ON SEPARABLE TESTS, CORRELATED PRIORS AND PARADOXICAL RESULTS IN MULTIDIMENSIONAL ITEM RESPONSE THEORY

Giles Hooker

Abstract

This note considers the problem of paradoxical results in multidimensional item response theory. This is the possibility that a deliberately incorrect response could be beneficial to an examinee. An example is created that demonstrates that, when three or more ability dimensions are considered, paradoxical results can occur for Bayesian estimates even for separable tests when all the dimensions have positive prior correlation. A discussion of the mathematics behind the result is given and a computationally feasible means to check for the existence of paradoxical results is developed.

Key words: item response theory, multidimensional, posterior, paradoxical result, MAP, EAP

1. Introduction

A number of recent papers (*****, 2009b,a; Hooker et al., 2009) have considered the phenomenon labeled "paradoxical results" in multidimensional item response theory. This is the possibility that an examinee in a test may obtain a higher ability estimate in one ability dimension if one of their correct responses was changed to incorrect. This situation may be regarded as being unfair – it would be in the examinee's best interest to deliberately answer that question incorrectly – and thus it would be useful to avoid this possibility. This paper extends the results of those cited above by examining the practice of "borrowing strength" between ability dimensions by using correlated priors in Bayesian estimates. The term is borrowed from Wainer et al. (2001) in the context of empirical Bayes shrinkage of univariate estimates of ability. It was used explicitly in Segall (2000) in a fully Bayesian framework and ***** (2009b) employed it as a means of reducing paradoxical results in two dimensions. We demonstrate that this practice can have the perverse effect of inducing paradoxical results in seemingly reasonable scenarios when three or more ability dimensions are used.

Hooker et al. (2009) demonstrated that *separable* tests, in which every item measures only one of the ability dimensions, cannot produce paradoxical results if an examinee's ability is estimated via maximum likelihood. The same result holds if a Bayesian estimate such as the expected *a posteriori* estimate (EAP) is employed with a prior in which all the ability dimensions are independent. That paper further showed that, in two dimensions, there are fewer items that can produce paradoxical results when the prior correlation between dimensions is increased. It may therefore seem reasonable to suppose that the use of separable tests with positively-correlated traits can be guaranteed to avoid paradoxical results. This is always true for estimates in two-dimensional ability spaces. This note presents a numerical example of a three-dimensional separable test with positively correlated ability estimates in which MAP and EAP estimates produce a paradoxical result. It also provides a mathematical explanation for how this occurs and a means to check on whether paradoxical results are possible in a given test.

2. A Numerical Example

Our example is intended to demonstrate the existence of paradoxical results in this situation rather than to examine their prevalence in practice. It has therefore been constructed with an emphasis on simplicity rather than realism. We consider a test that measures 3 abilities: $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)$. Each item in the test follows the two-parameter logistic model for the probability of a correct response:

$$P(y_i = 1|\boldsymbol{\theta}) = \frac{1}{1 + \exp\left[-a_i(\theta_{d_i} - b_i)\right]}$$

where y_i is an indicator of a correct response to item i, θ_{d_i} is the ability parameter of the dimension measured by this item, and a_i and b_i are the discrimination and difficulty parameters for the item, respectively. Using the above parametrization, b_i is the value of θ_i at which a subject has probability 0.5 of giving a correct response. These parameters and a response sequence are given for an example seven-item test in Table 1. We assume a prior distribution

$$\boldsymbol{\theta} \sim N(0, \Sigma), \ \Sigma = \begin{pmatrix} 2.00 & 1.40 & 0.45 \\ 1.40 & 1.50 & 0.90 \\ 0.45 & 0.90 & 1.50 \end{pmatrix}$$

We also assume that that the a_i and b_i are given and fixed and that conditional on $\boldsymbol{\theta}$, all responses are independent.

=========================

Insert Table 1 about here

=========================

Maximum *a posteriori* estimates for $\boldsymbol{\theta}$ were obtained using a gradient-based method (the `optim` function in `R`). For the this test and response sequence, the estimated ability vector was (-0.0148, 0.0379, 0.2690). When the response for the final item was changed from correct to incorrect, the estimate became (0.0059, -0.0149, -0.1057). Here the estimates for θ_2 and θ_3 decreased as would be expected, but the estimate for θ_1 *increased.* Were the first dimension to be of primary interest, an examinee who had obtained the responses given for the first six items would benefit themselves by deliberately giving an incorrect response to item seven.

