British Journal of Mathematical and Statistical Psychology (2009), 62, 439–456

The British Psychological Society
www.bpsjournals.co.uk

Expert Tutorial

Multilevel modelling: Beyond the basic applications

Daniel B. Wright[1]* and Kamala London[2]

[1]Department of Psychology, Florida International University, Miami, Florida, USA
[2]University of Toledo, Toledo, Ohio, USA

Over the last 30 years statistical algorithms have been developed to analyse datasets that have a hierarchical/multilevel structure. Particularly within developmental and educational psychology these techniques have become common where the sample has an obvious hierarchical structure, like pupils nested within a classroom. We describe two areas beyond the basic applications of multilevel modelling that are important to psychology: modelling the covariance structure in longitudinal designs and using generalized linear multilevel modelling as an alternative to methods from signal detection theory (SDT). Detailed code for all analyses is described using packages for the freeware R.

1. Introduction

There has been a large increase in the use of multilevel models, in some form and by different names, within the social and medical sciences over the past decade. The number of computer programs that have specialized multilevel modules has also increased from a few specialist programs in the early 1990s, like ML2 (a precursor to MLwiN) and HLM, to being included in general packages like SAS, SYSTAT, and SPSS.

Different authors use different notations. We will use two notations in this paper. First, we follow Goldstein's (2003) notation, which we feel is the simplest for presenting multilevel models when describing the statistics. The second notation is that used in R, which we use to describe the computations for the examples. Using the first notation, the standard linear multilevel model with a single predictor variable is:

$$y_{ij} = \beta 0 + \beta 1 x_{ij} + u_j + e_{ij}$$

where j is the subscript for the different groups and i is the subscript for the different individuals. For this model, the intercept for group j is estimated by $\beta 0 + u_j$, an

* *Correspondence should be addressed to Dr Daniel B. Wright, Department of Psychology, Florida International University, Miami, FL 33199, USA (e-mail: danw@fiu.edu).*

DOI:10.1348/000711008X327632

estimate of the population intercept (β0) and variation around it for the individual groups (u_j). The e_{ij} are the individual level residuals. In this equation all the groups are assumed to have the same slope β1. It is usually assumed that the u_j and the e_{ij} are normally distributed around 0 with unknown standard deviations, σ_u and σ_e. The equal slope assumption can be relaxed by letting the β1 vary such that $\beta 1_j = \beta 1 + u1_j$, where $u1_j$ is a random variable usually also assumed to be normally distributed. Similarly heteroscedasticity as a function of x_{ij} in the level-1 residuals can be modelled by including additional level 1 random variation that is function of x_{ij} (for example, $x_{ij} \cdot e1_{ij}$). Different assumptions can be made and tested for correlations among the variables. More predictor variables can be added for multiple multilevel regressions, and if x_{ij} is replaced with a series of dummy variables, then multilevel ANOVAs can be analysed. These, in a sense, are the standard multilevel models and there are several introductions to these for psychologists (e.g. Hoffman & Rovine, 2007; Wright, 1998).

The purpose of this paper is to go beyond the basic multi-level model and to describe some of the techniques that can be particularly beneficial for psychologists. This paper is written for psychologists with good quantitative skills and not for statisticians. Where there is some statistical debate about, for example, different algorithms, we direct readers to where they can read more about the debate. More detailed reviews, in ascending order of the statistical knowledge they assume, are: Kreft and de Leeuw (1998); Hox (2002); Singer and Willett (2009); Bryk and Raudenbush (2002); Pinhiero and Bates (2000); and Goldstein (2003). We take the view that most quantitative psychologists are more interested in how to run a particular model, than specifics about the algorithms. Because of this we focus more on computer implementation than mathematics.

Each multi-level package has a different interface and different capabilities; therefore the choice of which to use is important. To make this paper useable by the largest audience we chose to present these models using the freeware R (R Development Core Team, 2008), which can be downloaded from the web (see Appendix). While general packages like SAS, SPSS, S-Plus, and SYSTAT are widely available to most academics, they are expensive without a site licence. Other packages like aML, MIXOR (which is now superseded by the commercial SuperMix), and WinBUGS are free, but are more specialized than R so are less useful for other analyses. MLwiN (Rasbash, Steele, Browne, & Prosser, 2005) was another contender both because it is in our opinion the state-of-the-art multi-level package and also that it is currently free for UK users (http://www.cmm.bristol.ac.uk/MLwiN/ordering/ac-uk.shtml). However, given the international readership of this journal we opted for R. Crawley (2005, 2007) has written an introduction to R that is suitable for the same audience as this paper and a detailed manual. For more advanced coverage see Chambers (2008) and for an introductory book see Venables, Smith, and the R Development Core Team (2008). In ascending order of statistical knowledge assumed, Wright and London (2009), Faraway (2004, 2006), and Fox (2002) provide tutorials on running regressions using R. There are different packages for R that estimate multi-level models. We will use `nlme` (Pinheiro & Bates, 2000; Pinheiro, Bates, DebRoy, & Sarkar, 2008) and `lme4` (Bates, 2007). These two packages have similar syntax and can do similar things. `nlme` has in-built correlation structures which makes it better suited for our first example and `lme4` allows generalized linear models (GLM) which makes it better suited for our second example.

We use `courier` for R functions, objects, and output in the text. We assume that readers have R (2.6 or higher). To download R go to http://cran.r-project.org/ and following the instructions. `nlme` and `lme4` can be installed and loaded from within R by:
`install.packages(c("lme4","nlme"))`, `library(lme4)`, and `library(nlme)`.
Some knowledge of R is assumed (Venables *et al.*, 2008). All of the code for running the analyses is on this paper's website (http://www.sussex.ac.uk/Users/danw/MLM.htm). The example data are part of the `sdtalt` package (Wright, Horry, & Skagerberg, in press) which can be installed and loaded in the same way.

This is a pedagogical piece. This is why R code and output are included in the text. No new statistical procedures are described. We look at two situations, common in psychology, and show how extensions of the standard multi-level model presented above can be used to analyse the data. These are longitudinal designs and using multi-level logistic regression as an alternative to methods from SDT.

