickering and Branigan, 1998; among many others; or in research on speech errors). For language comprehension, examples of research on categorical outcomes include eye-tracking experiments (first fixations), picture identification tasks to test semantic understanding, and, of course, comprehension questions. More generally, any kind of forced-choice task, such as multiple-choice questions, and any count data constitute categorical data.

Despite this preponderance of categorical data, the use of statistical analyses that have long been known to be questionable for CDA (such as analysis of variance, ANOVA) is still commonplace in our field. While there are powerful modern methods designed for CDA (e.g. ordinary and mixed logit models; see below), they are considered too complicated or simply *unnecessary*. There is a widely-held belief that categorical outcomes can safely be analyzed using ANOVA, if the arcsine-square-root transformation (Cochran, 1940; Rao, 1960; Winer et al., 1971) is applied.

This belief is misleading: *even ANOVAs over arcsine-square-root transformed proportions of categorical outcomes* (see below) *can lead to spurious null results and spurious significances*. These spurious results go beyond the normal chance of Type I and Type II errors. The arcsine-square-root and other transformations (e.g. by using the *empirical logit* transformation, Haldane, 1955; Cox, 1970) are simply approximations that were primarily intended to reduce costly computation time. In an age of cheap computing at everyone's fingertips, we can abandon ANOVA for CDA. Modern statistics provide us with alternatives that are in many ways superior.

This paper provides an informal introduction to one such method: generalized linear mixed models with a logit link function, henceforth *mixed logit models* (Bates & DebRoy, 2004; Bates & Sarkar, 2007; Breslow & Clayton, 1993; see also conditional logistic regression, Dixon, this issue; for an overview of other methods, see Agresti, 2002). Mixed logit models are a generalization of logistic regression. Like ordinary logistic regression (Cox, 1958, 1970; Dyke & Patterson, 1952; henceforth ordinary logit models), they are well-suited for the analysis of categorical outcomes. Going beyond ordinary logit models, however, mixed logit models include random effects, such as subject and item effects. I introduce both ordinary and mixed logit models and compare them to ANOVA over untransformed and arcsine-square-root transformed proportions using data from a psycholinguistics study (Arnon, 2006, submitted). All analyses were performed using the statistics software package R (R Development Core Team, 2006). The R code is available from the author.

The inadequacy of ANOVA over categorical outcomes

Issues with ANOVAs and, more generally, linear models over categorical data have been known for a long time (e.g. Cochran, 1940; Rao, 1960; Winer et al., 1971; for summaries, see Agresti, 2002: 120; Hogg & Craig, 1995). I discuss problems with the interpretability of ANOVAs over categorical data and then show that these problems stem from conceptual issues.

Interpretability of ANOVA over categorical outcomes

ANOVA compares the *means* of different experimental conditions and determines whether to reject the hypothesis that the conditions have the same population means given the observed sample *variances* within and between the conditions. For continuous outcomes, the means, variances, and the confidence intervals have straightforward interpretations. But what happens if the outcome is categorical? For example, we may be interested in whether subjects answer a question correctly depending on the experimental condition. So, we may observe that of the 10 elicited answers, 8 are *correct* and 2 are *incorrect*. What is the mean and variance of 8 correct answers and 2 incorrect answers? We can code one of the outcomes, e.g. correct answers, as 1 and the other outcome, e.g. wrong answers, as 0. In that case, we can calculate a mean (here 0.8) and variance (here 0.18). The mean is apparently straightforwardly interpreted as the mean proportion of correct answers (or percentages of correct answers if multiplied by 100).

The current standard for CDA in psychology follows the aforementioned logic. Categorical outcomes are analyzed using subject and item ANOVAs (F1 and F2) *over proportions or percentages*. The approach is seemingly intuitive and, by now, so widespread that it is hard to imagine that there is any problem with it. Unfortunately, that is not the case. ANOVAs over

proportions can lead to hard-to-interpret results because confidence intervals can extend beyond the interpretable values between 0 and 1. For the above example, a 95% confidence interval would range from 0.52 to 1.08 (= 0.8 +/- 0.275), rendering an interpretation of the outcome variable as a proportion of correct answers impossible (proportions above 1 are not defined). One way to think about the problem of interpretability is that ANOVAs attribute probability mass to events that can never occur, thereby likely underestimating the probability mass over events that actually *can* occur. This intuition points at the most crucial problem with ANOVAs over proportions of categorical outcomes. ANOVA over proportions easily leads to spurious results.

Categorical outcomes violate ANOVA's assumption

The inappropriateness of ANOVAs over categorical data can be derived on theoretical grounds. Assume a binary outcome (e.g. correct or incorrect answers to *yes*/*no*-questions) that is binomially distributed; that is, for every trial there is a probability p that the answer will be correct. Then the probability of k correct answers in n trials is given by the following function:

$$(1) \qquad f(k,n,p) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

The population mean and variance of a binomially distributed variable X are given in (2) and (3).

$$(2) \qquad \mu_X = n[1p + 0(p-1)] = np$$

$$(3) \qquad \sigma_X^2 = n\left[(1-p)^2 p + (0-p)^2 (1-p)\right] = np(1-p)$$

The expected sample proportion P over n trials is given by dividing μ_X by the number of trials n, and hence is p. Similarly, the variance of the sample proportion is a function of p:

$$(4) \qquad \sigma_P^2 = \frac{p(1-p)}{n}$$

From (4) it follows that the variance of the sample proportions will be highest for $p = 0.5$ (the product of n numbers x that add up to 1 is highest if $x_1 = \ldots = x_n$) and will decrease symmetrically as we approach 0 or 1. This is illustrated in Figure 1. Note that the shape of the curve and the location of its maximum are determined by p alone.

[insert Figure 1 here]

Now assume that we have two samples elicited under different conditions. In one condition, the probability that a trial will yield a correct answer is p_1, in the other condition it is p_2. For example, if $p_1 = 0.45$ and $p_2 = 0.8$, then:

$$(5) \qquad \sigma_P^2(p_1) = \frac{p_1(1-p_1)}{n} = \frac{0.2475}{n} > \frac{0.16}{n} = \frac{p_2(1-p_2)}{n} = \sigma_P^2(p_2)$$

In other words, if the probability of an outcome differs between two binomially distributed conditions, the variances will only be identical if p_1 and p_2 are equally far away from 0.5 (e.g. $p_1 = 0.4$ and $p_2 = 0.6$). The bigger the difference in distance from 0.5 between the two conditions, the less similar the variances will be. Also, as can be see in Figure 1, the differences close to 0.5 will matter less than differences closer to 0 or 1. Even if p_1 and p_2 are unequally distant from 0.5, as long as they are relatively close to 0.5, the variances of the sample proportions will be similar. Sample proportions between 0.3 and 0.7 are considered close enough to 0.5 to assume homogeneous variances (e.g. Agresti, 2002: 120). Within this interval, $p(1-p)$ ranges from 0.21 for $p = 0.3$ or 0.7 to 0.25 for $p = 0.5$. Unfortunately, we usually cannot determine *a priori* the range of sample proportions in our experiment (see also Dixon, this issue). Also, in general,

variances in two binomially distributed conditions will not be homogeneous – contrary to the assumption of ANOVA.

The inappropriateness of ANOVA for CDA was recognized as early as Cochran (1940, referred to in Agresti, 2002: 596). Before I discuss the most commonly used method for CDA using ANOVA over *transformed* proportions, I introduce logistic regression, which is an alternative to ANOVA that was designed for the analysis of binomially distributed categorical data.