The same paradoxical result is observed for expected *a posteriori* estimates. These were obtained using a 25-point Gauss-Hermite quadrature rule in each dimension. EAP estimates avoid concerns over the convergence of estimates based on optimization. While the use of quadrature does introduce some bias, the estimates are exact for a distribution that places point masses at the quadrature points. For this test the EAP estimate was (-0.0205, 0.0502, 0.2993). When the response to the final item was changed from correct to incorrect, the estimate became (0.0101, -0.0248, -0.1479). Again, the estimate for θ_1 increased after changing a response from correct to incorrect.

3. Mathematical Explanation

We can explain the mathematical mechanisms creating this phenomenon through the theoretical results of Hooker et al. (2009). There it was observed that the phenomenon is attributable to the structure of item response functions creating negative posterior correlation between items. When this is the case, increasing the estimate of ability in one dimension – by giving a correct response to an item that loads heavily onto it – will tend to decrease the estimates in the other dimension. When separable tests are used and no prior correlation is given between items, each ability may be viewed as being estimated independently of the others and the posterior correlation between abilities remains zero. However, in three or more dimensions, positive prior correlation can have the perverse effect of inducing negative posterior correlation.

In order to make this precise we consider the more abstract setting of k ability dimensions with some positive-definite prior correlation matrix Σ with non-negative entries. We form the “design matrix” A_{-n} for a test of n items not including the final item such that A_{ij} specifies the discrimination of item i in dimension j. Due to the separability of the test, A will have zero entries apart from

$$[A_{-n}]_{id_i} = a_i.$$

That is, the ith row of A has only one non-zero entry: a_i appears in the column corresponding to the ability dimension, d_i being measured. We also let $W(\boldsymbol{\theta})$ be a diagonal matrix with diagonal entries

$$w_i(\boldsymbol{\theta}) = P(y_i = 1|\boldsymbol{\theta}) \left[1 - P(y_i = 1|\boldsymbol{\theta})\right]. \tag{1}$$

These are the second derivatives with respect to t of the log likelihood terms $\log[P(y_i = 1|a_i\theta_{d_i} = t)^{y_i} P(y_i = 0|a_i\theta_{d_i} = t)^{1-y_i}]$. In the common case of items following the logistic model the second derivative does not depend on y_i. For other models, this dependence complicates the means to check for paradoxical results below. For the sake of simplicity, we will focus on the final item which we assume to measure the d_nth ability; our estimates are clearly invariant to the item order and the ordering of the ability vector.

From a geometric perspective, we can consider the MAP for $\boldsymbol{\theta}_{-d_n}$ (the ability dimensions other than d_n) conditional on θ_{d_n}. Note that this conditional MAP will not change after the nth item since it loads only onto θ_{d_n}. When the final item is answered correctly, the MAP for θ_{d_n} must increase. If the conditional MAP for some θ_i decreases as θ_{d_n} increases its unconditional MAP will therefore be smaller after the additional correct response. The converse argument demonstrates that a larger MAP for θ_i will occur after an incorrect response.

In the case of compensatory models with log-concave item response functions as explored here, Hooker et al. (2009) provide the following sufficient condition for a MAP estimate to produce a paradoxical result. ***** (2009a) observed that a similar condition may also be given for when a MAP estimate can be guaranteed to avoid paradoxical results.

*Theorem 3.1. Hooker et al. (2009) Theorem 6.3 and ***** (2009a) Theorem 4.1*

A sufficient condition for the MAP estimate of θ_i to behave paradoxically when the final response is changed between correct and incorrect is that

$$\left[\left(A_{-n}^T W(\boldsymbol{\theta}) A_{-n} + \Sigma^{-1}\right)^{-1}\right]_{id_n} < 0 \tag{2}$$

for all $\boldsymbol{\theta}$.

A sufficient condition for the MAP estimate of θ_i not to behave paradoxically is that

$$\left[\left(A_{-n}^T W(\boldsymbol{\theta}) A_{-n} + \Sigma^{-1}\right)^{-1}\right]_{id_n} \geq 0 \tag{3}$$

for all $\boldsymbol{\theta}$.