1. Example #1 – Longitudinal designs: Child birth

To download data:

```
ayers <- read.table("http://www.sussex.ac.uk//Users//danw//MLM//
ayers.dat")
```

or

```
install.packages("sdtalt")
library(sdtalt)
```

then

```
attach(ayers)
```

Ayers' (1999) longitudinal study of women's mental health during and after pregnancy is used to illustrate examining possible variance–covariance structures. There were four time points (during pregnancy, 1 week, 6 weeks, and 6 months after birth) and 287 women in total. We will look at their anxiety scores, which range from 0 to 21, at these four time points: `anx1`; `anx2`; `anx3`; and `anx4`. Here are the values for the first few cases (the `dep` (depression) variables are not considered here):

```
ayers[1:3,]
    partno   dep1   dep2   dep3   dep4   anx1   anx2   anx3   anx4
1       67      0      0      0      0      3      2      1      2
2       22      0      0     NA     NA      2      0     NA     NA
3       40      0      0      0      0      6      5      6      2
```

Some of the values are missing and are labelled `NA` in the data file (e.g. participant #22 for the 3rd and 4th sessions). There are a variety of different methods for dealing with missing values (Little & Rubin, 2002), often estimating some value for each missing value. These can be computationally cumbersome with traditional within-subject

approaches to data analysis. Because of this, in the past researchers would go to great lengths to include all participants at each testing phase, make sure that the testing phases were all at the same time, and often would throw out participants with incomplete data. The multi-level approach solves some of these computational difficulties because the individual measurements are treated as randomly sampled elements nested within the individual and a variable for time can be used in the models. Conceptual difficulties still remain with the multi-level approach if the missing values are not missing at random, but the computational means for including incomplete cases are addressed.

The data were positively skewed for each anxiety variable (0.66, 0.90, 1.10, and 0.91, for `anx1` to `anx4`). The models in this example assume that the residuals are normally distributed. While having normally distributed response variables does not imply that the residuals will be normally distributed (nor vice versa), it was the case here and therefore we transformed the variables. Several transformations were tried and the square root of the variable plus .5 was used. Let $\text{anx1} = \sqrt{\text{anx1} + .5}$ or in R:

```
anx1 <- sqrt(anx1 + .5); anx2 <- sqrt(anx2 + .5)
anx3 <- sqrt(anx3 + .5); anx4 <- sqrt(anx4 + .5)
```

The new skewness values are: −0.12, 0.17, 0.23, and 0.10.

The numbers of women completing each wave were 251, 244, 219, and 201, and the means at these times were: 2.59; 2.19; 2.09; and 2.14, respectively. Complete data are available for only 177 of the 287 women. As mentioned above, with the traditional approach missing values must be dealt with. The two simplest approaches for dealing with missing data are: exclude incomplete cases and calculate measures for all pairs with data. The next step in analysing these data should be to look at the relationships between pairs of these variables, both graphically and numerically. Figure 1 shows the scatterplots, the histograms, and in the upper triangle the correlations and their 95% confidence intervals. The code for Figure 1 follows the example for `pairs` on the R help facility. From Figure 1, it is clear that the anxiety scores are all positively related.

Table 1 shows the correlations between data at the four time points in the lower triangle, the covariances in the upper triangle, and the variances along the diagonal. The left side of the table shows the values only for participants with complete data and the right side for all pairwise complete. The relationship between covariances, correlations, and variances is simple and is important for understanding R output. The covariance between variables x and y is: $\text{cov}_{xy} = \text{sd}_x\text{sd}_y\text{cor}_{xy}$. In R the correlation and covariance tables can be combined into a single table, like Table 1, with:

```
anxvars <- cbind(anx1,anx2,anx3,anx4)
coranx <- cor(anxvars,use = "complete.obs")
covanx <- cov(anxvars,use = "complete.obs")
cmat1 <- upper.tri(covanx,diag = T)*covanx + lower.tri
   (coranx)*coranx
coranx2 <- cor(anxvars,use = "pairwise.complete.obs")
covanx2 <- cov(anxvars,use = "pairwise.complete.obs")
cmat2 <- upper.tri(covanx2,diag = T)*covanx2 + lower.tri
 (coranx2)*coranx2
print(cbind(cmat1,cmat2),digits = 2)
```

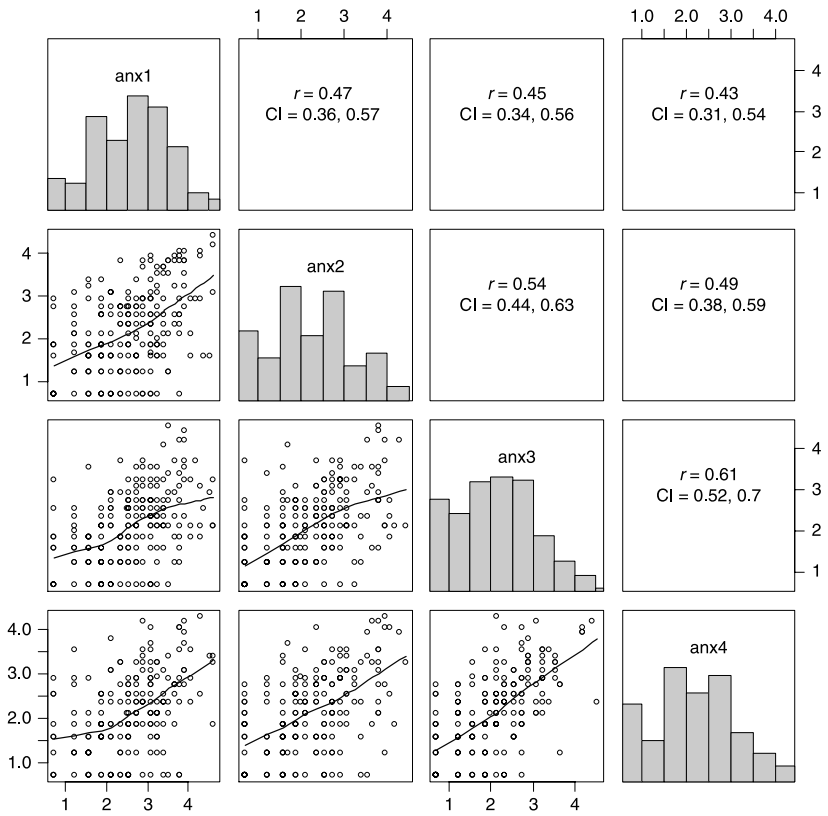

Figure 1. Scatterplots (lower triangle), histograms (diagonal), and correlations (upper triangle) for Ayers' (1999) data. All pairwise complete data are included.

