Exploiting test structure: Case series, case-control comparison, and dissociation

Michael Smithson a, Martin Davies b & Anne M. Aimola Davies a b c

a Department of Psychology, The Australian National University, Canberra, Australia
b Department of Experimental Psychology and Faculty of Philosophy, University of Oxford, Oxford, UK
c NIHR (National Institute for Health Research) Biomedical Research Centre, University of Oxford, Oxford, UK

Available online: 24 Nov 2011
Exploiting test structure: Case series, case–control comparison, and dissociation

Michael Smithson¹, Martin Davies², and Anne M. Aimola Davies¹,²,³

¹Department of Psychology, The Australian National University, Canberra, Australia
²Department of Experimental Psychology and Faculty of Philosophy, University of Oxford, Oxford, UK
³NIHR (National Institute for Health Research) Biomedical Research Centre, University of Oxford, Oxford, UK

Most neuropsychological tests consist of multiple items, and a subject’s test score is the sum of the item scores. The test results for each subject thus comprise multiple data-points, and any data-set with test results from more than one subject has at least a two-level structure, with the test item as the first level and the subject as the second level. This structure may be exploited to yield more nuanced statistical analyses than those that treat each subject’s test score as a single data-point. Exploiting this structure allows us to take into account the effect of test length and dispersion on score variance and may enhance statistical power. Focusing on tests for which the score can be regarded as a binomial random variable, and using the binomial general linear model, we describe appropriate statistical methods for exploiting test structure in analysing a case series, comparing a case with a control sample, and testing for dissociation. These methods also allow multiple predictors, both categorical and continuous, to be taken into account, thereby enhancing the capacity of researchers to test hypotheses in a case series and to investigate other explanatory factors, in addition to case–control status.

Keywords: Case series; Test length; Overdispersion; Binomial general linear model; Dissociation.

Exploiting test structure

Most neuropsychological tests consist of multiple items, and so the test results for each subject comprise multiple data-points, not just the subject’s test score. Those data-points have a variance, and therefore the test score, considered as the sum (or the mean) of the item scores, has a standard error. Thus, any data-set with test results from more than one subject has at least a two-level structure, with the test item as the first level and the subject as the second level. (In some cases, subtests may constitute a third intermediate level.) This structure may be exploited to yield more nuanced statistical analyses than those that treat each subject’s test score as a single data-point. Ignoring this structure may result in needless loss of statistical power.

In this paper we focus on tests whose items are binary, simply scored “correct” or “incorrect”. This test format is widespread, its underlying distribution is well known, and the standard error of a subject’s test score can be calculated from just the score itself and the test length. On a binary-item
test, the subject’s score is the number of items correct \((r, \text{say})\) out of the total number of items \((n)\). For the simplest statistical techniques, each item may be regarded as a Bernoulli random variable, and the test score as a binomial random variable with probability \(\pi\) of getting an item correct.\(^1\)

We develop statistical methods that exploit the structure of binary-item tests, and we show how to use these methods in analysing a case series. For illustrative purposes, we draw on some data from a study by Maguire and Ogden (2002).\(^2\) Over an 18-month period, patients with unilateral neglect persisting more than three months after a stroke were identified at two inpatient neurorehabilitation hospitals in Auckland (serving a population of one million). Nine patients were assessed, of whom the seven with complete data-sets were included in the statistical analysis. Even with a relatively small series of cases, the methods that we describe provide sufficient statistical power to reveal differences in test scores among the patients and to account for these differences using predictors. In later sections of the paper, we extend the approach for use in comparing a case with a control sample and in testing for dissociation.

It is important to note that we can exploit the structure of a binary-item test even if we do not have access to the item-by-item raw data. In most of the applications discussed in this paper, we assume that only the individual subjects’ test scores are available.\(^3\)

**Motivating the approach**

We now provide three illustrations to motivate the consideration of statistical models that exploit the structure of binary-item tests rather than treating each subject’s test score as a single data-point.

**Illustration 1: Assessing heterogeneity in a case series**

First, consider a situation in which we have a series of cases, all with the same diagnosis. Suppose that these patients have been assessed using two neuropsychological tests and that we have their test scores (but not the item-by-item raw data). Also suppose that, in reality, patients with this diagnosis are heterogeneous on Test 1 but homogeneous on Test 2. That is, their scores on Test 2 come from a population where everyone has the same mean (lower than the normal mean) and random variability around that mean, whereas their scores on Test 1 come from a population where individuals have their own mean, and their test score varies randomly around that.

If we treat each patient’s score on each test as a single data-point, then there is no way to discern heterogeneity on Test 1 but not on Test 2. We can calculate a sample mean and standard deviation for each test, but we cannot tell whether either standard deviation is “too large” for a sample of scores from a homogeneous population. Statistical models that exploit the structure of binary-item tests allow researchers to detect and account for heterogeneity in scores from a sample (series) of cases—or from a sample of control subjects.

**Illustration 2: Taking advantage of test length in case–control comparison**

For our second illustration, motivating statistical models that exploit the structure of binary-item tests, consider an example of comparing a single case with a control sample. Table 1 displays the scores on two hypothetical tests for five controls and one case. Test 1 has length \(n = 100\), and Test 2 has length \(n = 10\).

If we treat the control test scores as single data-points, then we can calculate their mean and standard deviation for each test. The mean number of

---

\(^1\) This approach has sometimes been recommended in the single-case research methods literature (e.g., Willmes, 1990). A useful analogy for the approach is meta-analysis: Each subject’s test score corresponds to the summary statistic from a study, and the test length corresponds to the sample size of that study.

\(^2\) The first author of Maguire and Ogden (2002) is the last author of the present paper.

\(^3\) If the item-by-item raw data are available, then the framework presented here can be generalized to model dependencies and item difficulty levels via item and subject effects. A full exegesis of this generalized framework is beyond the scope of this paper but several excellent texts are available (e.g., de Leeuw & Meijer, 2007).
correct items for the controls is 50 for Test 1 and 5 for Test 2 (so the mean proportion for both tests is $50/100 = 5/10 = .5$), and the standard deviation of the number of correct items is 5 for Test 1 and 1 for Test 2. If we treat the case test scores (35 for Test 1 and 3 for Test 2) as single data-points, then the tests yield $z$ scores for the case of $-3.0$ on Test 1 [$z = (35 - 50)/5$], and $-2.0$ on Test 2 [$z = (3 - 5)/1$]. If the convention were that any score at least two standard deviations from the control mean is to be regarded as outside the normal range, the case would be regarded as impaired on both tests. The case’s performance would not be regarded as significantly lower than that of the controls on either test.

If we treat test scores as single data-points, then we reach the same conclusion about the two tests in this example. According to the $z$ scores, the case is impaired on both tests; according to the Crawford–Howell $t$-statistics, the case is significantly different from the controls on neither test. However, Test 1 is 10 times the length of Test 2. Longer tests are more reliable and therefore should give greater statistical power to this kind of comparison. Statistical models that exploit the structure of binary-item tests enable researchers to take advantage of greater test length.

Illustration 3: The problem of non-normally distributed test scores
Traditional and more recently developed procedures rely on the assumption that test scores are normally distributed. It is well known that severe departures from normality (e.g., skew or kurtosis) render such procedures problematic. (Skew is the asymmetry of a distribution about its mean. Kurtosis is the peakedness or flatness of a distribution compared to the normal distribution.) But more importantly, severe departures from normality in control sample test scores are commonplace. As Crawford and colleagues point out (Crawford, Garthwaite, & Gray, 2003, p. 367): “Skewed control data are not uncommon in single-case studies when the tasks that measure abilities are largely within the competence of most healthy individuals.” Indeed, it is not uncommon for controls to achieve the most extreme score possible on a test (e.g., getting all of the items correct), which can result in uncorrectable skew. Crawford and colleagues (Crawford, Garthwaite, Azzalini, Howell, & Laws, 2006) also note that leptokurtic (narrow-peaked) distributions frequently crop up in control sample data.