An alternative: Ordinary logit models (logistic regression)

Logistic regression, also called ordinary logit models, was first used by Dyke and Patterson (1952), but was most widely introduced by Cox (1958, 1970; see Agresti, 2002: Ch. 16). For extensive formal introductions to logistic regression, I refer to Agresti (2002: Ch 5), Chatterjee and colleagues (2000: Ch. 12), and Harrell (2001). For a concise formal introduction written for language researchers, I recommend Manning (2003: 5.7).

Logit models can be seen to be a specific instance of a generalization of ANOVA. To see this link between logit models and ANOVA, it helps to know that ANOVA can be understood as linear regression (cf. Chatterjee, 2000: Ch. 5). Linear regression describes outcome y as a linear combination of the independent variables $x_1 \ldots x_n$ (also called *predictors*) plus some random error ε (and optionally an intercept β_0). This is usually stated in one of three equivalent ways:

$$(6) \qquad y = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n + \varepsilon \Leftrightarrow \mathrm{E}(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$

The first equation describes the value of y. The second and third equations describe the expected value of y. Note that categorical predictors have to be recoded into numerical values for (6) to

make sense (treatment-coding, also called dummy-coding, being the most common coding). We can abbreviate (6) using vector notation $E(y) = \boldsymbol{x\beta}$ (I use boldface for vectors), where $\boldsymbol{x}$ is a transposed vector consisting of 1 for the intercept, and all predictor values $x_1 \ldots x_n$, and $\boldsymbol{\beta}$ is a vector of coefficients $\beta_0 \ldots \beta_n$. A frequently used and even more compact notation describes an entire data set using matrix notation $E(\boldsymbol{y}) = \boldsymbol{X\beta}$ (I use capital letters for matrices). Each row of the matrix $\boldsymbol{X}$ corresponds to the transposed vector $\boldsymbol{x}$ of a case in the data and $\boldsymbol{y}$ is the vector of outcomes. The coefficients $\beta_0 \ldots \beta_n$ have to be estimated. This is done in such a way that the resulting model fits the data 'optimally'. Usually, the model is considered optimal if it is the model for which the actually observed data are most likely to be observed (the maximum likelihood model; for an informal introduction, see Baayen et al., this issue: Appendix A).

Now imagine that we want to fit a linear regression to proportions of categorical outcomes. So, we could define the following model of expected proportions:

(7) $E(p) = \boldsymbol{x\beta}$ or for the entire data set: $E(\boldsymbol{p}) = \boldsymbol{X\beta}$

Such a linear model, also called *linear probability model* (Agresti, 2002: 120), has many of the same problems mentioned above for ANOVAs over proportions. But, what if we transformed proportions into a space that is not bounded by 0 and 1 and that captures the intuition that changes around 0.5 weigh less than changes close to 0 or 1? Odds are such a space. They are easily derived from probabilities (and vice versa):

(8) $odds(p) = \frac{p}{1-p}$ and $p(odds) = \frac{odds}{1+odds}$

Thus, odds increase with increasing probabilities, with odds of 1 corresponding to a proportion of 0.5. Differences in odds are usually described multiplicatively (i.e. in terms of *x*-fold increases or decreases). For example, the odds of being on a plane with a drunken pilot are reported to be "1 to 117" (http://www.funny2.com/). In the notation used here, this corresponds to odds of 1 / 117 ≈ 0.0086. Unfortunately, these odds are 860 times higher than the odds of dating a supermodel (≈ 0.00001). Thus, we can describe the odds of an outcome as a product of coefficients raised to the respective predictor values (assuming treatment-coding, predictor values are either 0 or 1):

$$(9) \qquad \mathrm{E}\left[\frac{p}{1-p}\right] = \beta_0 * \beta_1^{x_1} * \cdots * \beta_n^{x_n}$$

By simply taking the natural logarithm of odds instead of plain odds, we can turn the model back into a linear combination, which has many desirable properties:

$$(10) \qquad \mathrm{E}\left[\ln\frac{p}{1-p}\right] = \ln\left(\beta_0 * \beta_1^{x_1} * \cdots * \beta_n^{x_n}\right) = \ln(\beta_0) + \ln(\beta_1)x_1 + \cdots + \ln(\beta_n)x_n$$

The natural logarithm of odds is called the *logit* (or log-odds). The logit is centered around 0 (i.e. $\mathrm{logit}(p) = -\mathrm{logit}(1-p)$), corresponding to a probability of 0.5, and ranges from negative to positive infinity. The $\ln \beta_0$ ... $\ln \beta_n$ in (10) are constants, so we can substitute β_0 ... β_n for them (or any other arbitrary variable name). This yields (11):

$$(11) \qquad \mathrm{E}\left[\ln\frac{p}{1-p}\right] = \mathrm{E}[\mathrm{logit}\ p] = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n = \boldsymbol{x\beta}$$

In other words, we can think of ordinary logit models as linear regression in logit space! The logit function defines a transformation that maps points in probability space into points in log-

odds space. In probability space, the linear relationship that we see in logit space is gone. This is apparent in (12), describing the same model as in (11), but transformed into probability space:

$$(12) \qquad \mathrm{E}[p] = \frac{e^{x\beta}}{1+e^{x\beta}} = \frac{1}{1+e^{-x\beta}}$$

Logit models capture the fact that differences in probabilities around $p = 0.5$ matter less than the same changes close to 0 or 1. This is illustrated in Figure 2, where the left panel shows a hypothetical linear effect of a predictor x in logit space ($y = -3 + 0.2x$), and the right panel shows the same effect in probability space. As can be seen in the right panel, small changes on the x-axis around $p = 0.5$ (i.e. $x = 15$ since $0 = -3 + 0.2 * 15 = \text{logit}(0.5)$) lead to large decreases or increases in probabilities compared to the same change on the x-axis closer to 0 or 1.

[insert Figure 2 here]

Thus logit models, unlike ANOVA, are well-suited for the analysis of binomially distributed categorical outcomes (i.e. any event that occurs with the same probability at each trial). Logit models have additional advantages over ANOVA. Logit models scale to categorical dependent variables with more than two outcomes (in which case we call the model a *multinomial model*; for an introduction, see Agresti, 2002). Among other things, this can help avoid confounds due to data exclusion. For example, in priming studies where researchers are interested in speakers' choice between two structures, subject sometimes produce neither of those two. If non-randomly distributed, such "errors" can confound the analysis because what appears to be an effect on the choice between two outcomes may, in reality, be an effect on the chance of an error. Consider a scenario in which, for condition X, participants produce 50% outcome 1, 45% outcome 2, and 5% errors, but, for condition Y, they produce 50% outcome 1, 30% outcome 2, and 20% errors. If an analysis was conducted after errors are excluded, we may conclude, given small enough

standard errors, that there is a main effect of condition (in condition X, the proportion of outcome 1 would be 50/95 = 0.53; in condition Y, 50/80 = 0.63). This conclusion would be misleading, since what really happens is that there is an effect on the probability of an error. We would find a spurious main effect on outcome 1 vs. 2. The problem is not only limited to errors. It also includes any case in which "other" categories are excluded from the analysis (e.g. when speakers in a production experiment produce structures that we are not interested in). Multinomial models make such exclusion unnecessary and allow us to test which of *all* possible outcomes a given predictor affects. For the above example, we could test whether the condition affects the probability of outcome 1 or outcome 2, or the probability of an error.

Logit models also inherit a variety of advantages from regression analyses. They provide researchers with more information on the directionality and size of an effect than the standard ANOVA output (this will become apparent below). They can deal with imbalanced data, thereby freeing researchers from all too restrictive designs that affect the naturalness of the object of their study (see Jaeger, 2006 for more details). Like other regressions, ordinary logit models also force us to be explicit in the specification of assumed model structure. At the same time, regression models make it easier to add and remove additional post-hoc control in the analysis, thereby giving researchers more flexibility and better post-hoc control. Another nice feature that logit models inherit from regressions is that they can include continuous predictors. Modern implementations of logit models come with a variety of tools to investigate linearity assumptions for continuous predictors (e.g. *rcs* for restricted cubic splines in R's *Design* library; Harrell, 2005). Ordinary logit models do, however, have a major drawback compared to ANOVA: they do not model random subject and item effects. Later I describe how *mixed* logit models

overcome this problem. First I present a case study that exemplifies the problems of ANOVA over proportions using a real psycholinguistic data set. The case study illustrates that these problems persist even if arcsine-square-root transformed proportions are used in the ANOVA.

