Paradoxical Results and Item Bundles

Giles Hooker and Matthew Finkelman

March 12, 2009

Abstract

Hooker et al. (2009) defined a paradoxical result as the attainment of a higher test score by changing answers from correct to incorrect and demonstrated that such results are unavoidable for maximum likelihood estimates in multidimensional item response theory. The potential for these results to occur leads to the undesirable possibility of a subject's best answer being detrimental to them. This paper considers the existence of paradoxical results in tests composed of item bundles when compensatory models are used. We demonstrate that paradoxical results can occur when bundle effects are modeled as nuisance parameters for each subject. However, when these are modeled as random effects, or used in a Bayesian analysis, it is possible to design tests comprised of many short bundles that avoid paradoxical results and we provide an algorithm for doing so. We also examine alternative models for handling dependence between item bundles and show that using fixed dependency effects is always guaranteed to avoid paradoxical results.

1 Introduction

Finkelman et al. (2007) demonstrated an unexpected phenomenon: in multidimensional item response theory it is possible to increase a subject's estimated ability in one dimension by changing one of their answers from "correct" to "incorrect". They labeled this phenomenon a *paradoxical result.* Finkelman et al. (2007) and Hooker et al. (2009) explored the situation where two subjects (A and B) discover that A failed and B passed despite A giving correct answers to every question B answered correctly, as well as some that B answered incorrectly. The existence of this situation can be justified statistically, but can raise concerns about the perceived fairness of the scoring system. It may therefore be desirable to avoid such phenomena and a score (or ability estimate) that always does so is described as *regular*. Hooker et al. (2009) investigated the phenomenon mathematically and proved the surprising result that in compensatory models, paradoxical results occur for any non-separable test and almost all answer sequences within that test when maximum likelihood estimates are used to estimate abilities. Hooker et al. (2009) observed that the problem is ameliorated when priors are placed on abilities, but that paradoxical results are still possible, especially for long tests. However, that paper assumed a test in which items are independent conditional on the vector of abilities; the potential for paradoxical results to occur without this assumption is an open question.

An anonymous referee for Hooker et al. (2009) suggested that one frequent instance where conditional independence is violated is the use of tests that are comprised of a number of testlets or *bundles* that have some features in common. A canonical example is a test that is made up of a number of reading passages each of which has several comprehension questions. Bundles may also have particular contexts or assumed knowledge in common and may be detected statistically (e.g. Douglas et al., 1996). In such situations, the usual assumption that a subject's answers are locally independent becomes problematic (Rosenbaum, 1988).

One method of accounting for this dependence is to add a “nuisance ability” for each bundle that might capture the extent to which a subject was familiar with the context of a reading passage, for example. This device retains a form of local independence (conditional on the “bundle abilities”) at the cost of using a multivariate ability vector and the potential for paradoxical results.

Because paradoxical results can lead to student claims of unfairness, it is important for practitioners to know whether their tests are prone to producing them. At present, this is not possible if the test contains item bundles and the test is scored using methods that model dependence within a bundle (e.g. Li et al., 2006). The goal of this paper is to remedy the situation by examining whether paradoxical results can occur when employing models for item bundles. As will be seen, paradoxical results are possible when constructing tests haphazardly, but may be avoided when items are chosen judiciously. This paper examines the use of nuisance abilities to control for bundle effects. We restrict attention to a single dimension of interest and use compensatory models and, where appropriate, Gaussian priors for a subject’s abilities. Following Hooker et al. (2009), maximum likelihood estimates always produce paradoxical results in this situation. However, maximum likelihood estimation in these situations is rarely appropriate. Instead, “bundle abilities” can be treated as random effects, or within a Bayesian analysis (Rijmen et al., 2003).

We show that while paradoxical results can occur in the models under consideration, it is also possible to design tests made up of large numbers of short bundles for which Bayesian estimates of the ability of interest are regular. Given the potential for paradoxical results, the practitioner may desire to use this to ensure the perceived fairness of a test. We give conditions for checking when regularity can be guaranteed and develop algorithms for test building that incorporate this restriction. Empirically, it does not appear that restricting to such tests results in a large reduction in precision.