Crawford and colleagues (Crawford et al., 2006) also demonstrated that, for the method proposed by Crawford and Howell (1998), the effects of skew and leptokurtosis on Type I errors were fairly modest. Their recommendation was that a more conservative Type I error criterion could be

Table 1. Hypothetical scores for 5 controls and 1 case on two tests

<table>
<thead>
<tr>
<th>Test length</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Control</td>
<td>46</td>
<td>4</td>
</tr>
<tr>
<td>Control</td>
<td>53</td>
<td>6</td>
</tr>
<tr>
<td>Control</td>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>Control</td>
<td>57</td>
<td>4</td>
</tr>
<tr>
<td>Control</td>
<td>49</td>
<td>5</td>
</tr>
<tr>
<td>Mean</td>
<td>50</td>
<td>5</td>
</tr>
<tr>
<td>SD</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Case</td>
<td>35</td>
<td>3</td>
</tr>
</tbody>
</table>

The likelihood that an individual whose probability of a correct response on Test 1 is .5 should score 35 or less out of 100 (about .002) is much lower than the likelihood that an individual whose probability of a correct response on Test 2 is .5 should score 3 or less out of 10 (about .172). We hasten to add that the Crawford–Howell $t$ test also is responsive to test length because, given a constant level of dispersion, the standard deviation of control scores will scale down as test length increases. Nevertheless, these likelihoods are much more contrastive than the $t$ test significance levels of .052 versus .142.
Binomial general linear model (GLM)

We now introduce statistical techniques that address the issues raised in our first two illustrations (assessing heterogeneity in a case series or control sample and taking advantage of test length in case–control comparison). When discussing these techniques, we usually assume that we have access to the test scores for individual subjects but not to the item-by-item raw data. The techniques are based on the binomial general linear model (GLM), which allows multiple predictors, both categorical (e.g., case versus control status and gender) and continuous (e.g., age).

The binomial GLM (e.g., McCullagh & Nelder, 1989) is implemented in a number of statistical packages (e.g., GenStat, R, SPlus, SAS, SPSS, Stata). The simplest version of this GLM may be written as

\[
\ln[\pi_i/(1 - \pi_i)] = \beta_0 + \sum_b \beta_b X_{bi},
\]

where \(\pi_i\) is the proportion being estimated for the \(i\)th subject (the subject’s test score given as the proportion of items that are correct), \(\beta_0\) is the intercept, \(b = 1, \ldots, H\), and the \(X_{bi}\) are predictors with the \(\beta_b\) their respective coefficients. The \(\ln[\pi_i/(1 - \pi_i)]\) expression in Equation (1) is the logit of the \(\pi_i\), and this is the link function connecting \(\pi_i\) to the weighted linear combination of predictors in the right-hand side of that equation. The predicted value of \(\pi_i\) is then the inverse of the link function:

\[
\pi_i = \logit^{-1}\ln[\pi_i/(1 - \pi_i)] \\
= \exp(\beta_0 + \sum_b \beta_b X_{bi})/[1 + \exp(\beta_0 + \sum_b \beta_b X_{bi})].
\]

Because this link function is very popular in this context, the term logistic regression is often used for this version of the binomial GLM.

In the simplest applications of the binomial GLM, the intercept term, \(\beta_0\), is simply a constant, and we have a fixed-effects model. There is no term for individual subjects’ random variation around a mean, and so the predicted value of \(\pi_i\) is the same for subjects who are classified in the same way by the predictors. Thus, consider the binomial GLM with a single predictor, \(X\):

\[
\ln[\pi_i/(1 - \pi_i)] = \beta_0 + \beta_1 X_i.
\]

If the predictor is age, then the predicted value of \(\pi_i\) is the same for all subjects of the same age. If the predictor is case–control status, then the predicted value of \(\pi_i\) is the same for all control subjects (and the same for all cases, if there is more than one).

Assumptions

All statistical methods rest on assumptions, and the simplest models that treat the test score as a binomial random variable assume (a) that test item responses are statistically independent of one another (Assumption 1; independent items) and (b) that all items in a test have the same level of difficulty (Assumption 2; item homogeneity). When combining test scores for analysis, the simplest models also assume (c) that subjects are independent (Assumption 3; independent subjects) and (d) that they all possess the same level of ability measured by the test (Assumption 4; subject homogeneity). We have just seen, for example, how an application of the fixed-effects model in Equation (3) for case–control comparison would treat the subjects in the control sample as homogeneous.

More sophisticated models relax one or more of these assumptions (for an early discussion, see Lord & Novick, 1968), but most of the models in this paper continue to assume independent items (Assumption 1) and item homogeneity (Assumption 2). The primary reason for making these two assumptions is that, when we do not
have access to the item-by-item raw data, it is impossible to test for violations of them. These two assumptions about items have implications for Type I error rates and standard errors, and we outline them here, beginning with the assumption of independent items.

**Independent items, homogeneity of difficulty, and practice effects**

Positive autocorrelation among test items provides a simple example of violation of Assumption 1, independent items. It *increases* the variance of the test score—that is, the variance of the sum of the item scores—via the positive covariance between item pairs (e.g., Altham, 1978). Positive autocorrelation might result from similarity between test items or, in extreme cases, from repetition of items within the same test. In such cases, the positive autocorrelation would be similar to a practice effect. In general, however, practice effects may increase or decrease the variance of the test score depending on their impact on item difficulty. In order to explain this further, we turn to the issue of heterogeneity in item difficulty—that is, violation of Assumption 2, item homogeneity.

Heterogeneity in item difficulty decreases the variance of the sum of the item scores. This fact is probably not as well known as the positive autocorrelation result, so we present a brief argument for it here (see also, for example, Crosbie, 1987). Suppose we have two independent items of equal difficulty, given by the probability, $p$, of getting the item correct. Then the sum of the scores on the two items has variance $2\pi(1 - \pi)$. Now, suppose we have two independent items of different difficulties, whose mean difficulty is $\pi$. Let the first item difficulty level be $\pi + \theta$, and the second $\pi - \theta$, where $0 < \theta < \min(\pi, 1 - \pi)$. Then the variance of their sum is $(\pi + \theta)(1 - \pi - \theta) + (\pi - \theta)(1 - \pi + \theta) = 2\pi(1 - \pi) - 2\theta^2$, which is less than $2\pi(1 - \pi)$. This argument generalizes to any finite sum of independent items, so we have proven that heterogeneity in item difficulty decreases the variance of the sum of the item scores.

We are now in a position to explain the effects of practice on the variance of the test score. On the one hand, suppose that practice makes item difficulty more heterogeneous (e.g., by making easy items easier but having no impact on the most difficult). Then the range of difficulty will be increased, which will decrease the variance of the sum of the item scores. On the other hand, suppose that practice homogenizes item difficulty (e.g., by making all items equally easy). Then the range of difficulty will be decreased, and this will tend to increase the variance of the sum. However, if mean difficulty is shifted towards 0.5 it will increase the variance, whereas if it is shifted away from 0.5 then this will decrease the variance of the sum. Depending on how they co-occur and which of these two effects is stronger, the practice effect may increase or decrease the variance of the sum of the item scores.

We conclude that positive autocorrelation among items increases the variance of test scores, heterogeneity in item difficulty among independent items decreases the variance, and practice effects may increase or decrease the variance. When there is positive autocorrelation, a significance test (e.g., for a case–control difference) that assumes independent items (Assumption 1) is liable to underestimate the probability of a Type I error. In contrast, when there is heterogeneity in item difficulty, a significance test that assumes item homogeneity (Assumption 2) is liable to overestimate Type I error probability. Thus, if we are testing for a difference between a case and a control sample, the assumption of homogeneity of item difficulty is conservative. Nevertheless, as we shall see, exploiting test structure yields increases in statistical power with longer tests and also enables the researcher to take subject heterogeneity into account.

---

5 Suppose, as a hypothetical example, that each subject’s responses to Items 5, 10, 15, and 20 of a 20-item test are positively correlated with each other. For simplicity, suppose that the correlation is perfect, so that subjects who respond correctly to any one of the four items respond correctly to all of them (and likewise for incorrect responding). Two subjects who respond differently on any one of these items will differ in the same way on all four of them, and so differences in their ability will be amplified in their test scores.
Subject heterogeneity and overdispersion of test scores

When we do not have access to subject-by-subject data, but only to the mean score and standard deviation for a sample (usually a control sample), the modeling techniques that we have proposed thus far are compelled to assume independent subjects (Assumption 3) and subject homogeneity (Assumption 4). The assumption of independent subjects is usually regarded as relatively uncontroversial. But adoption of the assumption of subject homogeneity poses a problem for the cautious researcher because violation of the assumption may result in underestimation of variance across individuals. In reality, the scores of individuals in the sample may be overdispersed—they may vary around the sample mean more than would be expected if the test were equally difficult for all members of the sample.