There are a few things to note from Table 1. First, the correlations are all substantial and positive. Second, the correlations tend to be slightly higher when the variables are nearer in time (for the pairwise complete data: .47; .54; .61, for one step away compared with .45; .49, for two steps away; and .43 for three steps away). In many correlation matrices these differences are more pronounced, but it is still worth examining if this pattern should be taken into account here.

Table 1. The correlations (lower triangle), variances (diagonal), and covariance (upper triangle) for the four waves of data for anxiety from Ayers (1999)

	Excludes incomplete				All pairwise data included			
	anx1	anx2	anx3	anx4	anx1	anx2	anx3	anx4
anx1	.87	.41	.35	.37	.87	.42	.38	.36
anx2	.46	.91	.48	.41	.47	.94	.49	.43
anx3	.43	.57	.76	.47	.45	.54	.87	.50
anx4	.44	.48	.60	.83	.43	.49	.61	.86

Although this is not a multi-level dataset *per se*, researchers now often conceptualize repeated measures data as multi-level, and use multi-level algorithms to circumvent some of the problems of the standard repeated measures ANOVA (Singer & Willett, 2003). The first step is restructuring the dataset so that there is a single response variable, `anx`, a session variable called `session`, and a participant number (called `partno2`).

```
anx <- c(anx1,anx2,anx3,anx4)
session <- rep(1:4,each = length(anx1))
partno2 <- rep(partno,4)
```

The missing values were also removed. In R this is done by:

```
detach(ayers)
ayers <- na.omit(data.frame(partno2,session,anx))
rm(partno2); rm(session); rm(anx)
attach(ayers)
```

Here are the first few lines for these data:

```
ayers[1:3,]
   partno2  session     anx
1    67        1       1.870829
2    22        1       1.581139
3    40        1       2.549510
```

There are other ways to restructure the data. In some packages when creating a new multi-level structure the original variables would no longer be active. Here they are, so `anx1` still exists. This is why `partno2` was used rather than `partno`, and why the remove function (`rm`) was used to clean-up the working environment (other variables could also be removed). Most R statistics functions allow for the inclusion or exclusion of missing values, but here all `NA`s were removed with the `na.omit` function. The new data file `ayers` has 915 lines for the 915 anxiety measurements (i.e. 251 + 244 + 219 + 201).

The main substantive question in Ayers (1999) was about differences in anxiety at the different points in time and how these relate to other factors. Here we concentrate on finding a good correlation structure. Sometimes researchers are specifically interested in the correlations among variables, but an obvious question is whether choosing a good correlation structure makes a difference for estimating the fixed effects. The short answer is that it usually does not greatly affect the estimates of the fixed effects, but it often affects the precision of these estimates. Therefore, it is important to explore if confidence intervals or p values are to be reported.

`gls` – generalized least squares from `nlme` library (Pinheiro & Bates, 2000)

```
gls(respvar ~ covariates, weight = variance function,
  correlation = correlation structure)
```

Example variance and correlation functions (also available for `lme`):

```
varIdent(form = ~ 1 | repeated measure)
corAR1(form = ~ 1 | level)
```

It is necessary to examine both the variances of the individual measures and the correlations among the measures. We will use the `gls` function from `nlme` (Pinheiro & Bates, 2000; Pinheiro *et al.*, 2008), and when using this function these two facets (variances and correlations) are considered separately using the `weight` and the `correlation` options. The `nlme` package has several built-in variance and correlation structures. For the variance terms, they can all be equal (assume homogeneity), all be different with no particular relationships among them (unstructured heterogeneity), be a function of the time series (e.g. variances closer in time more similar), or be a function of other variables (e.g. variances a function of predicted values).

Similarly, the variables can be assumed to be uncorrelated as in a between-subjects design, have equal correlations, all have different correlations without any particular pattern, or a variety of correlation structures. Most of the structures in which psychologists would be interested come with `nlme`, but the user can construct their own (Pinheiro *et al.*, 2008). The most common pattern for longitudinal studies is first order autoregressive (AR1) where the correlations steadily decrease with distance from the main diagonal. Table 2 shows the patterns of standard deviations and correlations for some of these models (assumed covariances can be calculated using $cov_{xy} = sd_x sd_y cor_{xy}$).

The default contrasts for an ordered variable (`session`) are polynomials (linear, quadratic, etc.) and these will be used for the fixed part of the model. The default estimation procedure for these models is restricted maximum likelihood (REML) and this will be used for this example. The main alternative is maximum likelihood (set `method = "ML"`). Here is the code for these 10 models. The model, `mij`, corresponds to the *i*th row and *j*th column of Table 2. The only difference between the first five and the last five models is the standard deviations are allowed to differ in the later models.

```
m11 <- gls(anx ~ as.ordered(session))
m12 <- gls(anx ~ as.ordered(session),
    correlation = corCompSymm(form = ~ 1 |partno2))
m13 <- gls(anx ~ as.ordered(session),
    correlation = corAR1(form = ~ 1 |partno2))
m14 <- gls(anx ~ as.ordered(session),
    correlation = corARMA(form = ~ 1 |partno2,p = 3,q = 0))
m15 <- gls(anx ~ as.ordered(session),
    correlation = corSymm(form = ~ 1 |partno2))
m21 <- gls(anx ~ as.ordered(session),
    weights = varIdent(form = ~ 1| session))
m22 <- gls(anx ~ as.ordered(session),
    weights = varIdent(form = ~ 1| session),
    correlation = corCompSymm(form = ~ 1 |partno2))
m23 <- gls(anx ~ as.ordered(session),
    weights = varIdent(form = ~ 1| session),
    correlation = corAR1(form = ~ 1 |partno2))
m24 <- gls(anx ~ as.ordered(session),
    weights = varIdent(form = ~ 1| session),
    correlation = corARMA(form = ~ 1 |partno2,p = 3,q = 0))
m25 <- gls(anx ~ as.ordered(session),
    weights = varIdent(form = ~ 1| session),
    correlation = corSymm(form = ~ 1 |partno2))
```

Table 2. Some possible variance and correlation structures for Ayers (1999) for the anxiety measures across the four time points