A case study: Spurious significance in ANOVA over proportions

Arnon (2006, submitted) conducts several experiments to test whether locality affects children's production and comprehension of relative clauses in the same way as it has been shown to affect adults' performance (e.g. Gibson, 1998). I consider only parts of the comprehension results of Arnon's Study 2. In this 2 x 2 experiment, twenty-four Hebrew-speaking children listened to Hebrew relative clauses (RCs). RCs were either subject or object extracted. The noun phrase in the RC (the object for subject extracted RCs and the subject for object extracted RCs) was either a first person pronoun or a lexical noun phrase (NP). An example item in all four conditions is given in Table 1 (taken from Arnon, 2006), where the manipulated NP is underlined.

[insert Table 1 approximately here]

Arnon hypothesized that children, just like adults (Warren & Gibson, 2003), should (a) have a harder time understanding object RCs than subject RCs, and (b) perform worse on the RCs with full lexical NPs than on RCs with pronoun NPs. The comprehension data Arnon collected support her hypothesis (Arnon's conclusions are based on the results of a mixed logit model). Table 2 summarizes the mean question-answer accuracy (i.e. the proportion of correct answers) and standard errors across the four conditions.

[insert Table 2 approximately here]

Note that, contrary to the assumption of the homogeneity of variances, but as expected for binomially distributed outcomes, the standard errors (and hence the variances) are bigger the

closer the mean proportion of correct answers is to 50%. The results in Table 2 also suggest that an ANOVA will find main effects of RC type and NP type as well as an interaction. Question-answer accuracy is higher for subject RCs than for object RCs (92.7% vs. 76.6%) and higher for pronoun NPs than for lexical NPs (90.0% vs. 79.3%). Furthermore, the effect of NP type on the percentage of correct answers seems to be bigger for object RCs (68.9% vs. 84.3%) than for subject RCs (89.7% vs. 95.7%), suggesting that an ANOVA will find an interaction.

ANOVA over untransformed proportions

Indeed, subject and item ANOVAs over the average percentages of correct answers return significance for both main effects and the interaction.

[insert Table 3 approximately here]

As expected the interaction comes out as highly significant in the ANOVA. Now, are these effects spurious or not? In the previous section, I discussed several theoretical issues with ANOVAs over proportions. But do those issues affect the validity of these ANOVA results? As I show next, the answer is *yes*, *they do*.

Ordinary logit model

Ordinary logit models are implemented in most modern statistics program. I use the function *lrm* in R's *Design* library (Harrell, 2005). The model formula for the R function *lrm* is given in (13).

(13) `Correct~ 1 + RCtype + NPtype + RCtype:NPtype`

The "1" specifies that an intercept should be included in the model (the default). Further shortening the formula, I could have written `Correct ~ RCtype*NPtype`, which in R implies inclusion of all combinations of the terms connected by "*" (I will use this notation below).

For the ordinary logit model, the analyzed outcomes are the correct or incorrect answers. Thus, all cases are entered into the regression (instead of averaging across subjects or items). Significance of predictors in the fitted model is tested with likelihood ratio tests (Agresti, 2002: 12). Likelihood ratio tests compare the data likelihood of a subset model with the data likelihood of a superset model that contains all of the subset model's predictors and some more. A model's data likelihood is a measure of its quality or fit, describing the likelihood of the sample given the model. The -2 * logarithm of the ratio between the likelihoods of the models is asymptotically χ^2-distributed with the difference in degrees of freedoms between the two models. Thus a predictor's significance in a model is tested by comparing that model against a model without the predictor using a χ^2-test.

Here I use the function *anova.Design* from R's *Design* library (Harrell, 2005). The function automatically compares a model against all its subset models that are derived by removing exactly one predictor. For Arnon's data, we find that a model without RC type has considerably lower data likelihood ($\chi^2(1)= 28.8$, $p< 0.001$), as does a model without NP type ($\chi^2(1)= 12.2$, $p< 0.001$). Thus RC and NP type contribute significant information to the model. The interaction, however, does not ($\chi^2(1)= 0.01$, $p> 0.9$). The summary of the full model in Table 4 confirms this.

[insert Table 4 approximately here]

Note that the standard summary of a regression model provides information about the size and directionality of effects (an ANOVA would require planned contrasts for this information). The first column of Table 4 lists all the predictors entered into the regression. The second column gives the estimate of the coefficient associated with the effect. The coefficients have an intuitive

geometrical interpretation: they describe the slope associated with an effect in log-odds (or logit) space. For categorical predictors, the precise interpretation depends on what numerical coding is used. Treatment-coding compares each level of a categorical predictor against all other levels. This contrasts with effect-coding, which compares two levels against each other. Here I have used treatment-coding, because it is the most common coding scheme in the regression literature. For example, for the current data set, subject RCs are coded as 1 and compared against object RCs (which are taken as the baseline and coded as 0). So, the coefficient associated with RC type tells us that the log-odds of a correct answer for subject RCs are 1.35 log-odds higher than for object RCs. But what does this mean? Recall that log-odds are simply the log of odds. So, the odds of a correct answer for subject RCs are $e^{1.35} \approx 3.9$ times higher than the odds for object RCs. Following the same logic, the odds for RCs with pronouns are estimated to be $e^{0.89} \approx 2.4$ times higher than the odds for RCs with lexical NPs.

The third column in Table 4 gives the estimate of the coefficients' standard errors. The standard errors are used to calculate Wald's z-score (henceforth Wald's Z, Wald, 1943) in the fourth column by dividing the coefficient estimate by the estimate for its standard error. The absolute value of Wald's Z describes how distant the coefficient estimate is from zero in terms of its standard error. The test returns significance if this standardized distance from zero is large enough. Coefficients that are significantly smaller than zero decrease the log-odds (and hence odds) of the outcome (here: a correct answer). Coefficients significantly larger than zero increase the log-odds of the outcome. Unlike the likelihood ratio test, however, Wald's z-test is not robust in the presence of collinearity (Agresti, 2002: 12). Collinearity leads to inflated estimates of the standard errors and changes coefficient estimates (although in an unbiased way). The model

presented here contains only very limited collinearity because all predictors were centered (VIFs < 1.5).[1] This makes it possible to use the coefficients to interpret the direction and size of the effects in the model.

The main effects of RC type and NP type are highly significant. We can also interpret the significant intercept. It means that, if the RC type is not 'subject RC' and the NP type is not 'pronoun', the chance of a correct answer in Arnon's sample is significantly higher than 50%. The odds are estimated at $e^{0.8} \approx 2.2$, which means that the chance of a correct answer for object RCs with a lexical NP is estimated as $p = \frac{2.2}{1+2.2} \approx 0.69$. Indeed, this is what we have seen in Table 2. Similarly the predicted probability of a correct answer for subject RC with a pronoun is calculated by adding all relevant log-odds, $0.8 + 1.35 + 0.89 = 3.04$, which gives $p = \frac{e^{3.04}}{1+e^{3.04}} \approx 0.95$ (compared to 95.7% given in Table 2).