***** (2009a) gave conditions of a similar form for paradoxical results to occur for EAP estimates, but these are somewhat more complex and will not be considered here.

The left hand side of (2) is the derivative of the MAP for θ_j conditional on θ_{d_n}. Since θ_{d_n} increases after a correct response, the MAP for θ_j decreases. The conditions as stated require the entries in this second derivative matrix to retain the same sign over all values of $\boldsymbol{\theta}$. This guarantees the presence (or absence) of a paradoxical result at the final item whatever the previous $n-1$ responses were. It is possible for (3) to be satisfied for some $\boldsymbol{\theta}$ and not others. So long as it is satisfied for the conditional MAP on all points between the estimates of θ_{d_n} with correct and incorrect responses, integrating between these two values demonstrates that a paradoxical result will still occur. However, if the entries change sign it is also possible that some response sequences yield paradoxical results while others do not. (3) is also conservative; its satisfaction guarantees that no paradoxical results can occur, but it may be the case that it does not hold for some test that still does not produce paradoxical results.

It is easy to see that for separable tests, $A_{-n}^T W(\boldsymbol{\theta}) A_{-n}$ is a diagonal matrix with positive entries. If Σ is also diagonal, so is the matrix on the left hand side of (3) and its off-diagonal elements are exactly zero. If $k=2$, it can be shown to be always positive. However it is not always the case that the entries of $(D+\Sigma^{-1})^{-1}$ are always positive for positive diagonal matrices D and positive definite matrices Σ with positive entries. Our test falls precisely into such a pairing.

4. Checking for Paradoxical Results

The results of Theorem 3.1 suggest a means of checking whether a paradoxical result is possible. If

$$\inf_{\boldsymbol{\theta}} \left[\left(A_{-j}^T W(\boldsymbol{\theta}) A_{-j} + \Sigma^{-1}\right)^{-1}\right]_{id_j} \geq 0, \forall j, \forall i \neq d_k \tag{4}$$

the theorem can be used to guarantee that the test will be free of paradoxical results. The condition can therefore be tested explicitly by removing each item k in turn and minimizing the left hand side of (4) for each i. This approach was taken in ***** (2009a) in the context of item bundles.

The context of separable tests leads to a very different approach, however. We define $\tilde{A}(\theta, j)$ to be the diagonal matrix $A_{-j}^T W(\boldsymbol{\theta}) A_{-j}$. Note that $\tilde{A}_{ii}(\theta, j)$ only depends on θ_i. We now observe that the standard adjunct formulation of a matrix inverse gives

$$\left[(\tilde{A}(\theta, j + \Sigma^{-1})^{-1}\right]_{id_j} = \frac{(-1)^{i+d_j}}{|\tilde{A}(\theta, j) + \Sigma^{-1}|} \left|[\tilde{A}(\theta, j) + \Sigma^{-1}]_{-i-d_j}\right|, \tag{5}$$

where C_{-i-d_j} represents the matrix C after deleting the ith row and d_jth column and $|C|$ is the determinant of C. Note that $|\tilde{A}(\theta, j) + \Sigma^{-1}| > 0$ so the sign of (5) is entirely determined by the (i, d_j)th cofactor of $\tilde{A}(\theta, j) + \Sigma^{-1}$. In particular, this co-factor does not incorporate $A_{d_j d_j}(\theta_{d_j}, j)$: the only entry that depends on the jth item and the only entry that depends on θ_{d_j}! It also does not incorporate $A_ii(\theta_j)$. Moreover, if one item loading onto d_k is found to violate (3) all other items that load onto that dimension will violate it also.

In the context of the example test above,

$$\Sigma^{-1} = \begin{pmatrix} 1.87 & -2.20 & 0.76 \\ -2.20 & 3.63 & -1.52 \\ 0.76 & -1.52 & 1.35 \end{pmatrix}$$

and

$$\left|[\tilde{A}(\theta, j) + \Sigma^{-1}]_{-i-d_j}\right| = 2.2 * 1.52 - 0.76\,(3.63 + A_{22}(\theta_2, j)) \tag{6}$$

and we see that whether an item that loads onto θ_3 causes θ_1 to behave paradoxically entirely depends on how much information is available about θ_2 as measured by

$$A_{22}(\theta_2) = \sum_{d_i = 2} a_i^2 w_i(\theta_2).$$

Formally, considering the minimum of (6) over θ_2 provides a conservative means of checking for paradoxical results; if the minimum is positive Theorem 3.1 guarantees that paradoxical results cannot occur. However, it is possible that (6) is positive at the MAP for θ_2 for all response sequences, only becoming non-zero in a short interval. In this case, Theorem 3.1 can be refined to demonstrate that no paradoxical results are possible; see Hooker et al. (2009). In the example test above, however, the case is exactly the opposite – all response vectors yield paradoxical results.