We assume a general linear structure in our models, rather than a restriction to the Rasch models in Wang et al. (2002) and Wang and Wilson (2005). Under the restrictions in those papers, ability estimates are always regular when the “bundle abilities” are independent, but may produce paradoxical results when they are positively correlated. The use of nuisance abilities is not the only means of controlling for within-bundle dependence. The dependence may also be modeled in a manner that is independent of the subject. This occurs in models using fixed dependency effects (e.g. Hoskens and de Boeck, 1997). These models can be shown to always produce regular tests.

The paper is structured as follows: we specify the model that we consider and the assumptions we use in Section 2. Section 3 derives conditions under which paradoxical results occur, and under which maximum likelihood with random effects, maximum *a posteriori* and expected *a posteriori* estimates are regular and shows that paradoxical results are inescapable for maximum likelihood estimates. Section 4 considers Bayesian estimates and random effects models using independence priors and derives algorithms to select tests for whether these are regular; Section 5 examines modifications when the bundle effects are positively correlated. Section 6 considers the application of our algorithms to test parameters estimated from real world data. Section 7 shows that estimates based on fixed dependency effects are always regular. We note that our analysis is confined to tests with a single dimension of interest with further factors used to control for bundle effects; non-separable tests in which there are two or more ability dimensions of interest may still be expected to produce paradoxical results.

2 Framework and Assumptions

We assume that we are interested in estimating a single dimension of ability θ in the presence of a test comprised of item bundles. Further, we assume that the dependence is modeled by assuming further individual “abilities” for each bundle. Our analysis may be applied to a number of estimates of θ and in this section we set out some notation and assumptions here that will be used throughout the paper. Firstly, we assume that the parameters of the test are given and defined with respect to fixed dimensions. We use the following assumptions and conventions:

A1 The test is comprised of N items. Responses to each item i are recorded as $y_i = 0$ (incorrect) or $y_i = 1$ (correct).

A2 We use θ for the subject's ability along the dimension of interest. $\gamma_1, \dots, \gamma_d$ are defined as factors corresponding to each of d item bundles. We write $\boldsymbol{\theta}$ for the concatenation $(\theta, \boldsymbol{\gamma})$. It is known that $\boldsymbol{\theta} \in \Omega \subseteq \mathbb{R}^{(d+1)}$.

A3 The probability of a correct response on item i given parameters $\boldsymbol{\theta}$ is

$$P(y_i = 1|\boldsymbol{\theta}) = g_i(\mathbf{a}_i^T \boldsymbol{\theta})$$

where $\log g_i(t)$ and $\log(1 - g_i(t))$ are both concave. This form includes commonly used compensatory models such as the multidimensional two-parameter logistic (Ackerman, 1996; Reckase, 1985) and normal ogive (Bock et al., 1988) models.

We write

$$l_i(t, y_i) = y_i \log g_i(t) + (1 - y_i) \log(1 - g_i(t))$$

and take $l_i'(t, y_i)$ and $l_i''(t, y_i)$ to be respectively the first and second derivative with respect to t.

A4 We define

$$A = [\mathbf{a}_1, \dots, \mathbf{a}_{N-1}]^T$$

to be the "design matrix" of the first $N - 1$ items on the test, and we let $\mathbf{b} = \mathbf{a}_N$ for notational convenience. Note that none of our estimates depends on the order of the items.

We use A_θ to designate the first column of A and $A_{\boldsymbol{\gamma}}$ the final d columns. Similarly, $b_{\boldsymbol{\gamma}}$ represents the "bundle loadings" for the final item and b_θ represents the loading of the final item onto the ability of interest.

A5 We assume that items are divided into d bundles with the mutually exclusive sets $B_1, \dots, B_d$ containing the indexes of the items in each bundle. We further assume that each row of $A_{\boldsymbol{\gamma}}$ only has one non-zero entry corresponding to its bundle. Without loss of generality, $b_{\boldsymbol{\gamma}}$ is zero except for b_{γ_d}.

A6 The set of responses to the first $N-1$ items in the test is contained in the vector $\mathbf{y}^{N-1} = (y_1, \dots, y_{N-1})$. We let $\mathbf{y}_0 = (\mathbf{y}^{N-1}, 0)$ and $\mathbf{y}_1 = (\mathbf{y}^{N-1}, 1)$. We use the notational convention that $\mathbf{y} \in \{0, 1\}^N$.

A7 The estimate of θ is $\hat{\theta}(\mathbf{y})$. This is one of a maximum likelihood estimate (MLE), a maximum likelihood estimate with $\boldsymbol{\gamma}$ modeled as being random (MR), a maximum *a posteriori* (MAP) estimate or a monotone functional of the marginal posterior distribution of θ such as the expected *a posteriori* estimate (EAP).