The numbers do not quite match because we did not include the coefficient for the interaction. However, notice that they *almost* match. This is the case because the interaction does not add significant information to the model (Wald's Z=0.01, P > 0.9). The effects are illustrated in Figure 3, showing the predicted means and confidence intervals for all combinations of RC and NP type (the plot uses *plot.Design* from R's *Design* library, Harrell, 2005):

[insert Figure 3 approximately here]

In sum, there is no significant interaction because the effect of NP type for different levels of RC type does not differ in odds (and hence neither does it differ in log-odds). Indeed, both the change from 68.9% to 84.3% associated with NP type for object RCs and the change from 89.7%

to 95.7% associated with NP type for subject RCs correspond to an approximate 2.5-fold odds increase. So, unlike ANOVA, logistic regression returns a result that respects the nature of the outcome variable.

The spurious results of the ANOVA should be of no further surprise given the before-mentioned conceptual problems. At this point, readers familiar with transformations for proportional data may find the argument against ANOVA spurious because they believe that ANOVAs will return correct results once the data is adequately transformed. In the next section I describe why this assumption is wrong for at least the most commonly used transformation.

The arcsine-square-root transformation and its failure

There are several problems with the reliance on transformation for ANOVA over proportional data. To begin with, it is unclear how strictly journals enforce the use of transformations – few psycholinguistic papers with categorical dependent variables mention transformation. There is also reason to doubt that transformations are always applied correctly. For example, the most popular transformation, the so called *arcsine transformation*, or more accurately *arcsine-square-root transformation* ($t(x) = \arcsin\left(\sqrt{x}\right)$; e.g. Rao, 1960; Winer et al., 1971) requires further modifications for small numbers of observations or proportions close to 0 or 1 (0 or 100% for percentages, respectively; e.g. Hogg & Craig, 1995). Bartlett (1937: 168, footnote) proposes that proportions of 0 should be converted to $1/4n$ before applying the transformation and proportions of 1 should be converted to $(n\text{-}1/4)/n$. In praxis these modifications are rarely applied (Victor Ferreira, p.c.), despite the fact that sample proportions close to 0 or 1 are common in behavioral research (e.g. in research on speech errors or when analyzing comprehension accuracies). Even

more worrisome is the lack of a theoretical justification for the arcsine-square-root transformation (cf. Cochran, 1940: 346). Most importantly, however, even ANOVA over transformed proportions can lead to spurious results. I illustrate this again using Arnon's data.

Spurious significance persists even after arcsine-square-root transformation

I limit myself to the subject analysis, since this is where the insufficiency of the arcsine-square-root transformation shows up most clearly. As can be seen in Table 5, the interaction is still incorrectly considered significant ($p< 0.01$). This is the case because several children in Arnon's experiment performed close to ceiling (the proportions of correct answers are 1 or close to 1). ANOVAs over arcsine-square-root transformed data are unreliable for such data sets.

[insert Table 5 approximately here]

One reason why the arcsine-square-root transformation is unreliable for such data becomes apparent once we compare the plots of logit and arcsine-square-root transformed proportions. Figure 5 shows the two transformations plotted against probabilities. Figure 6 shows the slope (1^{st} derivative) and curvature (2^{nd} derivative) of the two transformations. Both transformations have a saddle point at $p= 0.5$, but for all $p\neq 0.5$ the slope of the logit is always higher than the slope of the arcsine-square-root. The absolute curvature (the change in the slope) is also larger. In other words, as one moves away from $p= 0.5$, a change in probability p_1 to p_2 corresponds to more of a change in log-odds than to a change in arcsine-square-root transformed probabilities. This means that, compared to the logit, the arcsine-square-root transformation underestimates changes in probability more the closer they are to 0 or 1.

[insert Figure 5 and 6 approximately here]

Now consider the actual mean proportions for the four conditions in Arnon's data and the corresponding logit and arcsine-square-root transformed values given in Table 6 (cf. Table 2). While it does not make sense to compare the absolute transformed values, we can compare the differences in the differences. In logit space the effect of NP type corresponds to an increase of 0.88 for object RCs and 0.94 for subject RCs – a difference in the effects of 6.8%. This difference can be thought of as the interaction (it describes the super-additivity over the two main effects). Ignoring variance for now, we can say that the bigger this difference is, the more of a potential interaction effect there is. In arcsine-square-root space, the effect of NP type corresponds to an increase of 0.18 for object RCs and 0.12 for subject RCs – a difference of 50%! At the end, significance of difference is determined by the amount of variance within and between the conditions, but what the above comparison shows is that the arcsine-square-root transformation does not attribute as much 'weight' to changes in proportions close to 1 as the logit transformation does. Intuitively, while the arcsine-square-root transformation makes proportional data more similar to what it would look like in logit space (the difference of the differences for proportions would be over 150%), it does not go quite far enough.

[insert Table approximately here]

In sum, for proportional data with proportions close to 0 or 1, even ANOVA over arcsine-square-root transformed data can return spurious results, while logistic regression does not. As mentioned earlier, this problem is not limited to spurious *significances*. Imagine the effect of NP type would be identical in proportions for subject and object RCs (e.g. imagine Arnon's data but with 74.9% correct answers for object RCs with pronouns): in proportions there would seem to be no interaction, but we may find one in logit space (granted sufficiently small standard errors).

At this point, one may ask whether there are any better transformations that would allow us to continue to use ANOVA for CDA. Several such transformations have been proposed, the most well-known being the *empirical logit* (first proposed by Haldane, 1955, but often attributed to Cox, 1970). The idea behind such transformations is to stay as close as possible to the actual logit transformation while being defined for 0 and 1 (for an empirical comparison of different logit estimates, see Gart and Zweifel, 1967). Indeed, appropriate transformations combined with appropriate weighing of cases mostly avoid the problems of ANOVA described above (for weighted linear regression that deals with heterogeneous variances, see McCullagh and Nelder, 1989). However, it is important to note that even these transformations are still ad-hoc in nature (which transformation works best depends on the actual sample the researcher is investigating, Gart & Zweifel, 1967). Transformations for categorical data were originally developed because they provided a computationally cheap *approximation* of the more adequate logistic regression – approximations that are no longer necessary.

This leaves one potential argument for the use of ANOVA (with transformations) for CDA: the fact that ordinary logit models provide no direct way to model random subject and item effects. The lack of random effect modeling is problematic as repeated measures on the subject or item in our sample constitute violations of the assumption that all observations in our data set are independent of one another. Data from the same subject or item is often referred to as a cluster. Analyses that ignore clusters produce invalid standard errors and therefore lead to unreliable results. Next I show that *mixed* logit models address this problem (other methods include separate logistic regressions for each subject/item, see Lorch and Meyers, 1990, or bootstrap sampling with random cluster replacement, see Feng et al., 1996).

Mixed logit models

Mixed logit models are a type of *Generalized Linear Mixed Models* (Breslow & Clayton, 1993; Lindstrom & Bates, 1990; for a formal introduction, see Agresti, 2002). Mixed Models with different link functions have been developed for a variety of underlying distributions. Mixed *logit* models are designed for binomially distributed outcomes.

Linear mixed models (Pinheiro & Bates, 2000; for an introduction, see Baayen et al., this issue) describe an outcome as the linear combination of fixed effects (described by $\boldsymbol{X\beta}$) and conditional random effects associated with e.g. subject and items (described by $\boldsymbol{Zb}$) of plus noise $\boldsymbol{\varepsilon}$. The random effects are characterized by a multivariate normal distribution, the variances and covariances of which are described by $\boldsymbol{\Sigma}$ (for more detail, see Baayen et al., this issue). The random effects are assumed to be independent of the random noise $\boldsymbol{\varepsilon}$.