From these observations it is clear that checking whether an individual item can produce a paradoxical result will not be useful in designing tests that avoid them. In our example test, changing the response to every item loading onto θ_1 or θ_3 yielded paradoxical results, yet our analysis above demonstrates that it is those items that load onto θ_2 that are responsible for them.

In order to obtain a computationally efficient means of checking both whether a test can produce paradoxical results, and to select tests that are free of them, we restate these observations formally:

1. For all $\boldsymbol{\theta}$
$$\text{sgn}\left(\left[\left(A_{-j}^T W(\boldsymbol{\theta}) A_{-j} + \Sigma^{-1}\right)^{-1}\right]_{id_j}\right) = \text{sgn}\left((-1)^{i+d_j}\left|[A_{-j}^T W(\boldsymbol{\theta}) A_{-j} + \Sigma^{-1}]_{-i-d_j}\right|\right)$$
2. For any j
$$(-1)^{i+d_j}\left|[A_{-j}^T W(\boldsymbol{\theta}) A_{-j} + \Sigma^{-1}]_{-i-d_j}\right| = (-1)^{i+d_j}\left|[A^T W(\boldsymbol{\theta}) A + \Sigma^{-1}]_{-i-d_j}\right|$$
3. $[A^T W(\boldsymbol{\theta}) A]_{jj} = \sum_{d_i = j} a_i^2 w_i(\theta_j)$ can take any value in its range independently of the other entries in $A^T W(\boldsymbol{\theta}) A$. Typically, the minimum value is 0.

Combined, these observations state that checking (4) can be reduced to the following algorithm:

Algorithm 4.1. Checking for paradoxical results

1. For each i compute
$$w_i^+ = \max_{\theta_i} \sum_{d_j = i} a_j^2 w_j(\theta_i).$$

2. For each i and j test whether
$$\inf_{0 \leq w_i \leq w_i^+, i \in 1, \ldots, k} (-1)^{i+j} \left| [\text{diag}(w) + \Sigma^{-1}]_{-i-j} \right| > 0 \tag{7}$$

Note that while for items with a logistic response function $w_j(\theta_i)$ as given in (1) is independent of y_j, for other types of response functions, maximization may need to also be undertaken over the values of y_j. This can be done by simply choosing the y_j that maximizes $w_j(\theta_i)$ at each value of θ_i. Finding w_i^+ may also require optimizing a multi-modal criterion. However, since the optimizing value for each $w_j(\theta_i)$ is usually known and the problem is one-dimensional, a grid search will generally be sufficient.

In the case of a three-dimensional ability space, (4) can be reduced to a simple condition.

Theorem 4.1. Let $\Sigma^{ij} = [\Sigma^{-1}]_{ij}$. If
$$\Sigma^{s_1 s_3} \Sigma^{s_2 s_3} > \Sigma^{s_1 s_2} (\Sigma^{s_3 s_3} + w_{s_3}^+) \tag{8}$$
for all permutations s of the integers (1,2,3), then (4) is true.

Proof. (4) is equivalent to (7) being true for all distinct i and j. (7) is in turn equivalent to
$$\Sigma^{s_1 s_3} \Sigma^{s_2 s_3} > \max_{w_{s_3}} \Sigma^{s_1 s_2} (\Sigma^{s_3 s_3} + w_{s_3})$$
for $s_1 = i$ and $s_2 = j$. When $w_{s_3} = 0$, this condition must be true because
$$\text{sgn}(\Sigma_{s_1 s_2}) = \text{sgn}(\Sigma^{s_1 s_3} \Sigma^{s_2 s_3} - \Sigma^{s_1 s_2} \Sigma^{s_3 s_3}) > 0$$
by assumption. Hence if $\Sigma^{s_1 s_2} \leq 0$ (8) is always true for $w_{s_3} > 0$. If $\Sigma^{s_1 s_2} \geq 0$, the right hand side of (8) is maximized at $w_{s_3} = w_{s_3}^+$. □