	Uncorrelated (between-subjects)	Uncorrelated (within-subject)	Autoregressive 1 (AR1)	Autoregressive 3 AR3 (see also Toeplitz)	Unstructured
Homogeneous $\sigma_i = \sigma_j$	$\begin{bmatrix} \sigma & & & \\ 0 & \sigma & & \\ 0 & 0 & \sigma & \\ 0 & 0 & 0 & \sigma \end{bmatrix}$	$\begin{bmatrix} \sigma & & & \\ \rho & \sigma & & \\ \rho & \rho & \sigma & \\ \rho & \rho & \rho & \sigma \end{bmatrix}$	$\begin{bmatrix} \sigma & & & \\ \rho & \sigma & & \\ \rho\rho & \rho & \sigma & \\ \rho\rho\rho & \rho\rho & \rho & \sigma \end{bmatrix}$	$\begin{bmatrix} \sigma & & & \\ \rho 1 & \sigma & & \\ \rho 2 & \rho 1 & \sigma & \\ \rho 3 & \rho 2 & \rho 1 & \sigma \end{bmatrix}$	$\begin{bmatrix} \sigma & & & \\ \rho 1 & \sigma & & \\ \rho 2 & \rho 3 & \sigma & \\ \rho 4 & \rho 5 & \rho 6 & \sigma \end{bmatrix}$
df (fixed + random)	$df = 4 + 1$	$df = 4 + 2$	$df = 4 + 2$	$df = 4 + 4$	$df = 4 + 7$
AIC, BIC	2,504.86, 2,528.937	2,266.17, *2,295.06*	2,294.73, 2,323.62	*2,265.84*, 2,304.35	2,265.86, 2,318.82
−2 LL	1,247.43	1,127.09	1,141.37	1,124.92	1,121.93
Heterogeneous $\sigma_i \neq \sigma_j$	$\begin{bmatrix} \sigma 1 & & & \\ 0 & \sigma 2 & & \\ 0 & 0 & \sigma 3 & \\ 0 & 0 & 0 & \sigma 4 \end{bmatrix}$	$\begin{bmatrix} \sigma 1 & & & \\ \rho & \sigma 2 & & \\ \rho & \rho & \sigma 3 & \\ \rho & \rho & \rho & \sigma 4 \end{bmatrix}$	$\begin{bmatrix} \sigma 1 & & & \\ \rho & \sigma 2 & & \\ \rho\rho & \rho & \sigma 3 & \\ \rho\rho\rho & \rho\rho & \rho & \sigma 4 \end{bmatrix}$	$\begin{bmatrix} \sigma 1 & & & \\ \rho 1 & \sigma 2 & & \\ \rho 2 & \rho 1 & \sigma 3 & \\ \rho 3 & \rho 2 & \rho 1 & \sigma 4 \end{bmatrix}$	$\begin{bmatrix} \sigma 1 & & & \\ \rho 1 & \sigma 2 & & \\ \rho 2 & \rho 3 & \sigma 3 & \\ \rho 4 & \rho 5 & \rho 6 & \sigma 4 \end{bmatrix}$
df (fixed + random)	$df = 4 + 4$	$df = 4 + 5$	$df = 4 + 5$	$df = 4 + 7$	$df = 4 + 10$
AIC, BIC	2,510.23, 2,548.75	2,271.06, 2,314.40	2,298.41, 2,341.75	2,270.27, 2,323.23	2,271.21, 2,338.62
−2 LL	1,247.12	1,126.53	1,140.21	1,124.14	1,121.61

Note. The lowest AIC and BIC are in italic and underlined.
Other models were tested and are discussed on the web page. Calculating the values for ρ and σ for these matrices is somewhat cumbersome. Methods are shown on the web page for doing this.

These models can be compared with the `anova` function:

```
anova(m11,m12,m13,m14,m15,m21,m22,m23,m24,m25)
     Model df AIC      BIC      logLik     Test   L.Ratio   p-value
m11  1     5  2504.864 2528.937  −1247.432
m12  2     6  2266.170 2295.057  −1127.085 1 vs 2  240.69381  <.0001
m13  3     6  2294.733 2323.621  −1141.367
m14  4     8  2265.838 2304.354  −1124.919 3 vs 4  32.89553   <.0001
m15  5     11 2265.855 2318.815  −1121.928 4 vs 5  5.98259     .1125
m21  6     8  2510.231 2548.747  −1247.116 5 vs 6  250.37579  <.0001
m22  7     9  2271.064 2314.395  −1126.532 6 vs 7  241.16657  <.0001
m23  8     9  2298.414 2341.745  −1140.207
m24  9     11 2270.274 2323.234  −1124.137 8 vs 9  32.14057   <.0001
m25  10    14 2271.213 2338.616  −1121.606 9 vs 10 5.06100     .1674
```

This paper's web page includes details of the other models and how to extract the observed variance/covariance matrices from the output (see also http://www.ats.ucla.edu/stat/r/examples/alda/ch7.htm). The Akaike's information criterion (AIC) and Bayesian information criterion (BIC) are often used to choose among models (smaller numbers mean better fits); see Burnham and Anderson (2004) for details. All the models that assume equal standard deviations are better than their heterogeneous counterparts on these measures. Among the homogeneous models the one with all correlations equal (`m12`) and AR3 (`m14`) appear the two best. BIC penalizes complex models more than AIC, so it prefers the simpler model, and AIC prefers the more complex one. There is little to choose between these. Statistically, the more complex model is not significantly better, $\chi^2(2) = 4.33$, $p = .11$, [1] which suggests `m12` is better, but any theory of anxiety would predict correlations closer in time to be higher than those more distance. This suggests `m14` should be preferred. The choice between these will depend on the individual user's needs. Because *a priori* it would be expected that the different diagonals should have different correlations, we would opt for the more complex model.