We saw earlier that violation of Assumption 1, independent items, increases the variance of the test score. Consequently, violation of either Assumption 1, independent items, or Assumption 4, subject homogeneity, is a potential source of overdispersion. Where there is overdispersion, these two assumptions (in their most general form) are not both warranted. We therefore provide a test of overdispersion that can guide researchers in choosing between techniques that assume independent items and subject homogeneity and those that do not.

Overdispersion

A simple test for overdispersion of scores on a binary-item test is the chi-square approximation recommended by McCullagh and Nelder (1989, pp. 124–128):

$$\chi^2(N - 1) = \sum (r_i - n_i P)^2 / [n_i P(1 - P)],$$

(4)

where $N$ is the sample size, $r_i$ is the number of items correct for the $i$th subject, $n_i$ is the total number of test items, and $P$ is the total sample proportion of items correct. In our applications, everyone is doing the same test, so we may drop the $i$ subscript from $n_i$ and just use $n$. Note that the null hypothesis in this test is that the items are independent, and all members of the sample have the same ability level, $P$. If the $\chi^2(N - 1)$ statistic is significant, this null hypothesis is rejected, and there is overdispersion, whether due to positive autocorrelation or subject heterogeneity. This test is fairly conservative, but useful for identifying clear instances of overdispersion.

As McCullagh and Nelder (1989, p. 124) observed: “Over-dispersion is not uncommon in practice. In fact, some would maintain that over-dispersion is the norm.” Overdispersion could be expected to be even more common among cases than controls because, even if brain damage includes a specific site, the extent of the lesion will vary and may well result in quite different deficits for different patients. Thus, in neuropsychological assessments of case series, tests for overdispersion of test scores should be routinely conducted where possible.

Testing for overdispersion in a control sample

In some situations, we may propose to use statistical methods that require the assumption of subject homogeneity for a control sample (e.g., the fixed-effects binomial GLM in Equation (3) with case–control status as the only predictor). In such a situation, it is important to test for overdispersion of test scores in the control sample using the chi-square statistic in Equation (4). If overdispersion is detected, then we must either correct for it or take advantage of methods that relax the assumption of subject homogeneity. (A brief review of these alternatives is provided a few paragraphs below; see the section Accounting for overdispersion with the binomial GLM.)

It is also possible to test for overdispersion of binary-item test scores in a sample when we do not have the individual test scores but only the mean score and standard deviation, $s$, (or,
equivalently, the variance, \( s^2 \) for the sample. Given the sample size \( N \), test length \( n \), total sample proportion correct \( P \), and the variance of the scores \( s^2 \), the overdispersion test statistic in Equation (4) may be written as

\[
\chi^2(N - 1) = (N - 1)s^2/[nP(1 - P)].
\]

(5)

If overdispersion is detected using the chi-square statistic in Equation (5), then, as before, we must correct for it or adopt alternative methods. Even when the individual test scores are not available, the binomial GLM can still be used for a single test, provided that we correct for overdispersion by compensating for the model’s underestimation of the variance of the test scores that is consequent on the assumption of subject homogeneity.

Modelling overdispersion

Suppose that overdispersion of test scores is found in a sample for which the individual test scores are available—a control sample, a series of cases, or a mixed sample of controls and cases. As Hilbe (2009, pp. 320–322) notes, this overdispersion may be not “real” in the sense that it may be accounted for by predictors. A reasonable goal, then, is to account for the overdispersion using the predictors in an appropriate model.

Returning to the Table 1 example (Illustration 2), suppose we did not know which of the 6 subjects were cases and which controls. The overdispersion test in Equation (4) detects heterogeneity among the 6 subjects on Test 1, \( \chi^2(5) = 11.53 \), \( p = .042 \), but not on Test 2, \( \chi^2(5) = 2.95 \), \( p = .809 \). “Modelling” this overdispersion in the Test 1 scores by separating the case from the controls, and using case–control status as a predictor, accounts for the overdispersion. The overdispersion test does not show any significant overdispersion in the Test 1 scores for the five controls. It should be noted that the standard deviation of the control scores on Test 1 is exactly as would be expected if there were no overdispersion, whereas the standard deviation of the control scores on Test 2 is somewhat lower than would be expected (these scores are slightly, but not significantly, underdispersed).

Suppose now that, instead of a standard deviation of 5 for Test 1, the five control scores were {15, 21, 50, 79, 85}, for which the standard deviation is \( s = 32.14 \), and so the variance is \( s^2 = 1,033.69 \). Then the overdispersion statistic in Equation (5) would be \( \chi^2(4) = 4 \times 1,033/100 \times .5(1 - .5) = 165.69 \ (p < .0001) \), and we would conclude that the control scores were substantially overdispersed. Clearly the case score of 35 would no longer look out of place among the control scores, and the Crawford–Howell (Crawford & Howell, 1998) modified t test agrees with this impression, \( t(4) = -0.426 \), \( p = .692 \). In this context, the overdispersion test supplements the Crawford–Howell test by alerting the researcher or clinician to the heterogeneity of the control sample data and thereby to the potential existence of individual differences in the ability measured by the test.

Case series

The assumption of subject homogeneity is generally not warranted for a patient population (Caramazza, 1986), and we can use the overdispersion test to investigate heterogeneity in a series of cases. If overdispersion is detected in the scores on a test, then we can try to model the presumed heterogeneity using the predictors in a GLM.

Detecting overdispersion in a case series

Table 2 displays the scores on Immediate Recall and Delayed Recall of the Logical Memory subtest of the Wechsler Memory Scale–Revised (WMS–R; Wechsler, 1987) for a case series of 7 patients with unilateral neglect persisting more than three months after a right-hemisphere stroke (patients F2 and M1–M6 from Maguire & Ogden, 2002). The Logical Memory subtest has 50 items. The Delayed Recall scores appear

8 Here, we suppose that the overdispersion does not result from a violation of Assumption 1, independent items.
to be somewhat more variable than the Immediate Recall scores, and this is reflected in the dispersion statistics. The Immediate Recall scores are just short of showing significant heterogeneity, $\chi^2(6) = 12.328, p = .055$, but the Delayed Recall scores clearly show overdispersion, $\chi^2(6) = 18.713, p = .005$. Later, we consider how this “extra” variability in the Delayed Recall scores might be modelled using predictors.

Accounting for overdispersion with the binomial GLM

Hilbe (2009, pp. 342–345) summarizes six alternatives for dealing with overdispersion not accounted for by predictors in the binomial GLM:

1. Rescaling the standard errors with the Pearson/df statistic or an alternative estimate of overdispersion.
2. Fitting a beta-binomial or other generalized binomial GLM.
3. Robust variance estimators.
4. Bootstrap or jackknife standard error estimates.
5. Model reparameterizations.

6. Fitting a multilevel model with Gaussian random effects.

Rescaling the standard errors is the oldest and simplest alternative and can be effective for simple models. Hilbe (2009, p. 132) observes that simulation studies have suggested that the Pearson statistic (the statistic used in Equation (4) is the best alternative for rescaling. The beta-binomial model has a long history, but as Hilbe points out, the beta-binomial model often encounters difficulties in convergence and fitting. Robust estimators, bootstrap or jackknife methods, and model reparameterizations are beyond the scope of this paper and are not discussed here. Multilevel GLMs with Gaussian random effects are, however, generalizable and adaptive, and we offer them as the preferred method where they are feasible. When it is not possible to run a multilevel model, we recommend rescaling the standard errors to compensate for overdispersion.

Table 3 displays scores on a test of sustained attention—the Lottery subtest of the Test of Everyday Attention (TEA; Robertson, Ward, Ridgeway, & Nimmo-Smith, 1994)—for the same 7 patients as those in Table 2. The Lottery subtest has 10 items, and the overdispersion test indicates heterogeneity, $\chi^2(6) = 13.390, p = .037$. Impaired sustained attention has been shown to be associated with persisting unilateral neglect (Robertson, 2001), and so there is theoretical reason to investigate whether the results of an assessment of neglect in these patients predict their Lottery scores. Table 3 displays a neglect assessment score for each patient as the number of tests failed from a battery of 12 tests of unilateral neglect.