$\hat{\theta}(\mathbf{y})$ is *paradoxical* if $\hat{\theta}(\mathbf{y}_0) > \hat{\theta}(\mathbf{y}_1)$ for some $\mathbf{y}^{N-1}$ and some ordering of the items, otherwise it is *regular*.

A8 Any prior used for $\boldsymbol{\theta}$ is Normal, centered on zero, with covariance Σ. In the theorems below, we use the notation $K = -\Sigma^{-1}$.

A9 When the $\boldsymbol{\gamma}$ are modeled as random effects, these are modeled as being Normal, centered on zero with covariance matrix $\Sigma_{\boldsymbol{\gamma}}$ and we define

$$K = \begin{bmatrix} 0 & 0 \\ 0 & -\Sigma_{\boldsymbol{\gamma}}^{-1} \end{bmatrix}.$$

A10 When using maximum likelihood estimates, we assume that $\hat{\theta}(\mathbf{y}^{N-1})$, the estimate after the first $N-1$ items, is uniquely defined and use $K = 0$ below.

A11 When MLE or MAP estimates are used, we take $W(\boldsymbol{\theta}, \mathbf{y})$ to be the diagonal matrix with ith diagonal entry $l_i''(\mathbf{a}_i^T\boldsymbol{\theta})$. Note in the common case that $g_i(t)$ is a logistic function,

$$w_i(\boldsymbol{\theta}) = g_i(\mathbf{a}_i^T\boldsymbol{\theta})(1 - g_i(\mathbf{a}_i^T\boldsymbol{\theta}))$$

does not depend on y_i.

A12 When EAP or MR estimates are being used, we take $W(\boldsymbol{\theta}, \mathbf{y})$ to be the diagonal matrix with diagonal entries

$$w_i(\boldsymbol{\theta}, y_i) = E\left(l_i''(\mathbf{a}_i^T\boldsymbol{\theta}, y_i)|\theta, \mathbf{b}^T\boldsymbol{\theta}, \mathbf{y}\right) - \text{Var}\left(l_i'(\mathbf{a}_i^T\boldsymbol{\theta}, y_i)|\theta, \mathbf{b}^T\boldsymbol{\theta}, \mathbf{y}\right) < 0.$$

These are the conditional expectations with respect to the posterior distribution of $\boldsymbol{\theta}$, given θ and the linear combination $\mathbf{b}^T\boldsymbol{\theta}$.

We also take

$$K(\boldsymbol{\theta}) = K - \text{Cov}\left(K\boldsymbol{\theta}|\theta, \mathbf{b}^T\boldsymbol{\theta}\right).$$

In practice, these corrections will not be relevant for item-bundle models due to their structure when the priors, or the random effects distribution, treat the bundle effects as independent. We understand $K(\boldsymbol{\theta}) = K$ throughout when MAP or MLE estimates are used.

A13 We define $\mathcal{W} \subset \mathbb{R}^N$ to be the Cartesian product of the regions

$$\left[\min_{\boldsymbol{\theta}\in\Omega, y_i\in\{0,1\}} w_i(\boldsymbol{\theta}, y_i), \max_{\boldsymbol{\theta}\in\Omega, y_i\in\{0,1\}} w_i(\boldsymbol{\theta}, y_i)\right].$$

We use the notation $W \in \mathcal{W}$ to mean that W is a diagonal matrix whose's diagonal vector lies in $\mathcal{W}$.

We note that commonly-studied Rasch models (Wang et al., 2002; Wang and Wilson, 2005) are more restrictive than that given in A3. The models there are restricted, in our notation, to setting $a_{i\theta} = a_{i\gamma_j}$ for all i in the jth bundle. This restriction means that a maximum-likelihood estimate regarding both θ and $\boldsymbol{\gamma}$ as parameters of interest is not well-defined since A does not have full column rank. However, it is easy to see in our calculations below that MAP, EAP, and MR estimates, will yield regular estimates $\hat{\theta}(\mathbf{y})$ when Σ' is diagonal. Interestingly, if the elements of $\boldsymbol{\gamma}$ have positive correlation, it is possible to produce paradoxical results (see Section 5). Li et al. (2006) found in empirical studies that the more general model ($a_{i\theta} \neq a_{i\gamma_j}$) better represented a number of data sets and we assume such a model here.