(14) $$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}), \quad \boldsymbol{b} \sim N(0, \sigma^2\boldsymbol{\Sigma}), \quad \boldsymbol{b} \perp \boldsymbol{\varepsilon}$$

and, similarly for a mixed logit model:

(15) $$\text{logit}(\boldsymbol{p}) = \boldsymbol{X\beta} + \boldsymbol{Zb} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(0, \sigma^2\mathbf{I}), \quad \boldsymbol{b} \sim N(0, \sigma^2\boldsymbol{\Sigma}), \quad \boldsymbol{b} \perp \boldsymbol{\varepsilon}$$

Just as ordinary logit models are fit by finding the maximally likely coefficient estimates, mixed logit models are fitted to the data in such a way that the resulting model describes the data optimally. However, unlike for mixed *linear* models, there are no known analytic solutions for the exact optimization of mixed logit models' data likelihood (Hausman & Harding, 2006:2; Bates, 2007: 29). Instead, simulation methods, such as Monte Carlo simulations, are used to find optimal fits. Another computationally more efficient method (especially for larger data sets)

maximizes analytic approximations of the likelihood function, so called *quasi-log-likelihood* (for the trade-offs of both approaches, see Agresti, 2002: 523-524). Here I use the *lmer* function from R's *lme4* library (Bates & Sakar, 2007), using Laplace approximation to maximize quasi-log-likelihood (Bates, 2007: 29). Laplace approximation "performs extremely well, both in terms of numerical accuracy and computational time" (Hausman & Harding, 2006: 19).

A case study using mixed logit models

The model formula is specified in (16), where the term in parentheses describes the random subject effects for the intercept, the effects of RC and NP type, and their interaction. Random effects are assumed to be normally distributed (in log-odds space) around a mean of zero. The only parameter the model fits for the random effects is their variance (see also Baayen et al., this issue; for details on the implementation, see Bates & Sakar, 2007). The random intercept captures potential differences in children's base performance. The other random effects capture potential differences between children in terms of how they are affected by the manipulations.

(16) `Correct ~ 1 + RCtype * NPtype + (1 + RCtype * NPtype | child)`

The estimated fixed effects are summarized in Table 7. The number of observations and the quasi-log-likelihood of the model are given in the table's caption. The estimated variances of the random effects are summarized in Table 8.

[insert Table 7 and 8 approximately here]

In sum, a mixed logit model analysis of the data from Arnon (submitted) confirms the results from the ordinary logit model presented above. Even after controlling for random subject effects, the interaction between RC type and NP type is not significant. Note that the total correlation between the random interaction and effect of NP type for subjects in Table 8 suggests that the

model has been overparameterized (cf. Baayen et al., this issue) – one of the two random effects is redundant. I get back to this shortly, when I show that we can further simplify the model.

Additional advantages of mixed logit models

Mixed logit models combine all the advantages of ordinary logit models with the ability to model random effects, but that's not all. Mixed logit models do not make the frequently unjustified assumption of the homogeneity of variances. Also, the R implementation of mixed logit models used here (*lmer*) actually maximizes *penalized* quasi-log-likelihood (Bates, 2007: 29). Penalization adds a term punishing large coefficient values to the function that is being maximized by the fitting algorithm. This makes overfitting of the model to the sample less likely, thereby making it more likely that the model describes generalizations over the entire population (Agresti, 2002: 524). Guarding against overfitting is especially relevant for unbalanced data sets that result from data loss. Additionally, the specific method used for fitting here, Laplace approximation, is known to provide great numerically accuracy (Hausman & Harding, in press). Indeed, simulations show that *lmer*'s quasi-likelihood optimization outperforms ANOVA in terms of accurately estimating effect sizes and standard errors (Dixon, this issue). In other words, mixed logit models have greater power than ANOVA and therefore more likely to detect true effects.

Another advantage of mixed models is that they allow us to test rather than to stipulate whether a hypothesized random effect should be included in the model. The question of whether or to what extent random subject and items effects (especially the latter) are actually necessary has been the target of an ongoing debate (Clark, 1974; Raaijmakers et al., 1999, a.o.). As Baayen et al. (this

issue) demonstrate, mixed models can be used to test a hypothesized random effect. The test follows the same logic that was used above to test fixed effects: we simply compare the likelihood of the model with and the model without the random effect. Before I illustrate this for the mixed logit model from Table 7 and 8, a word of caution is in order. Comparisons of models via quasi-log-likelihood can be problematic, since quasi-likelihood are approximations (see above). This problem is likely to become less of an issue as the employed approximations become better (for discussion, see Bates & Sakar, 2007). In any case, we can use quasi-log-likelihood comparisons between models to get an idea of how much evidence there is for a hypothesized random effect.

As mentioned above, the correlation between the random subject effects in Table 8 shows that some of the random effects are redundant. Indeed, model comparisons suggest that neither the random effect for the interaction nor the random effect for NP type is justified. The quasi-log-likelihood decreases only minimally (from -256.8 to -258.5) when these two random effects are removed. A revised mixed logit model without random effects for NP type and the interaction is specified in (17). Table 9 and Table 10 give the updated results.

(17) `Correct ~ 1 + RCtype * NPtype + (1 + RCtype | child)`

[insert Table 9 and 10 approximately here]

Note that most fixed effect coefficients have not changed much – neither compared to the full mixed logit model in (16), nor compared to the ordinary logit model in (13). In all models the main effects are significant but the interaction is not. Only the coefficient of RC type differs between the current mixed logit model and the ordinary logit model: it is quite a bit larger in the

current model, but note that the standard error has also gone up. Wald's Z for RC type does not differ much between the two models. In summary, if there are random subject effects associated with NP type or the interaction of RC and NP type (e.g. if children in the sample differ in terms of how they react to NP type), they would seem to be subtle.

Finally, mixed logit models inherit yet another advantage from the fact that they are a type of generalized linear mixed model. They allow us to conduct *one combined* analysis for many independent random effects. For example, we could include random intercepts for both subjects and items in the model:

(18) `Correct ~ 1 + RCtype * NPtype + (1 + RCtype | child) + (1 | item)`

If a fixed effect is significant in such a model, this means it is significant after the variance associated with subject and items is simultaneously controlled for. In other words, mixed logit models can combine F1 and F2 analysis (for more detail and further examples for linear mixed models, see Baayen et al., this issue). Here only a random intercept (rather than random slopes for RC type, etc.) is included for items, because all further random effects are highly correlated with the random intercept (rs > 0.8) and hence unnecessary. The resulting model is summarized in Table 11 and Table 12. The minimal change in the quasi-log-likelihood, and the small estimates for the item variance, suggest that item differences do not account for much of the variance. Note that despite the fact that two items had missing cells and had to be excluded from the ANOVA, the current model uses all 8 items and 24 subjects in Arnon's data.

[insert Table 11 and 12 approximately here]

Combining subject and item analyses into one unified model is efficient and conceptually desirable (cf. Clark, 1973). Note that, in principle, mixed models are even compatible with

random effects beyond subject and item effects (e.g. if the children spoke different dialects and we hypothesized that this matters, we could include a random effect for dialect).

Conclusions

I have summarized arguments against the use of ANOVA over proportions of categorical outcomes. Such an analysis – regardless of whether the proportional data is arcsine-square-root transformed – can lead to spurious results. With the advent of mixed logit models, the last remaining valid excuse for ANOVA over categorical data (the inability of ordinary logit models to model random effects) no longer applies. Mixed logit models combine the strengths of logistic regression with random effects, while inheriting a variety of advantages from regression models. Most crucially, mixed models avoid spurious effects and have more power (Dixon, this issue). Finally, they form part of the generalized linear mixed model framework that provides a common language for analysis of many different types of outcomes.

Acknowledgments

This work was supported by a post-doctoral fellowship at the Department of Psychology, UC San Diego (V. Ferreira's NICHD grant R01 HD051030). For helpful comments on earlier drafts, I thank C. Kidd, A. Frank, D. Barr, P. Buttery, V. Ferreira, R. Levy, E. Norcliffe, H. Tily, and P. Chesley, as well as the audiences at the Center for Language Research (UC San Diego), the Center for Language Science (University of Rochester), and the LSA Summer Institute (held at Stanford University).