Since w_i^+ can be arbitrarily increased by adding more items that load onto dimension i, it is also evident that if any off-diagonal element of Σ^{-1} is positive, there will be a test for which (4) does not hold, and hence in which a paradoxical result is possible. Correspondingly, no three-dimensional test will produce paradoxical results if all the off-diagonal elements of Σ^{-1} are non-positive. The latter condition can be be reduced to

Corollary 4.1. No maximum *a posteriori* estimate of $\boldsymbol{\theta}$ will behave paradoxically in a separable three-dimensional test if the prior correlations ρ_{ij} are all positive and
$$\rho_{s_1 s_2} \geq \rho_{s_1 s_3} \rho_{s_2 s_3}$$
for all permutations s of (1,2,3).

Proof. We require that the off-diagonal elements $\Sigma^{s_1 s_2}$ of the correlation matrix Σ be non-positive. Since

$$\mathrm{sgn}(\Sigma^{s_1 s_2}) = \mathrm{sgn}(\Sigma_{s_1 s_3}\Sigma_{s_2 s_3} - \Sigma_{s_1 s_2}\Sigma_{s_3 s_3})$$

we require

$$\Sigma_{s_3 s_3} \geq \frac{\Sigma_{s_1 s_3}\Sigma_{s_2 s_3}}{\Sigma_{s_1 s_2}} = \frac{\rho_{s_1 s_3}\rho_{s_2 s_3}}{\rho_{s_1 s_2}}\Sigma_{s_3 s_3}$$

which is equivalent to the condition in the corollary. □

Corollary 4.1 amounts to a condition that no pair of dimensions have much weaker correlation than the other pairs. In our example, the 0.45 correlation between dimensions 1 and 3 violated this condition.

This result motivates the following theorem for higher dimensions which examines the "exchangeable" case in which all pairs of dimensions have the same correlation. Here we intend 1_k to represent a column vector of length k with entries that are all 1.

Theorem 4.2. If Σ is of the form $(\sigma^2 - \alpha)I + 1_k \alpha 1_k^T$, no maximum *a posteriori* estimate of $\boldsymbol{\theta}$ will behave paradoxically in a separable test.

Proof. We observe that $\Sigma^{-1} = gI - 1_k e 1_k^T$ for $g = (\sigma^2 - \alpha)^{-1}$ and $e = \alpha/[(\sigma^2 - \alpha)(\sigma^2 + (k-1)\alpha)]$. Let $B(\boldsymbol{\theta}) = A^T W(\boldsymbol{\theta}) A + gI$. By the Woodbury identity, we are interested in the off-diagonal elements of

$$\left(B(\boldsymbol{\theta}) - 1_k e 1_k^T\right)^{-1} = B(\boldsymbol{\theta})^{-1} - B(\boldsymbol{\theta})^{-1} 1_k \left(-g + 1_k^T B(\boldsymbol{\theta})^{-1} 1_k\right)^{-1_k} 1_k^T B(\boldsymbol{\theta})^{-1}.$$

Since $B(\boldsymbol{\theta})$ is diagonal with positive entries, the off-diagonal elements will be positive so long as the term inside the brackets on the right hand side is negative. That is

$$-e + \sum_{i=1}^{k} \frac{1}{B_{ii}(\boldsymbol{\theta})} < 0.$$

Now, observing that since $B_{ii}(\boldsymbol{\theta}) > g$, the summand is bounded above by d/g and that $-e + d/g = -\alpha < 0$. □

While these conditions suggests a class of "safe" prior correlations, it would be philosophically difficult to justify choosing a prior based on the properties of the resulting estimator, rather than on the actual distribution of the quantities in question.

For specific tests, more general covariance structures may still be guaranteed to avoid paradoxical results, as in Theorem 4.1. In higher dimensional spaces, an exact formulation of (5) becomes algebraically intractable for general covariances. However, the following theorem demonstrates a tractable solution.