3 A Regularity Criterion

We begin by extending a theorem in Hooker et al. (2009) to include tests with item bundles:

Theorem 3.1. *Under conditions A1-A13, a sufficient condition for* $\hat{\theta}_1(\mathbf{y}_1) < \hat{\theta}_1(\mathbf{y}_0)$ *is*

$$\mathbf{e}_\theta^T \left(A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A + K(\boldsymbol{\theta})\right)^{-1} \mathbf{b} > 0 \tag{1}$$

for all $\boldsymbol{\theta} \in \Omega$, $\mathbf{y} \in \{\mathbf{y}_1, \mathbf{y}_0\}$ *and* $\mathbf{e}_\theta = (1, 0, \ldots, 0)$ *the first Euclidean d-vector. A sufficient condition for* $\hat{\theta}_1(\mathbf{y}_1) > \hat{\theta}_1(\mathbf{y}_0)$ *is*

$$\mathbf{e}_\theta^T \left(A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A + K(\boldsymbol{\theta})\right)^{-1} \mathbf{b} < 0 \tag{2}$$

for all $\boldsymbol{\theta} \in \Omega$, $\mathbf{y} \in \{\mathbf{y}_1, \mathbf{y}_0\}$.

For MLE and MAP estimates, this is exactly Theorems 6.1 and 6.2 in Hooker et al. (2009), with the additional observation that the proof for regularity follows by an analogous argument. The extension to EAP estimates is given in Appendix A.

3.1 Maximum Likelihood Estimates

We begin by observing that Lemma 6.1 of Hooker et al. (2009) demonstrates that maximum likelihood estimates always produce paradoxical results. However, it is instructive to extend this analysis to the case of item bundles. We observe that

$$A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A = \begin{bmatrix} A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_\theta & A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_\theta \\ A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} & A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} \end{bmatrix}$$

and that the bottom right entry is diagonal. It now follows that

$$\mathbf{e}_\theta^T (A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A)^{-1}\mathbf{b} = \frac{1}{k}\left[b_\theta - A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} \left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}}\right)^{-1} b_{\boldsymbol{\gamma}}\right]$$

where

$$\begin{aligned}
k &= A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_\theta - A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} \left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}}\right)^{-1} A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_\theta \\
&= \sum_{i=1}^{N-1} a_{i\theta}^2 w_i(\boldsymbol{\theta}, y_i) - \sum_{j=1}^{d} \frac{\left[\sum_{i \in B_j} a_{i\theta} a_{i\gamma_j} w_i(\boldsymbol{\theta}, y_i)\right]^2}{\sum_{i \in B_j} a_{i\gamma_j}^2 w_i(\boldsymbol{\theta}, y_i)} \\
&\leq \sum_{i=1}^{N-1} a_{i\theta}^2 w_i(\boldsymbol{\theta}, y_i) - \sum_{j=1}^{d} \frac{\left[\sum_{i \in B_j} a_{i\theta}^2 w_i(\boldsymbol{\theta}, y_i)\right]\left[\sum_{i \in B_j} a_{i\gamma_j}^2 w_i(\boldsymbol{\theta}, y_i)\right]}{\sum_{i \in B_j} a_{i\gamma_j}^2 w_i(\boldsymbol{\theta}, y_i)} \\
&\leq 0
\end{aligned} \tag{3}$$

from the Cauchy-Schwartz inequality. We observe that only b_θ and b_{γ_d} are non-zero in $\mathbf{b}$ so that

$$\mathbf{e}_\theta^T (A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A + K)^{-1}\mathbf{b} = \frac{1}{k}\left[b_\theta - \frac{\sum_{i \in B_d} a_{i\theta} a_{i\gamma_d} w_i(\boldsymbol{\theta}, y_i)}{\sum_{i \in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta}, y_i)} b_{\gamma_d}\right].$$

It is now easy to show that choosing $\mathbf{b}$ to be the item *within the dth bundle* that gives least relative weight to θ produces a paradoxical result.

This result identifies a larger set of items that produce paradoxical results than the single item implied by Corollary 6.1 in Hooker et al. (2009). However, it also demonstrates that the criterion (2) can be separated into criteria for each bundle. We make extensive use of this below.