2. Example #2 – Response times and accuracy in memory recognition

Since the 1970s (Banks, 1970; Lockhart & Murdock, 1970), the norm is to use statistics from SDT (see Macmillan & Creelman, 2005; Stanislaw & Todorov, 1999), to analyse memory recognition data. While there are exceptions, most of this research has used the equal variance Gaussian/normal model where the two most used statistics are d' as a measure of memory accuracy and C as a measure of bias. Within R the following function calculates d' and C for any set of hits, false alarms, misses, and correct rejections, or for vectors of these quantities if there are multiple participants, in which case it returns d' and C values for each participant.[2]

```
sdt1 <- function (hits,fas,misses,cr)
```

[1] *The χ^2 value is two times the difference in their log-likelihoods: 2(1,127.085 − 1,124.919).*
[2] *The R package `sdtalt` calculates these and 13 other common statistics (e.g. A′, odds ratio, eta, and weighted kappa) from signal detection theory, along with the confidence intervals for the sample means of these statistics (Wright, Horry, & Skagerberg, in press).*

```
  {d <- qnorm(hits/(hits + misses)) - qnorm(fas/(fas + cr))
  csdt <- -.5*(qnorm(hits/(hits + misses))
            + qnorm(fas/(fas + cr)))
  return(cbind(d,csdt))}
```

DeCarlo (1998) described how this is equivalent to a GLM, the probit regression, for each participant (see Faraway, 2006; Wright & London, 2009, for introduction to GLMs in R). Thus, the following function produces the same result as `sdt1` for a single set of values using the `glm` function. The first two lines inside the function reorganized the data into two variables.

```
sdt2 <- function(hits,fas,misses,cr)
  {sayold <- c(rep(1,sum(hits,fas)),rep(0,sum(misses,cr)))
  old <- c(rep(.5,hits),rep(-.5,fas),rep(.5,misses),rep(-.5,cr))
  model <- glm(sayold ~ old,family = binomial(link = "probit"))
  d <- model$coef[2]; csdt <- -model$coef[1]
  return(c(d,csdt))}
```

Both `sdt1(65,6,34,75)` and `sdt2(65,6,34,75)` produce $d' = 1.85$ and $C = 0.52$. Pictorially, Figure 2 shows these values.

These two methods are equivalent for calculating d' and C, so if these statistics are all that is needed then either can be used. Using `sdt2` allows access to other information, so if `summary(model)` was embedded in `sdt2` then it would show that, for example, the standard error of d' for these data is 0.47. The flexibility of using the `glm` function means that other methods can also be used. In medicine, the logistic model is more common that than the Gaussian/normal model (Zho, Obuchowski, & McClish, 2002). This can be found by changing `sdt2` so that it uses the default `family = binomial` link function, the logit. The coefficient of interest for the logistic model is the log of the odds ratio, ln(OR). Conceptually, ln(OR) can be thought of in a very similar way to d'. While $d' = z(\text{hit rate}) - z(\text{false alarm rate})$, $\ln(\text{OR}) = \ln(\text{odds of a hit}) - \ln(\text{odds of a false alarm})$. An approximate relationship between them is $\ln(\text{OR}) \approx 1.6d'$, and this holds except when d' is either very large or very small (i.e. $|d'| > 4$). For the example above, $\ln(\text{OR}) = 3.17$ which is about 1.6 times the observed d'. We will use the logistic model because that is the *natural* link function for the binomial distribution (Hoffmann, 2004).

Another way in which the GLM approach is flexible is that predictor variables which vary trial-by-trial, like confidence or response time, can be included within the model.

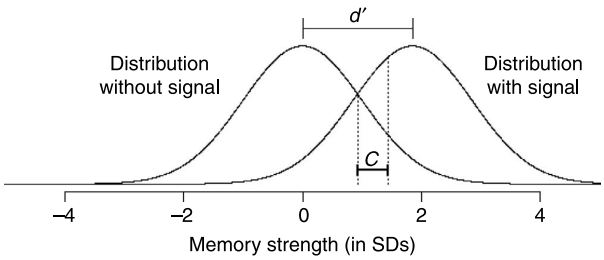

Figure 2. A pictorial representation of $d' = 1.85$ and $C = 0.52$ for the equal variance normal signal detection theory model.

This is difficult to do with the standard approach to SDT. The norm within most memory recognition research is to calculate d' (or other measures) for each participant, sometimes separately for different within-subject conditions, and compare these aggregate measures. One of the main arguments for multi-level modelling is to move away from analysing aggregate level data like these.

`lmer` – (generalized) linear mixed effects regression from `lme4` library (Bates, 2007)

```
lmer(respvar ~ covariates + (random variables | levels),
    family = binomial(link = "logit"))
```

The default `family` is Gaussian with identity link function. Use `family = binomial` for logistic regression.

In this example the function `lmer` is used rather than `lme` or `gls` because `lmer` allows generalized linear multi-level modelling. `lmer` was not used in the last example because it does not have built-in covariance structures. For `lmer` the random effects are placed within the model. Thus, `m12` from the first example could be written as:

```
lmer(anx ~ as.ordered(session) + (1|partno2))
```

The output is slightly different because different estimations procedures are used, but the basic models are equivalent.

With `lmer` for GLMs, maximum-likelihood estimates of the parameters are found with an iterative procedure. The `lme4` manual lists three approximation methods (Bates, 2007). In order of their level of accuracy and their computation time, these are: penalized quasi-likelihood (`PQL`); Laplacian approximation (`Laplace`); and adaptive Gaussian quadrature approximation (`AGQ`). Bates recommends using `Laplace`, so it is used here. Details of `lme4` are in Bates (2007, http://cran.r-project.org/web/packages/lme4/vignettes/Implementation.pdf).

We begin with a simple example where traditional SDT could also be used: showing that white people have better recognition memory for white faces than for black faces. We then add in a continuous predictor variable that varies by trial, the log of the response time. The data are from the white English participants of Wright, Boyd, and Tredoux (2003).

To download data:

```
memrec <- read.table(
"http://www.sussex.ac.uk//Users//danw//MLM//memrec.dat",header = T)
```

or

```
library(sdtalt)
```

then

```
attach(memrec)
```

The data are in `memrec`. The variables are: `face; saysold; faceold; facewhite; lntime;` and `partno`. Figure 3 shows the frequency of old responses in the different conditions. The proportion of hits is about the same for white and black faces, but the number of false alarms is much greater for black faces. This is consistent with the memory literature (Horry & Wright, 2008).