Table 2. Scores for 7 patients on Immediate and Delayed Recall of the Logical Memory subtest of the Wechsler Memory Scale—Revised

<table>
<thead>
<tr>
<th>Patients</th>
<th>Immediate</th>
<th>Delayed</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>38</td>
<td>37</td>
</tr>
<tr>
<td>M1</td>
<td>26</td>
<td>26</td>
</tr>
<tr>
<td>M2</td>
<td>34</td>
<td>20</td>
</tr>
<tr>
<td>M3</td>
<td>31</td>
<td>24</td>
</tr>
<tr>
<td>M4</td>
<td>37</td>
<td>32</td>
</tr>
<tr>
<td>M5</td>
<td>27</td>
<td>23</td>
</tr>
<tr>
<td>M6</td>
<td>36</td>
<td>33</td>
</tr>
</tbody>
</table>

9 Robertson et al. (1997) show that six measures of unilateral neglect are significantly correlated with patients’ scores on the Elevator Counting subtest of the Test of Everyday Attention (TEA). In the analysis of factor structure for the TEA, the Lottery and Elevator Counting subtests both load on sustained attention (Factor 3). The authors say (Robertson et al., 1994, p. 11): “These subtests go together in being sensitive to the ability to keep one’s mind on a relatively unchanging, even boring, task.”

10 The twelve tests are: Line Bisection (Schenkenberg, Bradford, & Ajaz, 1980), Line Erasure (Mark, Kooistra, & Heilman, 1988), Battersby Reading Test (Battersby, Bender, Pollack, & Kahn, 1956), Rey Complex Figure (Meyers & Meyers, 1995)–Copy condition with modified scoring (see Aimola, 1999), two Scene Tasks (Gainotti, D’Erme, Monteleone, & Silveri, 1986; Ogden, 1985), and six subtests (Line Crossing, Line Bisection, Letter Cancellation, Star Cancellation, Figure and Shape Copying, Representational Drawing) from the Behavioural Inattention Test (BIT; Wilson, Cockburn, & Halligan, 1987).
Traditionally, we could compute a correlation between sustained attention scores and neglect assessment scores, in which case we would find $r = -0.772 \ (p = .042)$. However, this procedure would fail to exploit the structure of the test. Using the fixed-effects binomial GLM in Equation (3) instead, we use the neglect assessment score as the predictor and test for an association between the sustained attention scores and neglect scores with the following model:

$$\ln\left[\frac{p_i}{1 - p_i}\right] = b_0 + b_1 X_i = 2.849 - 0.373 X_i,$$

where $X_i$ = neglect assessment score for the $i$th patient. If there is dispersion that is not accounted for by the predictor, then we can include rescaling of the standard errors of the coefficients using the Pearson dispersion statistic in Equation (4). Rescaling multiplies the standard errors by the square-root of $\chi^2(N - 1)/df$. If there is overdispersion, then this ratio will be larger than 1, and therefore the rescaled standard errors will be larger, making any significance test associated with them more conservative and widening the resulting confidence intervals around the coefficient estimates.

Without rescaling, the $b_1$ coefficient is significant, log-likelihood $\chi^2(1) = 8.847, \ p = .003$. However, we should now test whether significant overdispersion remains even after neglect scores have been taken into account. The dispersion statistic for the model with neglect scores as a predictor turns out to be $\chi^2(5) = 7.548$, with $p = .183$, so neglect has reduced overdispersion to a nonsignificant level.

Even though the remaining dispersion is nonsignificant, if we wish to be conservative we should still rescale the standard errors. With rescaling, the log-likelihood statistic still is significant but not as strongly so as before, $\chi^2(1) = 6.372, \ p = .012$. Without rescaling, the standard error for $b_1$ is 0.140. The dispersion statistic is $\chi^2(5) = 7.548$, so the rescaling factor is $(7.548/5)^{1/2} = 1.178$, and the rescaled standard error is $0.140 \times 1.178 = 0.165$. This larger standard error accounts for the decline in significance.

Although the number of patients in the case series is small, even adjusting for dispersion the Type I error probability is now considerably lower than for the correlation coefficient. The binomial GLM takes test length into account, whereas the correlation does not, and even such a short test length as 10 items makes an appreciable difference to statistical power.

Finally, we may apply Equation (2) to obtain the predicted scores shown in the rightmost column of Table 3. For example, patient M5’s neglect assessment score of 6 predicts a test item probability of $\exp(2.849 - 0.373 \times 6)/1 + \exp(2.849 - 0.373 \times 6) = .65$ and therefore a sustained attention test score of $10 \times .65 = 6.5$.

Consecutive predicted probabilities share the same odds-ratio, dictated by $1/\exp(-0.373) = 1.45$. Thus, the odds of getting a sustained attention test item correct are 1.45 times greater for a patient with a neglect score one unit lower. Of course, the fit between the predicted and actual sustained attention scores is imperfect, being worst for patients M4 and M6 but reasonably close for the others.

### Table 3. Scores for 7 patients on neglect assessment and sustained attention

<table>
<thead>
<tr>
<th>Patients</th>
<th>Neglect assessment</th>
<th>Actual</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>F2</td>
<td>1</td>
<td>10</td>
<td>9.2</td>
</tr>
<tr>
<td>M1</td>
<td>7</td>
<td>6</td>
<td>5.6</td>
</tr>
<tr>
<td>M2</td>
<td>8</td>
<td>4</td>
<td>4.7</td>
</tr>
<tr>
<td>M3</td>
<td>3</td>
<td>9</td>
<td>8.5</td>
</tr>
<tr>
<td>M4</td>
<td>7</td>
<td>8</td>
<td>5.6</td>
</tr>
<tr>
<td>M5</td>
<td>6</td>
<td>5</td>
<td>6.5</td>
</tr>
<tr>
<td>M6</td>
<td>4</td>
<td>6</td>
<td>8.0</td>
</tr>
</tbody>
</table>

Note: Lottery subtest of the Test of Everyday Attention.

Multilevel models

Earlier, we described four assumptions: Assumption 1, independent items; Assumption 2, item homogeneity; Assumption 3, independent subjects; and Assumption 4, subject homogeneity. The fixed-effects binomial GLM employed thus
far makes all of these assumptions. In the example of sustained attention scores, any variation among the subjects’ scores is assumed to be due to their neglect assessment score. That is, the model assumes subject homogeneity (Assumption 4) once neglect scores have been taken into account. Thus, for example, patients M1 and M4 have the same neglect assessment score, and so, because we are using a fixed-effects model, their predicted sustained attention scores are also the same. In the previous section, we have shown how to rescale coefficient standard errors when this assumption turns out not to be true.

We shall now introduce multilevel models that relax Assumptions 2 and 4, item and subject homogeneity. By modelling heterogeneity among items or subjects (or both), the multilevel GLM makes weaker and more plausible assumptions about item and subject independence than Assumptions 1 and 3. This type of model assumes conditional independence, which means that the average scores for individual items are considered independent of one another once the heterogeneity in item difficulty has been taken into account, and individual subjects’ scores are considered independent of one another once the heterogeneity in subject ability levels has been taken into account. These are the assumptions made by most item-response models in the psychometric literature (Agresti, 1990, pp. 399–400) because they are multilevel models in another form.

Adding random effects
The fixed-effects model of sustained attention scores assumes subject homogeneity once neglect scores have been taken into account. However, it is possible that the variability in the sustained attention scores could have two sources, one being neglect and the other resulting from differences in subjects’ ability to sustain attention that cannot be accounted for by neglect. A model that takes this additional source of variability into account is a random intercepts model, written as

\[ \ln[\pi_i/(1 - \pi_i)] = \beta_{0i} + \beta_1X_i, \]

where \( \beta_{0i} = \mu + \nu_i \). The right-hand side of Equation (6) is conceptually similar to a mixed-design analysis of variance (ANOVA) with one within-subjects and one between-subjects factor. The \( \beta_{0i} \) term is the intercept for the \( i \)th subject, the \( \mu \) term is the mean of the intercepts, and the \( \nu_i \) term is the subject effect. This term treats the subject effect as a random effect, whereas the \( \beta_1 \) term treats the effect of neglect as a fixed effect. Thus, \( \nu_i \) is normally distributed with a mean of 0 and its own standard deviation, and the standard deviation provides a measure of how dispersed the individual ability levels are around the line with intercept \( \mu \) and slope \( \beta_1 \). Because we have a fixed effect for the predictor, and the subject effect is treated as a random effect, this GLM is a mixed model, also known as a multilevel model (MLM).