Theorem 4.3. The minimizing value of

$$H(w) = (-1)^{i+j}\left|[diag(w) + \Sigma^{-1}]_{-i-j}\right| \tag{9}$$

over w in the hyper-rectangle $\bigotimes_{j=1}^{k}[0, w_j^+]$ occurs at a vertex.

Proof. We observe that $H(w)$ can be written in the form

$$H(w) = \sum_S c_S \prod_{i \in S} w_i \tag{10}$$

for some constants c_S where the sum is taken over subsets S of the integers $1, \dots, k$. $H(w)$ is therefore linear in each w_i with the others held fixed.

Therefore, for any interior point w of $\bigotimes_{j=1}^{k}[0, w_j^+]$, $H(w)$ must be increased by changing w_1 to one of 0 or w_1^+. With this value of w_1 fixed, $H(w)$ is now a linear function of w_2 and hence is maximized for w_2 either 0 or w_2^+. Continuing, until w_k, we observe the result. □

Using Theorem 4.3 we can check (4) by simply testing the values 0 and w_l^+ for each dimension l in each condition i and j. Using the algorithm described in the proof, this requires $2k$ evaluations for each of the $k(k-1)/2$ pairs of ability dimensions. Since k is typically small, this remains a computationally easy task.

5. Selecting Regular Tests

If a test fails condition (4) it is reasonable to ask whether a small modification would produce a test that passed it. In order to find tests that avoid paradoxical results the following greedy strategy can be used.

Algorithm 5.1. Item selection in separable tests

Repeat until (4) is true

1. Let i^* and j^* be the indices giving the smallest minimum value of (9) with corresponding minimum w^*. Let l^* be the index of the maximum element in w^*.
2. For each item t that loads onto dimension l^*, let

$$r_t = \max_{\theta} \sum_{d_s = l^*,\ s \neq t} a_s^2 w_s(\theta_{l^*})$$

3. Remove the item with the smallest value of r_t.

Each step in the above algorithm is greedy; the first step tries to fix the largest violation of (4). The second step then looks for the dimension that most contributes to that violation, given heuristically by the maximum value in w^*; note that this will not be either of i^* or j^*. Finally, we seek the largest reduction possible in this maximum value. The algorithm attempts to retain large tests by only throwing out one item at a time.

It is worthwhile noting that Algorithm 5.1 takes no account of design criterion that may favor more items loading onto one dimension or with a particular spread of difficulty and discrimination parameters. It may also not result in the longest possible test that avoids paradoxical results. In the context of “borrowing strength” it will tend to make short tests even shorter. For short tests, it is, of course, feasible to check all possible subsets of the items and use a design criteria to select these. For very short tests it would also be possible to explicitly enumerate all possible response sequences and check for paradoxical results directly.

For the example test above, observe that Σ^{13} is the only negative entry in the off-diagonal elements of Σ^{-1} and (9) has minimum value -4.19. Step 1 therefore selects $i^* = 1$, $j^* = 3$ and the only remaining index is $l^* = 2$ and $w^* = 6.291$. For Step 2, there are only two items that can be removed, each of which reduces w^* to 4. However, with either one removed (9) still has minimum value -2.45. We must therefore remove both, at which point, (9) has minimum 0.584 and the test is regular, but no items now load onto θ_2.

6. Assessing Conservatism

Our test for paradoxical results is conservative in the sense that tests satisfying (4) are guaranteed to avoid them. However, an overly-conservative criterion will result in the needless rejection of tests that would not have produced them.

We begin by examining the test above. Geometrically, in order to fail (4) but nonetheless avoid paradoxical results, we require that at least one of the conditions be violated but that all violations occur in a region of ability-space in which MAP estimates do not fall for any response vector. Following Theorem 4.1 for our example, this will be the case if w_2^+ is much larger than the value of $w_2(\theta_2)$ for any MAP estimate θ_2.

Figure 1 presents an exploration of conservativism. The example test is used with only one item for dimension 2. The figure plots difficulty and discrimination parameter combinations for the item loading onto dimension 2 along with whether paradoxical results occur for the test and whether (4) rejects the test. A typical example of conservativism is also given; (4) needlessly rejects for large difficulty parameters when the maximal curvature of the item response falls outside the range of MAP estimates and where the discrimination parameter is still small enough that the inequality for (4) does not hold for parameter values at the MAP estimates.