4 Random Effects Models and Bayesian Estimates with Independence Priors

In this section we consider using Bayesian estimates and random effects models in which the priors for the parameters are modeled as being jointly independent Gaussian random variables. For random effects models, only the bundle parameters are given priors. In this case we note that K is also diagonal and that

$$\mathbf{e}_\theta^T (A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A + K(\boldsymbol{\theta}))^{-1}\mathbf{b} = \frac{1}{k'}\left[b_\theta - A_\theta^T W(\boldsymbol{\theta},\mathbf{y}^{N-1})A_{\boldsymbol{\gamma}}\left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta},\mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} + K(\boldsymbol{\theta})_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\right)^{-1} b_{\boldsymbol{\gamma}}\right] \tag{4}$$

$$= \frac{1}{k'}\left[b_\theta - \frac{\sum_{i\in B_d} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta}, y_i)}{\sum_{i\in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta}, y_i) - \frac{1}{\sigma_{\gamma_d}^2}} b_{\gamma_d}\right] \tag{5}$$

for

$$\begin{aligned} k' &= \sum_{i=1}^{N-1} a_{i\theta}^2 w_i(\boldsymbol{\theta}, y_i) + K(\boldsymbol{\theta})_{11} \\ &\quad - A_\theta^T W(\boldsymbol{\theta},\mathbf{y}^{N-1})A_{\boldsymbol{\gamma}}\left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta},\mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} + K(\boldsymbol{\theta})_{\boldsymbol{\gamma},\boldsymbol{\gamma}}\right)^{-1} A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta},\mathbf{y}^{N-1})A_\theta \\ &< 0 \end{aligned}$$

by the same arguments as in (3) with $K(\boldsymbol{\theta})_{\boldsymbol{\gamma},\boldsymbol{\gamma}}$ the submatrix of $K(\boldsymbol{\theta})$ corresponding to $b_{\boldsymbol{\gamma}}$. We note that the correction to $K(\boldsymbol{\theta})$ in A12 only occurs in the factor $1/k'$, but defer this calculation to Appendix B. Similarly, the corrections to $w_i(\boldsymbol{\theta}, y_i)$ in A12 are also irrelevant in this case: since $l_i'(\boldsymbol{\theta}, y_i)$ only depends on θ and γ_d for $i \in B_d$, it has zero conditional variance.

While it is clear from the form of (5) that it is always possible to find a vector $\mathbf{b}$ so that a paradoxical result occurs, it is also possible to find tests such that no item will produce a paradoxical result. Hooker et al. (2009) observed this but commented that using priors to ensure regularity would ”restrict to either short tests, low discrimination parameters or strong priors.” In this case, we can think of item bundles as being multiple short tests. We show in the next section that it is possible to make use of (5) to provide a conservative criterion for testing whether a bundle can produce paradoxical results.

We note that when $a_{i\theta} = a_{i\gamma_d}$ for all $i \in B_d$, and hence $b_\theta = b_{\gamma_d}$, (5) is always positive since the ratio inside the brackets is always less than one. As noted earlier, this demonstrates that $\hat{\theta}$ is always regular under the Rasch models considered in Wang et al. (2002) and Wang and Wilson (2005).

4.1 A Selection Algorithm for Regular Bundles

The form of (5) demonstrates that when independence priors are used for parameters or random effects, (2) can be re-expressed only in terms of the bundle corresponding to the current item. Moreover, (5) may be bounded to provide a criterion by which to conclude ability estimates are regular. We make the following observations applied, without loss of generality, to the dth bundle.

Observation 4.1. *For all $j \in B_d$,*

$$\max_{\boldsymbol{\theta}\in\Omega,\mathbf{y}\in\{0,1\}^N} \frac{\sum_{i\in B_d, i\neq j} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta},y_i)}{\sum_{i\in B_d, i\neq j} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta},y_i) - \frac{1}{\sigma_{\gamma_d}^2}} \leq \max_{\boldsymbol{\theta}\in\Omega,\mathbf{y}\in\{0,1\}^N} \frac{\sum_{i\in B_d} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta},y_i)}{\sum_{i\in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta},y_i) - \frac{1}{\sigma_{\gamma_d}^2}}$$

and therefore

$$\frac{a_{j\theta}}{a_{j\gamma_d}} > \max_{\boldsymbol{\theta}\in\Omega,\mathbf{y}\in\{0,1\}^N} \frac{\sum_{i\in B_d} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta},y_i)}{\sum_{i\in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta},y_i) - \frac{1}{\sigma_{\gamma_d}^2}}$$

implies

$$\frac{a_{j\theta}}{a_{j\gamma_d}} > \frac{\sum_{i\in B_d, i\neq j} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta},y_i)}{\sum_{i\in B_d, i\neq j} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta},y_i) - \frac{1}{\sigma_{\gamma_d}^2}}, \ \forall \boldsymbol{\theta}\in\Omega, \mathbf{y}\in\{0,1\}^N.$$