For the example of sustained attention scores, the new model with the subject effect term turns out to be nearly identical to the old model with only the fixed-effect term. The \( \beta_1 \) coefficient has the same value as before and is still significant, log-likelihood \( \chi^2(1) = 7.06, p = .008 \), at a similar level to that obtained with rescaling the standard errors. This model has less statistical power than its fixed-effects unrescaled counterpart because of the extra \( \nu_i \) term. The estimated standard deviation of the \( \nu_i \) term is 0.00008, so the dispersion of individuals’ sustained attention scores around the line with intercept \( \mu \) and slope \( \beta_1 \) (that is, the dispersion once the neglect assessment scores have been taken into account) is tiny.\(^{11}\) A likelihood-ratio test between this model and the old one gives \( \chi^2(1) < 0.0005, p > .999 \), indicating that the new model adds nothing to the old one. In this instance, we could be content with a model that omits the subject effect term, just as we could be content without rescaling the standard errors in the fixed-effects GLM. But, as with rescaling the standard errors, including the subject effect term ensures that we have a conservative test of our model.

\(^{11}\) Note that the \( \nu_i \) term models random variation in a population parameter \( (\pi_i) \), not random variation in a sample of scores, which contain an additional variation component due to sampling error.
For an example where it would be advisable to incorporate the subject effect term or at least rescale the standard errors, we consider models in which the neglect assessment score is used to predict the Delayed Recall scores on the Logical Memory subtest of the WMS–R (Table 2). We already know that the Delayed Recall scores are overdispersed. A fixed-effect binomial GLM as in Equation (3) yields

\[
\ln[\pi_i/(1 - \pi_i)] = \beta_0 + \beta_1 X_i = 0.869 - 0.123X_i.
\]

The \( \beta_1 \) coefficient is again negative and significant, log-likelihood \( \chi^2(1) = 7.07, p = .008 \). But significant overdispersion remains even after neglect scores have been taken into account, \( \chi^2(5) = 12.070, p = .034 \). Moreover, if we rescale the standard errors using the Pearson dispersion statistic, the log-likelihood statistic is no longer significant, \( \chi^2(1) = 2.928, p = .087 \).

A random-intercepts model as in Equation (6) gives a similar picture, with \( \beta_1 = -0.126 \), with \( \chi^2(1) = 3.300, p = .069 \). Not only does \( \beta_1 \) have a slightly different value (-0.126 instead of -0.123), it is also no longer significant. The reason for this is that some of the extra dispersion has been found to be due to the subject effect. The estimated standard deviation of the \( \psi_i \) term is 0.248, which is not negligible. Even though the likelihood-ratio test between this model and the fixed-effects model is not significant, \( \chi^2(1) = 1.28, p = .129 \), it would be wise to adopt this more conservative model for testing whether neglect assessment scores predict Delayed Recall scores. As with rescaling standard errors, there seems to be no rule of thumb for deciding when to use a MLM instead of a fixed-effects GLM in this setting, but a researcher who wishes to err on the side of conservatism rather than liberalism would opt for the MLM or at least rescaling. However, significant overdispersion remaining after predictors have been entered into a fixed-effects model is sufficient reason to rescale or move to a MLM.

**Heterogeneity in item difficulty**

On some occasions, a clinical researcher may have access, not only to individual scores, but also to the item-by-item raw data from a test administered to a series of cases (e.g., the Logical Memory subtest of the WMS–R or the Lottery subtest of the TEA). If so, then heterogeneity in item difficulty can be taken into account as well as subject heterogeneity. A model with a single predictor, such as the neglect assessment score in our examples, and with an item effect term as well as a subject effect term, would be written as

\[
\ln[\pi_{ij}/(1 - \pi_{ij})] = \beta_{0ij} + \beta_1 X_{ij},
\]

where the subscript \( i \) indexes the cases, \( j \) indexes the items, and \( \beta_{0ij} = \mu + \psi_i + \theta_j \). The \( \theta_j \) term captures the deviation of the average score for the \( j \)th item from the grand mean of the item scores, thereby providing an estimate of how difficult that item is relative to the entire collection. We do not provide examples or further elucidation of this model, with an item effect term as well as a subject effect term, because that is not the focus of this paper. Henceforth, we return to our usual

12 The models were estimated in MLwiN 2.1 (Rasbash et al. 2009), Stata10 (StataCorp, 2007; using xologit and GLLAMM), SAS 9.2 (SAS Institute Inc., 2008), and R 2.10.0 (R Development Core Team, 2011; using the lme4 package, Bates, Maechler, and Bolker, 2011). All routines gave very similar results, and we report the SAS results here. The MLMs presented in this paper can be estimated in the versions of SAS, Stata, MLwiN, R, and WinBUGS current at the time this paper was written. We recommend SAS, Stata, R, and WinBUGS over MLwiN, which does not report a log-likelihood statistic for these models. The Appendix offers further guidance; worked examples in R, SAS, and Stata are available online at: http://dl.dropbox.com/u/1857674/Binomial_GLM/binom.html

13 It is worth noting that item-by-item raw data for seven cases on Delayed Recall of the Logical Memory subtest of the Wechsler Memory Scale–Revised would provide 350 data-points while, with an item effect term as well as a subject effect term, there would be 59 parameters to estimate. In the case of the Lottery subtest of the Test of Everyday Attention, item-by-item raw data for seven cases would provide 70 data-points with 19 parameters to be estimated if an item effect term and a subject effect term were included in the model. Maximum likelihood estimation usually requires at least moderately large samples (a minimum of about 100 data-points). Another rule of thumb that is sometimes invoked is that there should be at least 10 observations per parameter to be estimated (Long, 1997).
assumption that we have access to the test scores for individual subjects but not to the item-by-item raw data. Our MLMs include a subject effect term but not an item effect term, and Assumptions 1 and 2 about items are made.

Case–control comparison

Now suppose we have the individual subjects’ test scores for a sample of controls and a single case, as is true of the data in Table 1. We may apply the MLM with a subject effect term as in Equation (6) to the task of comparing the case with the control sample, by using case–control status as a categorical predictor. Note that using case–control status as a predictor does not require us to include a subject effect term, but doing so gives us a more conservative model. It relaxes Assumption 4, subject homogeneity, and allows that there may be variability in subjects’ test scores that cannot be accounted for by case–control status. Specifically, the MLM with case–control status as its single predictor allows that there may be variability among the subjects in the control sample.

Testing for a case–control difference

Recalling that we found overdispersion in the scores (for 5 control subjects plus the case) on Test 1 in Table 1 but not in the scores on Test 2, we should expect that the MLM will find a case–control difference on Test 1 but not on Test 2.

Applying the model in Equation (6) to Test 1 with $X_j = 0$ for the controls and 1 for the case, the estimate of the $\beta_1$ coefficient is $-0.619$, which significantly differs from 0, log-likelihood $\chi^2(1) = 7.64, p = .006$. However, the $\beta_1$ estimate for Test 2 fails to differ significantly from 0, $\chi^2(1) = 1.29, p = .256$, echoing our inability to find significant overdispersion in the Test 2 scores. Considered as proportions, the case score of 35/100 on Test 1 is closer to the control Test 1 mean of 50/100 ($0.5 – 0.35 = 0.15$) than the case score of 3/10 on Test 2 is to the control Test 2 mean of 5/10 ($0.5 – 0.3 = 0.2$). Nevertheless, the standard error of the $\beta_1$ estimate for Test 1 is only 0.228 whereas for Test 2 it is 0.746, more than three times as large, and this is due to the considerably greater length of Test 1.

Case–control comparison on two tests

What happens if we model the two tests together? The MLM required to do this can be written in the following way (with the $\beta_{0i}$ term henceforth unpacked as the sum $\mu + \psi_i$):

$$\ln[\pi_{ijk}/(1 – \pi_{ijk})] = \mu + \psi_i + \beta_1 X_j + \beta_2 Z_k + \beta_3 X_j Z_k.$$  

(8)

On the left-hand side of this formula, $\pi_{ijk}$ is the $i$th subject’s expected score (given as the proportion of items that are correct) on the $j$th test, and $k$ indexes whether the subject is a case or a control. Thus, $j = 1$ or 2, and $k = 1$ or 2. As before, the right-hand side of Equation (8) is conceptually similar to a mixed-design ANOVA with one within-subjects and one between-subjects factor. The $\mu$ term is the intercept, the $\psi_i$ term is the subject effect, and the predictors are treated as fixed effects.14

The $\beta_1$ coefficient is the main effect for test, and $X_j$ is a variable that equals 0 when $j = 1$ and 1 when $j = 2$, so $\beta_1 X_j$ is 0 for Test 1 and is $\beta_1$ for Test 2. Likewise, $\beta_2$ is the main effect for case–control status, and $Z_k$ is a variable that equals 1 when $k = 1$ (case) and 0 when $k = 2$ (control), so $\beta_2 Z_k$ is $\beta_2$ for the case and is 0 for the controls. That is, $\beta_2$ plays the role that $\beta_1$ played in the earlier model of Test 1 scores. Finally, $\beta_3$ is the interaction effect between case–control status and test (Test 1 versus Test 2), so that the sum $\beta_2 + \beta_3$ plays the role that $\beta_1$ played in the earlier model of Test 2 scores.