References

Agresti. A. (2002). *Categorical data analysis* (2nd Edition). New York, NY: John Wiley & Sons.

Arnon, I. (2006). Re-thinking child difficulty: The effect of NP type on child processing of relative clauses in Hebrew. Poster presented at *The 9th Annual CUNY Conference on Human Sentence Processing*, CUNY, March 2006

Arnon, I. (submitted). *Re-thinking child difficulty: The effect of NP type on child processing in Hebrew*.

Baayen, R. H., Davidson, D. J. and Bates, D. M. (submitted). *Mixed-effects modeling with crossed random effects for subjects and items*. Submitted to JML.

Bartlett, M. S. (1937). Some examples of statistical methods of research in agriculture and applied biology. *Supplement to the Journal of the Royal Statistical Society 4(2)*, 137-183.

Bates, D. M. (2007). *Linear mixed model implementation in lme4*. Ms., University of Wisconsin - Madison, August 2007.

Bates, D. M. and Sarkar, D. (2007). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.9975-12.

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association 39*, 357-365.

Bock, J. K. (1986). Syntactic Persistence in Language Production. *Cognitive Psychology 18*, 355–387.

Breslow, N.E. and Clayton, D. G. (1993). Approximate Inference in Generalized Linear Mixed Models. *Journal of the American Statistical Society 88(421)*, 9-25.

Chatterjee, S., Hadi, A., and Price, B. (2000). *Regression analysis by example*. New York: John Wiley & Sons, Inc.

Clark, H.H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research, *Journal of Verbal Learning and Verbal Behavior, 12,* 335–359.

Cochran, W.G. (1940). The analysis of variances when experimental errors follow the Poisson or binomial laws. *The Annals of Mathematical Statistics 11,* 335-347.

Cox, D. R. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society, Series B, 20*, 215-242.

Cox, D. R. (1970). *The analysis of binary data (2nd ed. 1989 by D. R. Cox and E. J. Snell).* London: Chapman & Hall.

DebRoy, S. and Bates, D. M. (2004). Linear mixed models and penalized least squares. *Journal of Multivariate Analysis 91(1)*, 1-17.

Dyke, G. V. and Patterson, H. D. (1952). Analysis of factorial arrangements when the data are proportions. *Biometrics 8*, 1-12.

Gart, J.J. and Zweifel, J.R. (1967). On the bias of various estimators of the logit and its variance with application to quantal bioassay. *Biometrika 54 (1 & 2)*, 181-187.

Haldane, J. B. S. (1955). The estimation and significance of the logarithm of a ratio of frequencies. *Annals of Human Genetics 20*, 309-311.

Hausman, J. and Harding, M.C. (in press). Using a Laplace Approximation to Estimate the Random Coefficients Logit Model by Non-linear Least Squares. *International Economic Review, 48(4).*

Harrell, F. E. Jr. (2001). *Regression modeling strategies*. New York: Springer.

Harrell, F. E. Jr. (2005). *Design: Design Package. R package version 2.0-12.* http://biostat.mc.vanderbilt.edu/s/Design, http://biostat.mc.vanderbilt.edu/rms

Hogg, R. and Craig, A. T. (1995) *Introduction into mathematical statistics*. Englewood Cliffs, NJ: Prentice Hall.

Jaeger, T. F. (2006). *Redundancy and Syntactic Reduction in Spontaneous Speech*. Ph.D. thesis, Stanford University.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics 46(3)*, 673-687.

Lorch, R. F. and Myers, J. L. (1990). Regression analyses of repeated measures data in cognitive research. *Journal of Experimental Psychology: Learning, Memory, and Cognition 16(1)*, 149—57

Manning, C. D. (2003). Probabilistic Syntax. In Bod, R., Hay, J. and Jannedy, S. (eds.) *Probabilistic Linguistics*, 289-341. Cambridge, MA: MIT Press.

McCullagh. P., and J. A. Nelder. (1989). *Generalized linear models*. New York, NY: Chapman and Hall.

Pickering, M. J. and Branigan, H. P. (1998). The Representation of Verbs: Evidence from Syntactic Priming in Language Production. *Journal of Memory and Language 39*, 633–651.

Pinheiro, J. C. and Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS*. New York: Springer.

Quené, H. & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: a tutorial. *Speech Communication 43*, 103–121.

Raaijmakers, J. G. W., Schrijnemakers, J. M. C., and Gremmen, F. (1999). How to deal with ”the language-as-fixed-effect-fallacy”: Common misconceptions and alternative solutions. Journal of Memory and Language, 41, 416–426.

R development core team (2005). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, http://www.R-project.org.

Rao, M. M. (1960). *Some asymptotic results on transformations in the analysis of variance*. ARL Technical Note, 60-126. Aerospace Research Laboratory, Wright-Patterson Air Force Base.

Wald, A. (1943). Test of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society 54(3)*, 426-482.

Winer, B.J., Brown, D. R., and Michels, K. M. (1971). Statistical *principles in experimental design*. New York: McGraw-Hill.

Appendices

[Insert appendices here]

Footnotes

[1] Collinearity is more of a concern in unbalanced data sets, but even in balanced data sets it can cause problems (for example, interactions and their main effects are often collinear even in balanced data sets). R comes with several implemented measures of collinearity (e.g. the function *kappa* as a measure of a model's collinearity; or the function *vif* in the *Design* library, which gives variance inflation factors – a measure of how much of one predictor is explained by the other predictors in the model). R also provides methods to remove collinearity from a model: from simple centering and standardizing (see the functions *scale*) to the use of residuals or principal component analysis (PCA, see the function *princomp*).

Table 1

Materials from Study 2 in Arnon (2006; Comprehension experiment)

Subject RC,	Eize tzeva ha-naalaim shel ha-yalda she metzayeret et ha-axot?
Lexical NP	Which color the-shoes of the-girl that draws the nurse-ACC
	What color are the shoes of the girl that is drawing <u>the nurse</u>?
Object RC,	Eize tzeva ha-naalaim shel ha-yalda she ha-axot metzayeret?
Lexical NP	Which color the-shoes of the-girl that the nurse draws?
	What color are the shoes of the girl that <u>the nurse</u> is drawing?
Subject RC,	Eize tzeva ha-naalaim shel ha-axot she metzayeret oti?
Pronoun	Which color the-shoes of the-nurse that draws me-ACC?
	What color are the shoes of the nurse that is drawing <u>me</u>?
Object RC,	Eize tzeva ha-naalaim shel ha-axot she ani metzayeret?
Pronoun	Which color the-shoes of the-nurse that I-NOM draw?
	What color are the shoes of the nurse that <u>I</u> am drawing?