The model to test for what is called the *own race bias* compares `model1` and `model2` below. `model1` includes the main effects for whether the face was previous shown (`faceold`, which measures accuracy), how many of these were white faces (`facewhite`), and it allows the intercept to vary for people. In SDT terminology this corresponds to people having different response criteria. In statistical notation it is:

$$\text{logit}(p[\text{saysold}_{ij}]) = \beta 0 + \beta 1\,\text{facewhite}_{ij} + \beta 2\,\text{faceold}_{ij} + u_j + e_{ij}$$

where the e_{ij} are assumed to be binomially distributed and the u_j are assumed to be normally distributed. In R this is:

```
model1 < - lmer(saysold ~ facewhite + faceold + (1|partno),
family = binomial, method = "Laplace")
```

Writing `summary(model1)` produces statistics for the overall fit of this model and the individual coefficient estimates: $\beta 0 = -0.99$; $\beta 1 = -0.51$; $\beta 2 = 2.02$; and $\text{var}(u_j) = 0.12$. We can conclude that people say 'old' more to black faces and to items that are old. Both of these effects are evident from Figure 3.

`model2` adds the interaction between the two predictor variables so it tests whether accuracy depends on whether the face was white or black. The `update` function is convenient for incrementally building models.

```
model2 < - update(model1,. ~ . + facewhite:faceold)
anova(model1,model2)
        Df  AIC     BIC     logLik    Chisq   Chi df  Pr(>Chisq)
model1  4   3455.4  3479.5   - 1723.7
model2  5   3422.2  3452.3   - 1706.1  35.225  1       2.937e - 09***
```

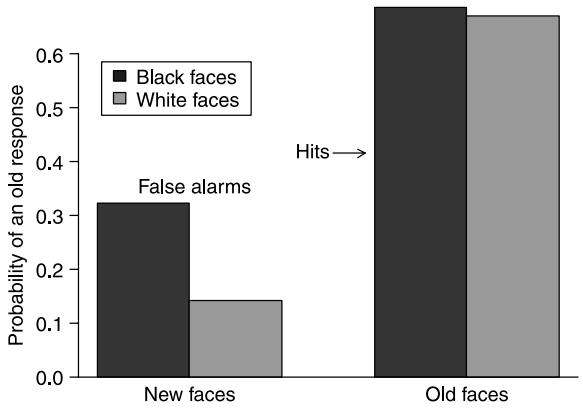

Figure 3. The probability of an 'old' response for the data from white English participants in Wright *et al.* (2003).

The own race bias is observed, $\chi^2(1) = 35.23$, $p < .001$. Participants were more accurate responding to white faces. The coefficient for whether something was previously shown (here `faceold`) measures discriminability (a measure of memory) and interactions between this and other variables show whether these other variables moderate accuracy. The fixed effects for `model2` can be found either by typing `model2` (which produces the fixed effects and a lot of other output) or by `fixef(model2)`.

```
fixef(model2)
     (Intercept)     facewhite       faceold       facewhite:faceold
     − 0.7633022     − 1.0868220     1.5656096     1.0168033
```

The estimated parameter for `faceold` is 1.57. This estimates $\ln(OR)_{black}$. The interaction was 1.02, so the estimate of $\ln(OR)_{white}$ is $1.57 + 1.02 = 2.59$. The probit model can also be used and the multi-level estimates for d'_{black} and d'_{white} by using the probit link.

```
model2a < - update(model2, family = binomial(link = probit))
fixef(model2a)
     (Intercept)     facewhite       faceold       facewhite:faceold
     − 0.4719156     − 0.6335219     0.9652778     0.5930885
```

The estimated value for d'_{black} is 0.97 and for d'_{white} is $0.97 + 0.59 = 1.56$. Notice that the $\ln(OR) \approx 1.6\ d'$ approximation holds.

If we had calculated ln(OR) and d' separately for each individual for each race, the means of these would be $\ln(OR)_{black} = 1.69$, $\ln(OR)_{white} = 5.47$, $d'_{black} = 1.02$, and $d'_{white} = 2.32$. Notice that these values are very different from those found with the multi-level model for white faces and that the $\ln(OR) \approx 1.6\ d'$ approximation does not hold for the white faces ($5.47/2.31 = 2.37$). This is because some individuals had extremely high values for these (many $d' > 4$) and that the mean is not a robust statistic. The 20% trim means are: $\ln(OR)_{black} = 1.74$; $\ln(OR)_{white} = 3.02$; $d'_{black} = 1.06$; and $d'_{white} = 1.77$. The approximation now holds and these values are closer to those found with the multi-level approach.

It is worth exploring if participants' accuracy varies. This can be done by changing the random part of the model to `(faceold|partno)`. This adds both a term for the variance of accuracy and the covariance between accuracy and responding old, and therefore there is an increase of two degrees of freedom in the model. Including this term increases the fit of the model, $\chi^2(2) = 12.59$, $p = .002$. It can be written with the following and compared with `model2`. The difference is statistically significant, $\chi^2(2) = 12.59$, $p = .002$, although the BIC value is increased.

```
model2b < - lmer(saysold ~ faceold*facewhite + (faceold|partno),
family = binomial,method = "Laplace")
anova(model2,model2b)
         df   AIC      BIC      logLik     Chisq    Chi df   Pr(>Chisq)
model2    5   3422.2   3452.3   − 1706.1
model2b   7   413.6    3455.7   − 1699.8   12.592     2      0.001844**
```

Up to this point traditional SDT could have been used to reach the same basic conclusion, providing care was taken to use robust estimators. The next step involves adding a continuous variable which varies by trials: the log of the response time

(`lntime`). The theoretical question is about the relationship between response time and accuracy for the different faces. Much research shows quicker responses tend to be more accurate, but the exact relationship is unclear (Weber, Brewer, Wells, Semmler, & Keast, 2004). Our question is whether the relationship is similar for old and new faces, and white and black faces.

```
model3 <- update(model2b,. ~ . + lntime)
anova(model2b,model3)
         df   AIC      BIC      logLik     Chisq    Chi df   Pr(>Chisq)
model2b   7   3413.6   3455.7   -1699.8
model3    8   3415.4   3463.4   -1699.7   0.2781     1     0.5979
```

The main effect of time was non-significant, $\chi^2(1) = 0.28$, $p = .60$, but is retained as interactions including this term are added to the model. The best model in terms of AIC, BIC, and significance tests, includes only adding the interaction between `lntime` and `faceold`: improvement $\chi^2(1) = 52.43$, $p < .001$.

```
model4 <- update(model3,. ~ . + lntime:faceold)
anova(model3,model4)
        df   AIC      BIC      logLik     Chisq    Chi df   Pr(>Chisq)
model3   8   3415.4   3463.4   -1699.7
model4   9   3364.9   3419.0   -1673.5   52.429     1     4.461e-13***
```

The coefficients for this model are:

```
fixef(model4)
         (Intercept)         faceold                  facewhite
         -7.5124112          13.3062753               -1.0654064
         lntime              faceold:facewhite        faceold:lntime
         0.8583524           0.9949194                -1.5010712
```

The coefficient associated with the `faceold:lntime` interaction is negative and therefore accuracy decreases with increased response time. The lack of other interactions improving the model shows that the relationship between time and accuracy is similar for white and black faces.