14 Note that, although Equation (8) includes a subject effect term, it does not include an additional random effect term ($\omega_{ij}$, say) to allow the difference between the two test scores to vary across subjects. The reason for this is that, with just two tests, there are insufficient degrees of freedom for this term to be included (see, e.g., Kenny, Kashy, & Cook, 2006, p. 89). However, if we were fortunate enough to have two collections of tests (as might be the case when there are alternative forms of the same test), then we could extend the model in Equation (8) to include this random term.
Applying the model in Equation (8) to the Test 1 and Test 2 scores in Table 1 yields an estimate for the case–control main effect $\beta_2$ of –0.619, identical to the $\beta_1$ estimate in the earlier model of Test 1 scores and as strongly significant. The main effect for Test is very close to 0 and not significant, as should be expected because the control means, considered as proportions, are identical (50/100 and 5/10), and only the case scores differ (35/100 versus 3/10). Finally, the interaction effect, $\beta_3$, is estimated as –0.228, and it fails to reach significance ($p = .770$); again this is because its standard error is large (0.780).

Note that $\beta_2 + \beta_3 = –0.619 – 0.228 = –0.847$, identical to the $\beta_1$ estimate in the earlier model of Test 2 scores. Our new model makes the same predictions as the earlier separate models did. So, why should we bother modelling both tests together? The answer is that we obtain a test of an interaction effect, which tells us whether the case–control difference differs between Test 1 and Test 2. In a later section, we show how to use the model in Equation (8) to test for dissociation of impairments on two tests, following the pattern established by Crawford and colleagues (Crawford & Garthwaite, 2005; Crawford et al., 2003), and then how to generalize the model to test for double dissociation.

Statistical power, test length, and control sample size

Before we proceed, recall that the significance test for the interaction term in the MLM in Equation (8) is unable to detect a difference between a case score of 35 versus a control mean of 50 on Test 1 and a case score of 3 versus a control mean of 5 on Test 2. While it is possible that there is really no interaction to detect (or that it is very small), it is also possible that the interaction is real but goes undetected because of insufficient statistical power.

At this point, it would be natural to ask how long a test would need to be, and how large the control sample should be, in order to provide sufficient statistical power. Unfortunately, a thorough investigation of these questions is beyond our scope, in part because it would require specifying dispersion levels, effect sizes, and confidence levels. Nevertheless, we can provide some important and practical intuitions on these matters, using a few simple simulations.

The same questions about statistical power—but now about power to detect a simple case–control difference, rather than an interaction—could be raised by the fact that the MLM in Equation (6) is unable to detect a case–control difference on Test 2. In fact, even the fixed-effect binomial GLM in Equation (3), which is less conservative than the corresponding MLM, still finds a non-significant coefficient for case–control difference on Test 2 ($\beta_1 = –0.847, p = .256$).

The data-set for Test 2 has 10 data-points from each of 6 subjects, so it falls well short of the 100-data-point rule of thumb that we mentioned in Footnote 13. The three ways to increase data-points are by increasing test length, the number of controls, and the number of cases. We now present a brief and impressionistic overview of the effectiveness of each of these options in case–control comparison using the methods described in this paper. Our review is based on runs of 10,000 simulated samples from a control population where the probability of a correct response is .6 and a case population where it is .3. To simplify matters, we begin by assuming no overdispersion and then add overdispersion in a second round of simulations.

First, assuming a test length of 10, a sample of 10 controls, and a single case, the probability of detecting the difference between the two populations turns out to be approximately .283. Doubling the test length (i.e., from 10 to 20) increases that power to about .672. However, doubling the size of the control sample does not have nearly the same impact. Power increases only to approximately .317. One reason for this is that we still have only one case. Doubling the control sample adds data only to the controls, whereas
doubling the test length adds data to both cases and controls. Adding 10 people to the total sample by adding 9 more controls and just 1 more case instead raises power to .682, about the same effect from doubling test length.\footnote{If we are considering a fixed-effect model with case–control status as the single predictor, then the cases—if more than one—will be assumed to have the same level of ability on the test, allowing for random variability.}

The picture changes slightly with overdispersion. In the second round of simulations, we assume that the control probability of a correct response follows a beta(3, 2) distribution, thus with mean probability .6 and precision equal to 5. The 10th percentile of this distribution is about .32, and the 90th percentile is about .86, so the control probabilities are indeed overdispersed. Assuming again a test length of 10, a sample of 10 controls, and a single case, the probability of detecting the difference between the two populations turns out to be approximately .389. Doubling the test length (i.e., from 10 to 20) increases power to about .588, which, because of overdispersion, is not as great an increase as before. Doubling the size of the control sample increases power to .418, about the same impact as before but still substantially less than doubling test length. Adding 10 people to the total sample by adding 9 more controls and just 1 more case instead raises power to .598, again about the same effect as doubling test length.

**Interim summary**

We have described appropriate statistical methods for exploiting the internal structure of a binary-item test, and we have explained the assumptions on which those methods rely. Taking account of test length increases statistical power even when the control sample is small or when a case series involves relatively few patients. The methods described here enable a neuropsychologist to test for overdispersion of test scores in a case series or control sample and to model overdispersion using categorical and continuous predictors such as gender, case–control status, age, or neglect assessment score. The methods can also be used to compare cases with a control sample on one or more tests. A neuropsychologist who is comparing a single case with a control sample should, if possible, use longer rather than shorter tests.

Ideally, a neuropsychologist should retain item-by-item raw data for analysis rather than simply aggregating test item responses into a test score. With sufficient data, this would allow the fullest use of the methods described in this paper, including even taking heterogeneity of item difficulty into account. Nevertheless, it is possible to exploit the structure of a binary-item test without having the item-by-item raw data, and the applications that we have described assume that only the individual subjects’ test scores are available.

In the remainder of the paper, we extend the methods for case–control comparison, in order to test for dissociation and double dissociation of impairments.

**Testing for dissociation**

The approach described in this paper can be elaborated to test for dissociation of impairments on two tests, following the pattern established by Crawford and colleagues (Crawford & Garthwaite, 2005; Crawford et al., 2003; see the section below, *Following the pattern*). In our framework, the statistical tests for dissociation amount to testing for an interaction effect between case–control status and test (Test 1 versus Test 2) and then carrying out what is known in ANOVA as a “simple effects” analysis. In this instance, the analysis proceeds by testing for a case–control difference on each test separately, provided that a significant interaction effect has been found.\footnote{We note two limitations of our approach. First, it assumes that Test 1 and Test 2 share the same kind of distribution. The question of how case–control differences on two tests with different kinds of underlying distribution can be compared raises issues beyond the scope of this paper. Second, our approach does not directly address the issue of dissociations between an impairment measured by a test and another impairment that is assessed in a categorical way. (Nevertheless, the binomial GLM can be applied to this problem if the test has an appropriate structure.)}
Procedure for identifying dissociation

The MLM required to test for dissociation of impairments on two tests is identical to Equation (8), repeated here for convenience:

$$\ln[\pi_{ijk}/(1 - \pi_{ijk})] = \mu + v_i + \beta_1X_j + \beta_2Z_k$$
$$+ \beta_3X_jZ_k.$$  

(8)

Recall that $\pi_{ijk}$ is the $i$th subject’s score on the $j$th test, and $k$ indexes whether the subject is a case or a control. The $\beta_1$ coefficient is the main effect for test, $\beta_2$ is the main effect for case–control status, and $\beta_3$ is the interaction effect. We now describe a procedure for identifying dissociation:

**Step 1.** Test whether the model in Equation (8) fits the data significantly better than a no-effects model—the no-effects model is $\ln[\pi_{ijk}/(1 - \pi_{ijk})] = \mu + v_i$. Often this can be done via the log-likelihood chi-square test. If the model fits better than the no-effects model, proceed to Step 2.