Table 2

Percentage of correct answers and standard errors by condition

	Lexical NP	Pronoun NP
Subject RC	89.7% (.02)	95.7% (.02)
Object RC	68.9% (.04)	84.3% (.03)

Table 3

Summary of the ANOVA results over untransformed data

	Subject analysis		Item analysis		Combined	
	F1(1,23)	P	F2(1,5)	P	minF(1,10)	P
RC type	24.2	<0.01	10.2	<0.03	7.2	<0.03
NP type	16.1	<0.01	19.7	<0.01	8.9	<0.01
Interaction	9.7	<0.01	12.6	<0.02	5.5	<0.04

Table 4

Summary of the ordinary logit model (N= 696; model Nagelkerke r^2= 0.126)

Predictor	Coefficient	SE	Wald Z	P
Intercept	0.80	(0.167)	4.72	<0.001
RC type=*subject RC*	1.35	(0.295)	4.58	<0.001
NP type=*pronoun*	0.89	(0.272)	3.26	<0.001
Interaction=*subject RC & pronoun*	0.05	(0.511)	0.10	>0.9

Table 5

Summary of the ANOVA results over arcsine-square-root transformed data

	F1(1,23)	P
RC type	28.5	<0.01
NP type	17.3	<0.01
Interaction	8.5	<0.01

Table 6

Proportions and their logit and arcsine-square-root transforms for the four conditions in Arnon (2006: Study 2)

	Object RC		Subject RC	
	Lexical NP	Pronoun NP	Lexical NP	Pronoun NP
Proportions *P*	0.689	0.843	0.897	0.957
Logit(*P*)	0.80	1.68	2.16	3.10
Arcsine $\sqrt{P}$	0.98	1.16	1.24	1.36

Table 7

Summary of the fixed effects in the mixed logit model (N= 696; log-likelihood= -256.2)

Predictor	Coefficient	SE	Wald Z	P
Intercept	0.84	(0.203)	4.17	<0.001
RC type=*subject RC*	1.82	(0.365)	4.97	<0.001
NP type=*pronoun*	1.05	(0.288)	3.66	<0.001
Interaction=*subject RC & pronoun*	0.59	(0.580)	1.02	>0.3

Table 8

Summary of random subject effects and correlations in the mixed logit model

Random subject effect	s^2	Correlation with random effect for Intercept	RC type	NP type
Intercept	0.283			
RC type=*subject RC*	0.645	0.625		
NP type=*pronoun*	0.010	0.800	0.459	
Interaction=*subject RC & pronoun*	0.221	0.800	0.459	1.000

Table 9

Summary of the fixed effects in the mixed logit model (N= 696; log-likelihood= -256.8)

Predictor	Coefficient	SE	Wald Z	P
Intercept	0.86	(0.212)	3.99	<0.001
RC type=*subject RC*	1.90	(0.380)	5.01	<0.001
NP type=*pronoun*	0.96	(0.278)	3.44	<0.001
Interaction=*subject RC & pronoun*	0.10	(0.544)	0.18	>0.8

Table 10

Summary of random subject effects and correlations in the mixed logit model

Random subject effect	s^2	Correlation with random effect for Intercept	RC type	NP type
Intercept	0.399			
RC type=*subject RC*	0.744	0.629		

Table 11

Summary of the fixed effects in the mixed logit model (N= 696; log-likelihood= -256.0)

Predictor	Coefficient	SE	Wald Z	P
Intercept	0.85	(0.244)	3.49	<0.001
RC type=*subject RC*	1.97	(0.385)	5.11	<0.001
NP type=*pronoun*	0.99	(0.283)	3.49	<0.001
Interaction=*subject RC & pronoun*	0.07	(0.550)	0.13	>0.8

Table 12

Summary of random subject and item effects and correlations in the mixed logit model

		Correlation with random effect for		
Random effect	s^2	Intercept	RC type	NP type
Subject intercept	0.420			
Subject RC type=*subject RC*	0.770	0.620		
Item intercept	0.086			

Figure captions

Figure 1: Variance of sample proportion depending on p (for n= 1)

Figure 2: Example effect of predictor x on categorical outcome y. The left panel displays the effect in logit space with $\ln \frac{1}{1-p(x)} = -3 + 0.2x$. The right panel displays the same effect in probability space with $p(x) = \frac{1}{1+e^{3-0.2x}}$

Figure 3: Estimated effects of RC type and NP type on the log-odds of a correct answer.

Figure 4: Proportions plotted against their logit transform (left panel) and arcsine-square-root transform (right panel)

Figure 5: Slope and curvature of the logit and arcsine-square-root transformation

Figure 1

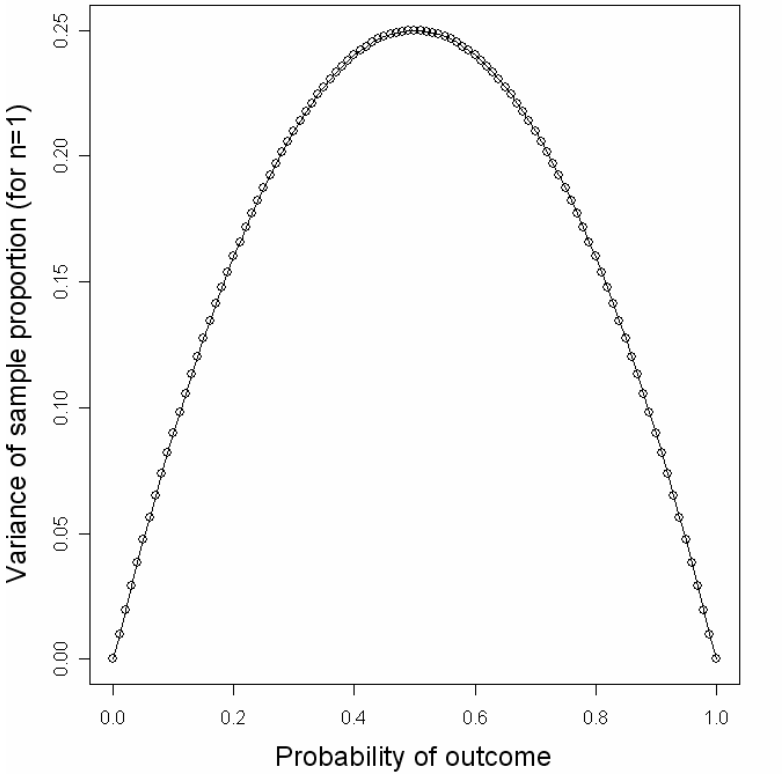

Variance of sample proportion (for n=1)
0.00
0.05
0.10
0.15
0.20
0.25
0.0
0.2
0.4
0.6
0.8
1.0
Probability of outcome

Figure 2

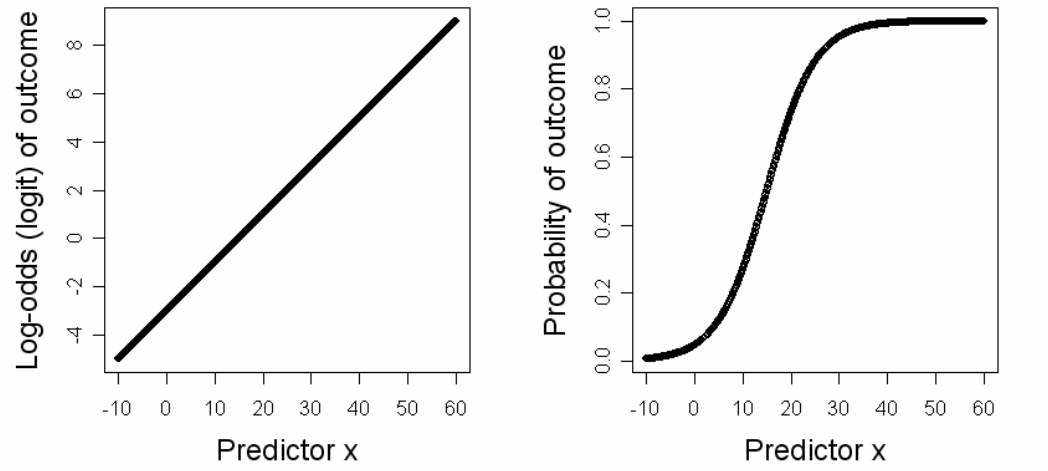

Log-odds (logit) of outcome
-4
-2
0
2
4
6
8
-10
0
10
20
30
40
50
60
Predictor x
Probability of outcome
0.0
0.2
0.4
0.6
0.8
1.0
-10
0
10
20
30
40
50
60
Predictor x