In Figure 4 the probability of a correct response by whether the face was new (where it is a probability of a correct rejection) or old (where it is a probability of a hit) and the race of the face is plotted with response time. We calculated the predicted probabilities using the estimates above and transformed the predicted values with $e^x/(1 + e^x)$, the inverse of the logit transformation. We then changed these predicted values to 1 minus themselves for new faces, so that the probabilities were for correct responses.

```
mod4 <- -7.51 + faceold*13.31 + facewhite* - 1.07 + lntime
  *0.86 + faceold*facewhite*.99 + faceold*lntime* - 1.50
predprob <- exp(mod4)/(1 + exp(mod4))
rightprob <- predprob*faceold + (1 - faceold)*(1 - predprob)
```

This shows that after controlling for response time the probability of correct response is highest for new white faces.

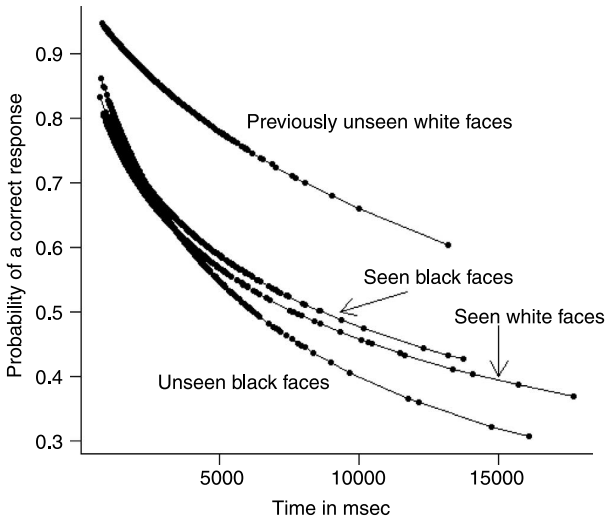

Figure 4. The probability of a correct response for previously seen and previously unseen white and black faces, with the raw response times, for the white English participants from Wright *et al.* (2003).

There are several other extensions to the multi-level GLMs that can be explored, including tests of the variance/covariance matrix at each level of the model, as discussed with the first example. It is also often worth examining the size of var(e_{ij}) to see if there is more or less variation than predicted from the binomial distribution (Browne, Subramanian, Jones, & Goldstein, 2005; Wright, 1997). This is reported with the `summary` information and for these data it is very similar to that predicted by the binomial distribution. Another extension is to model more flexible relationships using multi-level generalized additive models (see Wood, 2006, for details, and also Ng, Carpenter, Goldstein, & Rasbash, 2006). Wood (2006) has written a specialist package within R to run these, but it can also be done with the `bs` function from the `splines` library. `bs` is a B-spline which is a set of polynomials linked together smoothly at knots (see Wood, 2006, for more details). The following model allows the relationship between response time and responding old to be modelled in a more flexible way which increases the chances of detecting other effects. This is a multi-level gamcova (Wright & London, 2009).

```
model8 < - lmer(saysold ~ faceold*facewhite + bs(lntime,df = 4) +
  faceold + lntime:faceold + (faceold|partno), family = binomial,
  method = "Laplace")
anova(model4,model8)
        df   AIC      BIC      logLik     Chisq    Chi df   Pr(>Chisq)
model4  9    3364.9   3419.0    − 1673.5
model8  12   3359.2   3431.2    − 1667.6   11.751   3        0.008288**
```

Finally, the usual practice in memory recognition research is not to worry about differences among the stimuli within any category. Within linguistics there is a long tradition of taking into account the differences among stimuli (Clark, 1973). Baayen, Davidson, and Bates (2008) show how to include a random variable in their models using `lmer`. This creates what is called a cross-classified model. In this data set there is a variable, `face`, for the face number.

```
model9 < - lmer(saysold ~ faceold*facewhite + lntime*faceold + (1|face)
 + (faceold|partno), family = binomial, method = "Laplace")
anova(model4,model9)
          df  AIC     BIC     logLik     Chisq  Chi df   Pr(>Chisq)
model4    9   3364.9  3419.0  - 1673.5
model9    10  2896.6  2956.7  - 1438.3  470.27    1     < 2.2e - 16***
```

As can be seen, this greatly improves the model. This shows that there were more 'old' responses for some faces than for others.

2.1. Summary – Memory recognition

While traditional methods from SDT are often used for memory recognition studies, they present two difficulties. First, they are usually done by calculating measures (like d') for each individual and then using these aggregate measures in analysis. Outliers, particularly those based on few cases, can have a large impact. This is one of the main reasons for multi-level modelling. Second, if interested in a covariate which can take many possible values, like response time, the standard SDT methods are difficult to implement since the covariate has to be split into bins and SDT measures created for each person for each bin. This can create lots of problems particularly when there are few or unequal numbers of observations per person per bin because the estimates for the individual bins can be unstable.

The multi-level modelling approach is well suited to overcome these problems. Because the standard SDT approach is equivalent to a GLM (DeCarlo, 1998), we use GLMs (in particular logistic regression, although `link = "probit"` could be used throughout this section) for analyses, treating trials as nested within participants. Although multi-level GLMs have been used with memory recognition data for many years (e.g. Wright & McDaid, 1996), they are still not very common. While SDT has a long history within memory research, and has been very useful in theory construction, it is likely that the flexibility of the approach used here will mean multi-level models become the norm for recognition data in years to come. These have the methodological advantage that researchers can use more variables that differ by trial within their designs. In this example response time was included which given the number of studies now conducted via computer is a variable that could be included in most studies.

3. Summary

Multilevel modelling is one of the hot statistical methods in several areas of science, including psychology. While the traditional example has been with people nested within larger clusters (e.g. pupils nested within classrooms), because of the great amount of medical research with multiple measurements per person, multi-level models with the person as the higher order level are now common (perhaps more common). Harvey Goldstein, one of the pioneers of this approach, talks about how there are hierarchies everywhere. Multilevel modelling is now one of the tools expected for social and psychological scientists.