**Step 2.** Test whether the $\beta_3$ (interaction) term is significant. That is, compare the model in Equation (8) against a model without the interaction term:

$$\ln[\pi_{ijk}/(1 - \pi_{ijk})] = \mu + v_i + \beta_1X_j$$
$$+ \beta_2Z_k.$$  

If $\beta_3$ is not significant, then no dissociation can be said to have occurred. If $\beta_3$ is significant, then some kind of dissociation may have occurred. The specific kind is then determined by the simple effects analysis in Step 3.

**Step 3.** For each of Test 1 and Test 2, ascertain whether there is a significant difference between the case and controls and, if so, whether that difference indicates a deficit for the case. This is known as a “simple effects” analysis because we are testing the case–control main effect for each test separately.

For $j = 1$ (Test 1), this amounts to comparing two models:

$$\ln[\pi_{ijk}/(1 - \pi_{ijk})] = \mu + v_i + \beta_2Z_k$$

versus

$$\ln[\pi_{ijk}/(1 - \pi_{ijk})] = \mu + v_i$$

using Test 1 scores only (recall that, for Test 1, $\beta_1X_j = 0$ and $\beta_3X_jZ_k = 0$). The result must be a significant negative coefficient $\beta_2$ to indicate a deficit for the case on Test 1.

Likewise for $j = 2$ (Test 2), the models to be compared are:

$$\ln[\pi_{ijk}/(1 - \pi_{ijk})] = \mu + v_i + \beta_4Z_k$$

versus

$$\ln[\pi_{ijk}/(1 - \pi_{ijk})] = \mu + v_i$$

using Test 2 scores only, where conceptually $\beta_4$ is the simple effects equivalent of $\beta_2 + \beta_3$ in Equation (8). The result must be a significant negative $\beta_4$ to indicate a deficit for the case on Test 2.\footnote{Users may wish to follow Crawford and Garthwaite (2005) in applying one-tailed tests instead of the two-tailed tests that typify a GLM significance test for coefficients. We follow the more conservative two-tailed alternative, which allows researchers to identify and take into account unexpected findings.}

If the case shows a deficit on one test but is not significantly different from controls on the other test, then we have a classical dissociation. If the case shows deficits on both tests, then we have a nonclassical dissociation.\footnote{Our procedure tests for case–control by test interaction and then for case–control difference on each test and allows us to detect significant differences that are required for dissociation of impairments. We recognize that very long tests could yield significant but small effects, but it is beyond the scope of this paper to give the issue of effect size the treatment that it merits (see Crawford, Garthwaite, & Porter, 2010). Many neuropsychological tests are not very long, and so the issue of trivial but significant effect sizes may arise fairly rarely in practice.}

**Applying the procedure**

We illustrate this procedure for identifying dissociation by applying it to the hypothetical data
in Table 4, consisting of two sets of test scores with a sample of 10 controls and 1 case. The length of Test 1 is 50 items, and the length of Test 2 is 20 items. The scores on the two tests are correlated (the correlation for the controls is .401). Test 1 scores were randomly generated from a binomial distribution with probability of a correct answer $p = .5$, whereas Test 2 scores were randomly generated from a binomial distribution with $p = .75$, and these were reordered so as to induce a modest correlation between the tests.

Table 4 shows that, on Test 1, the case score is $17/50 = 0.340$, compared with the control average of 0.498. The numerical difference of 0.158 with a standard deviation of 0.080 results in a $z$ score for the case of $(0.340 - 0.498)/0.080 = -1.98$. On Test 2, the case score is $5/20 = 0.250$, compared with the control average of 0.745. The numerical difference of 0.495 with a standard deviation of .107 results in a $z$ score for the case of $(0.250 - 0.745)/.107 = -4.63$. The initial impression given by the numerical differences and $z$ scores is that the case performs worse relative to the controls on Test 2 than on Test 1.

The MLM in Equation (8) confirms dissociation of impairments on the two tests, and the results are summarized in Table 5. Step 1: Two of the three coefficients are significant, so clearly the model is significantly better than a no-effects model. Step 2: The interaction effect is significant according to the Wald test for the $\beta_3$ coefficient ($p = .015$). The Wald test is effectively a $t$ test for each coefficient (Agresti, 1990, p. 89). Another way of thinking about it is as a comparison of goodness of fit between a model that includes the $\beta_3$ coefficient and one without it. A better test statistic for comparing two models’ goodness of fit is twice the difference between the negative log-likelihoods for the models, but in this instance it gives a nearly identical $p$-value, $\chi^2(1) = 6.4021$, $p = .011$. So we can move to Step 3:

For Test 1, a simple effects analysis gives $\beta_2 = -0.656$ ($SE = 0.321$, two-tailed $p = .041$) but the model comparison log-likelihood chi-square test gives $\chi^2(1) = 3.703$, $p = .054$. If we are cautious and prefer the chi-square test, we would conclude that this effect has not quite reached significance. Note that the $-0.656$ estimate is close but not identical to the $-0.658$ value obtained for $\beta_2$ in the complete model in Table 5.

For Test 2, a simple effects analysis yields $\beta_4 = -2.171$ ($SE = 0.541$, two-tailed $p < .0005$), and the model comparison log-likelihood chi-square

---

**Table 4. Testing for dissociation: Hypothetical scores for 10 controls and 1 case on two tests**

<table>
<thead>
<tr>
<th>Test</th>
<th>Raw data</th>
<th>Summary data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Control 1</td>
<td>Test length</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>23</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>21</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>25</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>24</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>33</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>29</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>50</td>
</tr>
</tbody>
</table>

Note: $P = \text{mean/length}$.

---

20 In situations where only the control mean, standard deviation, sample size, and test length are available, one might consider using the overdispersion test in Equation (5) to obtain the Pearson chi-square statistic and then applying the binomial GLM with rescaled standard errors. (The reason that the control mean, sample size, and test length are sufficient information for the binomial GLM is that the sample estimate of the mean of a binomial distribution is $N \times P$, and the variance, without overdispersion, is $N \times P \times (1 - P)$, where $N$ is the sample size, and $P$ is the proportion of correct answers in the sample.) However, where two tests are involved, it is impossible to estimate an appropriate rescaling factor. Even in situations involving only one test, researchers should be cautious about using this approach because under some conditions it may not account for the full extent of overdispersion.
test gives $\chi^2(1) = 10.586$, $p = .001$. Note that the $-2.171$ estimate for $b_4$ is close but not identical to the $-0.658 - 1.523 = -2.181$ value obtained for $b_2 + b_3$ in the complete model in Table 5.

Step 2 of the procedure reveals a significant case–control by test interaction, and the simple effects analysis in Step 3 shows that the dissociation is near the borderline between classical and nonclassical. The case shows a clear deficit on Test 2 (in the simple effects analysis, the $b_4$ coefficient is negative and highly significant) and a borderline deficit on Test 1 (in the simple effects analysis, the $b_2$ coefficient is negative, and two tests of significance yield $p = .041$ and $p = .054$). The predicted proportions for every combination of case–control status and test are shown in the bottom row of Table 5, demonstrating that the complete model nearly perfectly reproduces the control means and the case scores (converted into proportions).

**Following the pattern**

The procedure that we have described for identifying dissociation follows the pattern established by Crawford and colleagues. As we have mentioned earlier, Crawford and Howell (1998) describe a modified $t$ test, to assess whether the performance of a case is significantly different from the performance of a control sample (without treating the control sample as if it were a population). Crawford, Howell, and Garthwaite (1998) also describe a modified paired-samples $t$ test for comparing the difference between a single case’s scores on two tests with the corresponding inter-test differences in the control sample. Crawford and Garthwaite (2005) provide an improved method for making this comparison, the Revised Standardized Difference Test (RSDT).  

Drawing on these tests, Crawford and colleagues (Crawford & Garthwaite, 2005; Crawford et al., 2003) offer explicit criteria for classical and nonclassical dissociation. Their criteria for classical dissociation, for example, are as follows (Crawford & Garthwaite, 2005, Table 5, p. 326):  

1. The case’s performance on one test (Test X) is significantly lower than that of the controls according to the Crawford–Howell (Crawford & Howell, 1998) modified $t$ test.
2. According to the same Crawford–Howell test, the case’s performance on the other test

---

**Table 5. Testing for dissociation: Multilevel model results**

<table>
<thead>
<tr>
<th>Coef.</th>
<th>SE</th>
<th>Wald test $p$</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 1</th>
<th>Test 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu$</td>
<td>-0.008</td>
<td>0.103</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$v_i$</td>
<td>0.157</td>
<td>0.119</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$b_2$</td>
<td>1.086</td>
<td>0.186</td>
<td>$&lt;.0001$</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$b_3$</td>
<td>-0.658</td>
<td>0.353</td>
<td>0.062</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$b_4$</td>
<td>-1.523</td>
<td>0.626</td>
<td>0.015</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Predicted proportion .498 .746 .339 .249

*Note: $Z_k = 0$ if it is a control, = 1 if it is a case. $X_j = 0$ if it is Test 1, = 1 if it is Test 2.