Figure 3

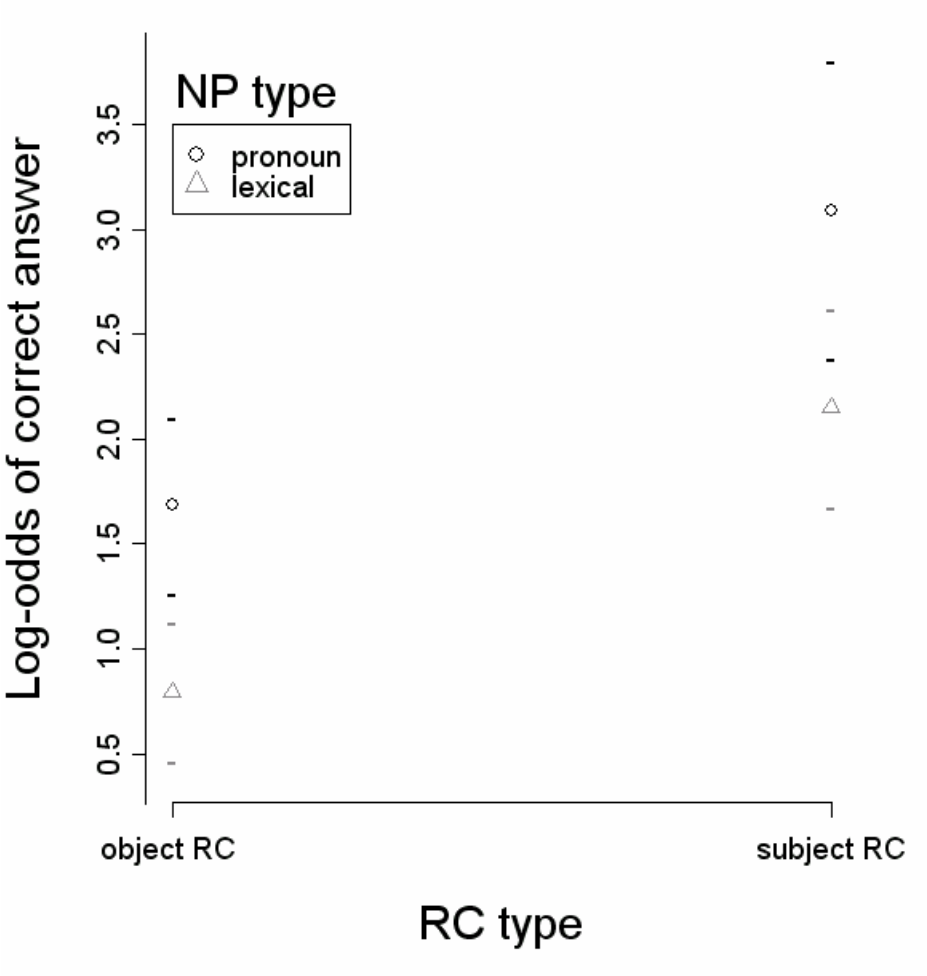

NP type
pronoun
lexical
Log-odds of correct answer
0.5
1.0
1.5
2.0
2.5
3.0
3.5
object RC
subject RC
RC type

Figure 4

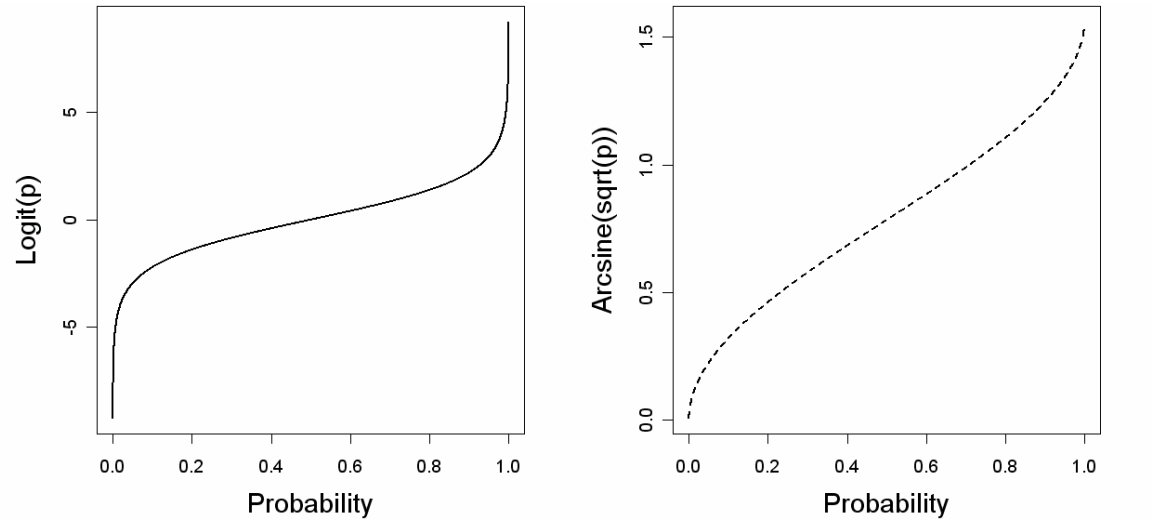
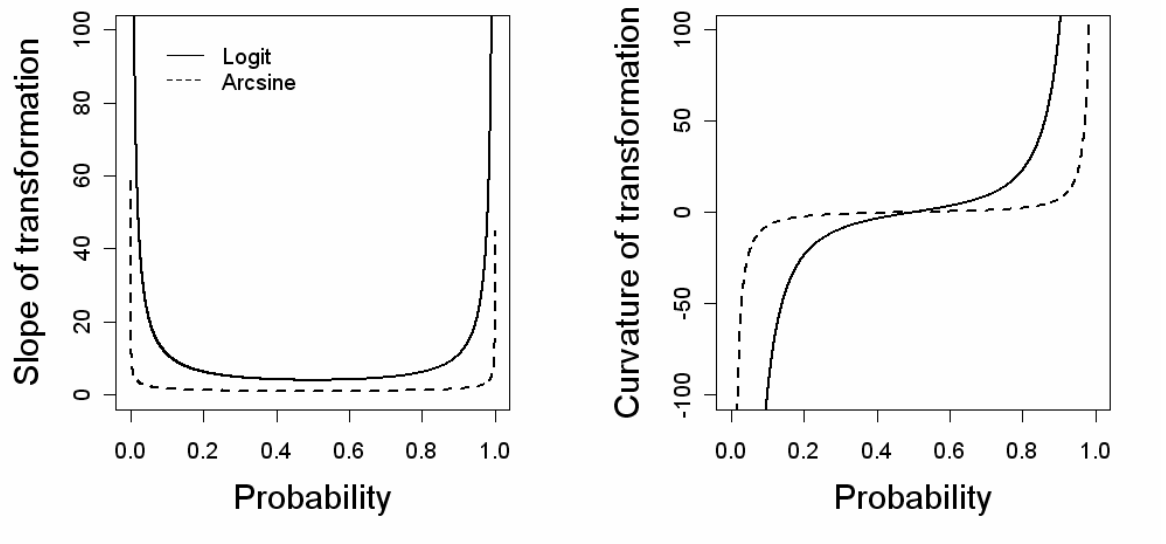

Logit(p)
5
0
-5
0.0
0.2
0.4
0.6
0.8
1.0
Probability
Arcsine(sqrt(p))
1.5
1.0
0.5
0.0
0.0
0.2
0.4
0.6
0.8
1.0
Probability

Figure 5


Slope of transformation
0
20
40
60
80
100
Logit
Arcsine
0.0
0.2
0.4
0.6
0.8
1.0
Probability
Curvature of transformation
-100
-50
0
50
100
0.0
0.2
0.4
0.6
0.8
1.0
Probability