We end with a caveat. While multi-level models are now expected to be used in areas where the hierarchical structure is obvious, more research is necessary to see how useful they are when the levels are not such clean structures and where the components at different levels cannot be viewed as some random sample of those at that level. This was Cohen's (1976) main criticism of Clark's (1973) language as a fixed effect fallacy.

Perhaps some of the resampling techniques (and local causal inference) will be applied to these situations. As with all statistical procedures, it is critical to examine the data carefully and consider the alternatives before running any statistical test. No amount of statistical sophistication can fix a bad design.

References

Ayers, S. E. (1999). Post-traumatic stress disorder following childbirth. PhD dissertation. University of London.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*.

Banks, W. P. (1970). Signal detection theory and human memory. *Psychological Bulletin*, *74*, 81–99.

Bates, D. (2007). lme4: Linear mixed-effects models using S4 classes. R package version 0.99875-9.

Browne, W., Subramanian, S., Jones, K., & Goldstein, H. (2005). Variance partitioning in multilevel logistic models that exhibit over-dispersion. *Journal of the Royal Statistical Society: A*, *168*, 599–613.

Bryk, A. S., & Raudenbush, S. W. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods and Research*, *33*, 261–304.

Chambers, J. M. (2008). *Software for data analysis: Programming with R*. London: Springer.

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.

Cohen, J. (1976). Random means random. *Journal of Verbal Learning and Verbal Behavior*, *15*, 261–262.

Crawley, M. J. (2005). *Statistics: An introduction using R*. Chichester, UK: Wiley.

Crawley, M. J. (2007). *The R book*. Chicester, UK: Wiley.

DeCarlo, L. T. (1998). Signal detection theory and generalized linear models. *Psychological Methods*, *3*, 186–205.

Faraway, J. J. (2004). *Linear models with R*. Boca Raton, FL: Chapman and Hall/CRC.

Faraway, J. J. (2006). *Extending the linear model with R: Generalized linear, mixed effects and nonparameteric regression models*. Boca Raton, FL: Chapman and Hall/CRC.

Fox, J. (2002). *An R and S-Plus companion to applied regression*. Thousand Oaks, CA: Sage Publications.

Goldstein, H. (2003). *Multilevel statistical methods* (3rd ed.). London: Edward Arnold.

Hoffmann, J. P. (2004). *Generalized linear models: An applied approach*. Boston, MA: Pearson Education.

Hoffman, L., & Rovine, M. J. (2007). Multilevel models for the experimental psychologist: Foundations and illustrative examples. *Behavior Research Methods*, *39*, 101–117.

Horry, R., & Wright, D. B. (2008). I know your face but not where I saw you: Context memory is impaired for other race faces. *Psychonomic Bulletin and Review*, *15*(3), 604–609.

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. London: Erlbaum.

Kreft, I. I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. London: Sage Publications.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). New York: Wiley.

Lockhart, R. S., & Murdock, B. B. (1970). Memory and the theory of signal detection. *Psychological Bulletin*, *74*, 100–109.

MacMillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Mawhaw, NJ: Erlbaum.

Ng, E. S. W., Carpenter, J. R., Goldstein, H., & Rasbash, J. (2006). Estimation in generalised linear mixed models with binary outcomes by simulated maximum likelihood. *Statistical Modelling*, *6*, 23–42.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-Plus*. New York: Springer.
Pinheiro, J. C., Bates, D. M., DebRoy, S., & Sarkar, D. (2008). *nlme: Linear and nonlinear mixed effects models*. R package version 3.1-87.
R Development Core Team (2008). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, http://www.R-project.org.
Rasbash, J., Steele, F., Browne, W., & Prosser, B. (2005). *A user's guide to MLwiN*. http://www.cmm.bristol.ac.uk/MLwiN/download/userman_2005.pdf
Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.
Singer, J. D., & Willett, J. B. (2009). *Applied multilevel data analysis*. Manuscript in preparation.
Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior Research Methods, Instruments, and Computers*, *31*, 137–149.
Venables, W. N., Smith, D. M., & the R Development Core Team (2008). *An introduction to R*. ISBN 3-900051-12-7. http://cran.r-project.org/doc/manuals/R-intro.pdf
Weber, N., Brewer, N., Wells, G. L., Semmler, C., & Keast, A. (2004). Eyewitness identification accuracy and response latency: The unruly 10–12 second rule. *Journal of Experimental Psychology: Applied*, *10*, 139–147.
Wood, S. N. (2006). *Generalized additive models: An introduction with R*. Boca Raton, FL: Chapman and Hall/CRC.
Wright, D. B. (1997). Extra-binomial variation in multilevel logistic models with sparse structures. *British Journal of Mathematical and Statistical Psychology*, *50*, 21–29.
Wright, D. B. (1998). Modelling clustered data in autobiographical memory research: The multilevel approach. *Applied Cognitive Psychology*, *12*, 339–357.
Wright, D. B., Boyd, C. E., & Tredoux, C. G. (2003). Inter-racial contact and the own race bias for face recognition in South Africa and England. *Applied Cognitive Psychology*, *17*, 365–373.
Wright, D. B., Horry, R., & Skagerberg, E. M. (in press). Functions for traditional and multilevel approaches to signal detection theory. *Behavior Research Methods*.
Wright, D. B., & London, K. (2009). *Modern regression techniques using R: A practical guide for students and researchers*. London: Sage Publications.
Wright, D. B., & McDaid, A. T. (1996). Comparing system and estimator variables using data from real line-ups. *Applied Cognitive Psychology*, *10*, 75–84.
Zho, X.-H., Obuchowski, N. A., & McClish, D. K. (2002). *Statistical methods in diagnostic medicine*. New York: Wiley.

Received 28 March 2008; revised version received 28 May 2008

Appendix

R was used for these analyses. To download R go to http://cran.r-project.org/ and following instructions for Linux, MacOS X, or Windows.

The data and detailed code for R 2.7 are available on http://www.sussex.ac.uk/Users/danw/MLM.htm. The output is annotated. The code and pages will be updated as needed. You may need to access libraries including `nlme` and `lme4`. To install and to load these use the `install.packages` and `library` functions. For example:

```
install.packages("lme4")
library(lme4)
```

The data are both on the paper's website and part of the `sdtalt` package which can be installed and loaded.