The use of a single item implies a straight line at the 1.754 where (8) changes value. There is a thin band of unnecessary rejections for values of a_i up to 1.768 where (8) is violated in a small region in which few or no MAP estimates occur. The broader area of unnecessary rejections occur only for $|b_i| > 1.4$. By way of interpretation, items with this value of b_i are located so that subjects with probability 1/2 of giving a correct response are 1.16 prior standard deviations away from the prior mean. Note that the range of values for $a_i \in [1.6\ 2.2]$ is also very narrow and chosen to focus on regions where the conservativism of (4) becomes apparent.

==========================
Insert Figure 1 about here
==========================

While providing a geometric intuition, the use of a single item above is unrealistic and the geometry of conservative regions will be increasingly complicated for larger numbers of items. Moreover, item parameters for the other ability dimensions were held fixed. While these other items will not affect (8), they will affect the location of MAP estimates for dimension 2 and hence may change whether the test was needlessly rejected. In order to better assess how frequently needless rejected occurs, a simulation study was conducted using tests of length six with two items loading onto each dimension. Each a_i was randomly generated uniformly in the range [10^{-4}, 5] with the b_is chosen uniformly on the range [-5, 5]. The prior given for our example test was used. For each test (4) was evaluated, all possible response sequences were also generated and the existence of paradoxical results was then checked directly.

Of 100,000 simulated tests, 2,734 were rejected by (8) but did not produce paradoxical results, meaning the conservatism of the condition affected 2.75% of the volume of the 12-dimensional hypercube containing the test parameters. Nine tests were found to produce paradoxical results that were not detected by our methods. On closer examination, all of these were the result of convergence failures when obtaining the MAP: creating “false” paradoxical results. Seven were readily verified as being non-paradoxical by reducing the convergence tolerance. The remaining two each had at least one item with $|b_i| > 4.98$ and associated $a_i > 4.5$. Such item response curves are essentially flat in regions of large posterior probability and the estimates of ability vectors were strongly dependent on initialization values for the optimization routines.

Figure 2 plots a histogram of the maximum absolute difficulty of the items in each falsely rejected test that loaded onto θ_2. The figure also presents an analysis of conservativism as a function of the maximum allowable $|b_i|$. This was calculated by the number of needlessly rejected tests in the simulation with $|b_i| < c$

for each c normalized by $(c/5)^6$ – the ratio of the volumes of the hypercubes defined by $|b_i| < c$ and $|b_i| < 5$. Only 104 tests (or 0.1% of the simulation) had $|b_i| < 3$ for both items that load onto θ_2; meaning that for almost all erroneously rejected tests at least one item had a difficulty parameter outside the central 98% of prior mass and would be relatively uninformative about ability for the majority of the subject pool.

==========================
Insert Figure 2 about here
==========================

This study focuses on short tests because the combinatorial explosion in the number of possible response sequences renders a study of longer tests computationally infeasible. The use of two items on each dimension allows for some interaction between items; we argue that adding more items will tend to broaden the range of values for $\boldsymbol{\theta}$ that violate (4) making the criterion less conservative as tests become longer.

7. Conclusion

This paper presents a numerical example of a separable test with positive prior correlation between abilities that nonetheless produces paradoxical results. The existence of such results has important implications for attempts to "borrow strength" from known prior correlation between ability dimensions to improve the estimation accuracy of a test.

In addition to providing a numerical example, we have given a mathematical justification for the existence of these results and a means of verifying whether they can occur for a given test. Our results do not cover all popular methods; item response functions with guessing parameters are not covered by the theorems in Hooker et al. (2009), for example and conditions for the existence of paradoxical results for expected *a posteriori* estimates are more complex. While the use of correlation between dimensions will reduce the potential for paradoxical results in a non-separable bivariate tests, we expect that this will not always be the case for higher dimensions.

References

***** (2009a). Paradoxical results and item bundles. *Psychometrika*, in press.

***** (2009b). Prevalence and severity of paradoxical results in multidimensional item response theory. under review.

Hooker, G., M. Finkelman, and A. Schwartzman (2009). Paradoxical results in multidimensional item response theory. *Psychometrika 74*(3), 419–442.

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden and C. A. W. Glas (Eds.), *Comperterized Adaptive Testing: Theory and Practise*, pp. 53–73. Boston, MA: Kluwer Academic Publishers.