*The $v_i$ term is an estimated standard deviation.

---

21 More recently, Crawford and Garthwaite (2007) have proposed a Bayesian Standardized Difference Test and Bayesian criteria for dissociation. It is also possible to use a Bayesian approach to the models discussed in this paper.

22 Crawford and Garthwaite (2007, p. 364) suggest that a pattern of performance meeting these criteria should be described as “a dissociation, putatively classical”. The reason is that, for lack of statistical power, “many patients who in reality have a [nonclassical] dissociation will be classified as exhibiting a classical dissociation because the deficit on the less impaired of the two tasks will frequently not be detected” (see also Crawford & Garthwaite, 2006).
(Test Y) is not significantly lower than that of the controls.

3. The case’s performance on Test X is significantly worse than that on Test Y according to the Crawford–Garthwaite (Crawford & Garthwaite, 2005) RSDT.

Thus, Step 2 in our procedure, where we test whether the interaction term is significant, corresponds to Criterion 3 in Crawford and colleagues’ pattern. They assess whether the case’s scores on Test 1 and Test 2 differ significantly, using the Crawford–Garthwaite (2005) RSDT. If the scores differ, then some kind of dissociation, classical or nonclassical, may have occurred.

The simple effects analysis in Step 3 of our procedure corresponds to Criteria 1 and 2 in Crawford and colleagues’ pattern. They use the Crawford–Howell (Crawford & Howell, 1998) modified t test to assess, for each of Test 1 and Test 2, whether the case’s score is significantly lower than that of the controls—that is, whether the case has a deficit. If the case shows a deficit on only one test, then we have a classical dissociation; if the case shows deficits on both tests, then we have a nonclassical dissociation.

It may be of some interest to apply Crawford and colleagues’ tests to the data from Table 4. The Test 1 mean and standard deviation for the controls are 24.90 and 4.01, respectively, the Test 2 mean and standard deviation are 14.90 and 2.13, and the sample correlation between the tests is $r = .401$. Applied to Test 1, the Crawford–Howell (Crawford & Howell, 1998) modified t test yields $t(9) = -1.877$, $p = .0932$, and so does not find a significant difference between the case and the control sample. Applied to Test 2, the modified t test yields $t(9) = -4.428$, $p = .0932$, thereby finding a significant case–control difference. These results are broadly in line with our simple effects analysis. However, in contrast to our finding of a case–control by test interaction, the Crawford–Garthwaite (Crawford & Garthwaite, 2005) RSDT yields $t(9) = 2.124$, $p = .06265$, and so does not find a significant difference in the case’s performance between Test 1 and Test 2.

**Testing for double dissociation**

The key to testing for double dissociation is a MLM that generalizes the one in Equation (8) to handle two distinct cases, Case 1 and Case 2 (with the $\beta_0$ term again unpacked as $\mu + \upsilon_i$):

$$\ln[\pi_{ijk}=(1-\pi_{ijk})] = \mu + \upsilon_i + \beta_1 X_j + \beta_2 Z_{1k} + \beta_3 Z_{2k} + \beta_4 X_j Z_{1k} + \beta_5 X_j Z_{2k}.$$  

(9)

The main additional feature of this model is that $Z_{1k}$ is 1 for Case 1 and 0 otherwise, and $Z_{2k}$ is 1 for Case 2 and 0 otherwise. Thus the coefficients for both cases compare them to the controls. We can see this by considering what Equation (9) reduces to for each case.

For Case 1, $Z_{1k} = 1$ whereas $Z_{2k} = 0$ so Equation (9) becomes

$$\ln[\pi_{ijk}=(1-\pi_{ijk})] = \mu + \upsilon_i + \beta_1 X_j + \beta_2 Z_{1k} + \beta_4 X_j Z_{1k}.$$  

For Case 2, $Z_{1k} = 0$ whereas $Z_{2k} = 1$ so Equation (9) becomes

$$\ln[\pi_{ijk}=(1-\pi_{ijk})] = \mu + \upsilon_i + \beta_1 X_j + \beta_3 Z_{2k} + \beta_5 X_j Z_{2k}.$$  

Both of these equations have the same form as Equation (8). Consequently, their respective pairs of coefficients ($\beta_2$ and $\beta_4$ for Case 1; $\beta_3$ and $\beta_5$ for Case 2) may be tested for dissociation and deficit using the three-step procedure outlined above.

For Case 1, we test the interaction term $\beta_4$ for significance to discover whether there is dissociation of impairments on the two tests. If there is dissociation, then we use a simple effects analysis to test $\beta_2$ (for Test 1) and the simple effects equivalent of $\beta_2 + \beta_4$ (for Test 2) to identify deficits and determine whether the dissociation
is classical or nonclassical. For Case 2, similarly, we test the interaction term $\beta_5$ to discover whether there is a dissociation and use a simple effects analysis to test $\beta_3$ (for Test 1) and the simple effects equivalent of $\beta_3 + \beta_5$ (for Test 2) to identify deficits. If both cases demonstrate dissociation of impairments, then we check that the interaction terms $\beta_4$ and $\beta_5$ have opposite signs (positive and negative). This confirms that the two dissociations are complementary—that is, that one case performs better on Test 1 than on Test 2 while the other case shows the reverse pattern.

REFERENCES


StataCorp (2007). *Stata Statistical Software: Release 10*. College Station, TX: StataCorp LP.


APPENDIX

Brief guide to implementing the binomial GLM and MLM

Several statistical computing environments have the capability to implement the models discussed in this paper. This guide is not exhaustive but instead focuses on the more popular and readily available packages, primarily R, SAS, SPSS, and Stata. Code and instructions for fully worked examples from this paper in each of these packages may be obtained at this web address: http://dl.dropbox.com/u/1857674/Binomial_GLM/binom.html

Binomial GLMs with rescaling for overdispersion

Binomial GLMs whose coefficients’ standard errors can be rescaled with the Pearson statistic are available in many packages such as GenStat, R, SAS, SPSS, and Stata. Some of them offer a choice of alternatives to the Pearson statistic (usually the deviance chi-square or the Williams procedure).

Example 1: Binomial GLM for sustained attention scores (Table 3. The example we present here uses SAS, but most of the command elements are common to the other packages. The requisite data include the predictor(s) (in this case, neglect scores), the number of correct test items for each subject (denoted by "r"), and the test length (denoted by "n"):

```sas
title 'sustained attention';
data sus;
input subno neglect r n;
cards;
1 1 10 10
2 7 6 10
3 8 4 10
4 3 9 10
5 7 8 10
6 6 5 10
7 4 6 10;
run;
```

The binomial GLM procedure then specifies the model, with \( r/n \) being predicted by neglect, and the Pearson rescaling statistic (the latter by the subcommand "scale = P"):

```sas
proc logistic data=sus;
model r/n=neglect / SCALE=P;
run;
```

The output from SAS includes the deviance and Pearson scaling parameter estimates along with significance tests for them (these significance tests are not offered by the other packages). The output from SAS also offers a choice of the likelihood-ratio, score, and Wald significance tests for comparing this model against its null-effect alternative (the other packages are limited to the likelihood ratio and Wald tests).

Binomial MLMs

Binomial MLMs can be implemented in R, SAS, and Stata, among others. We recommend these three environments as the most transparent and user-friendly for this purpose. SAS and Stata are slightly more convenient because they permit “grouped” data in the form of scores for the controls and cases, whereas the lme4 package in R requires “unpacking” those scores into pseudo item-by-item data (instructions on how to do this are available online at: http://dl.dropbox.com/u/1857674/Binomial_GLM/binom.html).

We recommend using more than one package, if possible, to estimate binomial MLMs and to check convergence messages and indicators, to ensure that different packages and procedures produce similar results. For instance, comparisons among R, SAS, and Stata output for our examples will reveal minor numerical variations and differences in the parameters that are included in the output. The most similar outputs will be obtained by ensuring that the estimation methods used in the packages are as similar as possible. To this end, the examples provided online (MLMs in R, SAS, and Stata for Delayed Recall scores on Logical Memory, Table 2; and Testing for Dissociation, Table 4) all use the Laplacian approximation.