Wainer, H., J. Vevea, F. Camacho, B. Reeve, K. Rosa, L. Nelson, K. Swygert, and D. Thissen (2001). Augmented scores—"borrowing strength" to compute scores based on small numbers of items. In D. Thissen and H. Wainer (Eds.), *Test Scoring*, pp. 343–387. Mahwah, NJ: Lawrence Erlbaum Associates.

b_i	a_{i1}	a_{i2}	a_{i3}	y_i
-1/4	4	0	0	1
1/4	4	0	0	0
-1/4	0	4	0	1
1/4	0	4	0	0
-1/4	0	0	4	1
1/4	0	0	4	0
-1/4	0	0	4	1

Table 1.
Parameters and responses for a separable 3-dimensional test which produces paradoxical results under a positively correlated prior correlation when the final response is changed from 1 to 0.

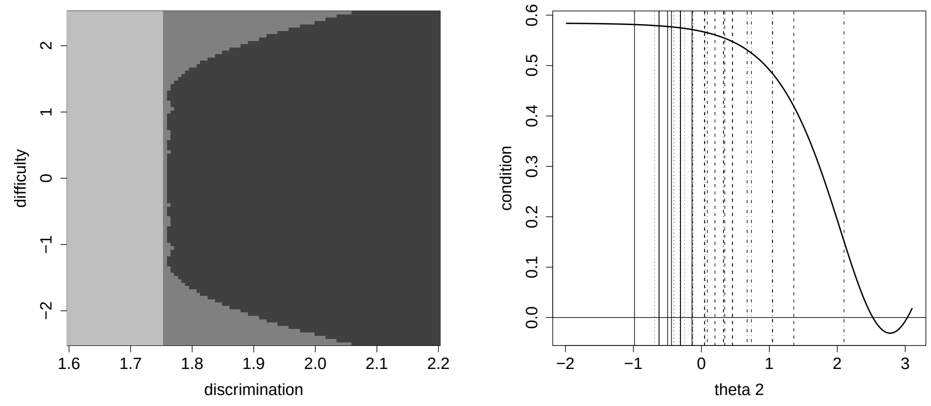

Figure 1.
An exploration of conservativism in testing for paradoxical results. Left: a plot of item difficulty and discrimination parameters indicating where paradoxical results occur (black), (4) is violated but paradoxical results do not occur (grey) and (4) is not violated and no paradoxical results occur (white). Right: an example of conservativism. The solid line gives $\Sigma^{12}\Sigma^{23} - \Sigma^{13}(\Sigma^{22} + A_{22}(\theta_2))$. (4) is violated if this is function is every less than zero: in a small region at the right of the plot. Vertical lines give MAP estimates for all possible response sequences. All these fall outside in a region where the violation occurs.

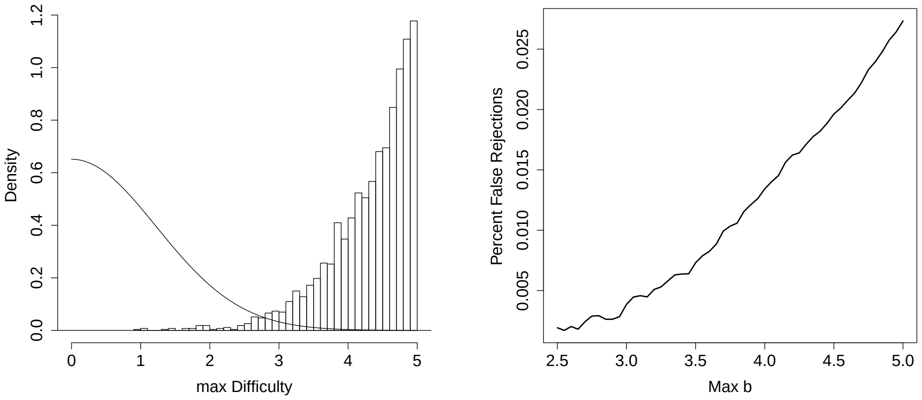

Figure 2.

Left: A histogram of $\max_{i_d=2} |b_i|$ for tests rejected unnecessarily. For comparison, the curve represents the prior density of $|\theta_2|$. Right: the proportion of tests needlessly rejected as the range of possible $|b_i|$ is increased.