That is, while (5) is a condition on an item that is not part of the items already in the test, including the item in the test only makes the condition more conservative. This conservative condition may reject items that would not produce paradoxical results, but does allow us to create an overall check on a bundle's regularity:

Observation 4.2. *Let*

$$i^0 = \operatorname*{argmin}_{i\in B_d} \frac{a_{i\theta}}{a_{i\gamma_d}}.$$

Then

$$\frac{a_{i^0\theta}}{a_{i^0\gamma_d}} > \max_{\boldsymbol{\theta}\in\Omega,\mathbf{y}\in\{0,1\}^N} \frac{\sum_{i\in B_d} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta},y_i)}{\sum_{i\in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta},y_i) - \frac{1}{\sigma_{\gamma_d}^2}} \tag{6}$$

implies

$$\frac{1}{k}\left[a_{j\theta} - \frac{\sum_{i\in B_d, i\neq j} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta},y_i)}{\sum_{i\in B_d, i\neq j} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta},y_i) - \frac{1}{\sigma_{\gamma_d}^2}} a_{j\gamma_d}\right] > 0, \ \forall \boldsymbol{\theta}\in\Omega, j\in B_d, \mathbf{y}\in\{0,1\}^N.$$

Observation 4.2 provides a means by which to check whether an item bundle can produce paradoxical results: find the item that places least relative weight on θ and check the condition (6). Checking this condition requires maximizing

$$F_{B_d}(\boldsymbol{\theta},\mathbf{y}) = \frac{\sum_{i\in B_d} a_{i\theta}a_{i\gamma_d}w_i(\boldsymbol{\theta},y_i)}{\sum_{i\in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta},y_i) - \frac{1}{\sigma_{\gamma_d}^2}}$$

over both $\boldsymbol{\theta}\in\Omega$ and $\mathbf{y}\in\{0,1\}^N$. This represents a mixed integer programming problem which can be computationally expensive. Evaluating $W(\boldsymbol{\theta},\mathbf{y})$ may also be difficult for EAP and MR estimates. As an alternative, we maximize over the weight value $\mathbf{w}$, ignoring the dependence of $\mathbf{w}$ on $\boldsymbol{\theta}$ and $\mathbf{y}$:

$$\max_{\boldsymbol{\theta}\in\Omega,\mathbf{y}\in\{0,1\}^N} F(\boldsymbol{\theta},\mathbf{y}) \leq \max_{\mathbf{w}\in\mathcal{W}} F_{B_d}(\mathbf{w}) = \max_{\mathbf{w}\in\mathcal{W}} \frac{\sum_{i\in B_d} a_{i\theta}a_{i\gamma_d}w_i}{\sum_{i\in B_d} a_{i\gamma_d}^2 w_i - \frac{1}{\sigma_{\gamma_d}^2}}. \tag{7}$$

This is now a linear fractional programming problem for which computationally efficient methods exist (e.g. Craven, 1988).

In the special case that $g_i(t)$ is a logistic function for each i and MAP estimates are used, we observe that $w_i(\boldsymbol{\theta})$ does not depend on y_i. In this case we only need to maximize $F_{B_d}(\boldsymbol{\theta}, \mathbf{y})$ over $\boldsymbol{\theta}$. Moreover, we note that for $i \in B_d$, $w_i(\boldsymbol{\theta}, y_i)$ only depends on θ and γ_d so that the maximization need only be undertaken in a two-dimensional space and may be accomplished efficiently by a number of iterative routines. Since using the left hand side of (7) represents a tighter bound in (6), we recommend doing so where possible.

If an item bundle cannot be shown to produce regular results, we may desire to find subsets of the bundle that can be. For small bundles, it is possible to enumerate all possible subsets and check (6). This will quickly become infeasible as the size of the bundle increases. In order to reduce the number of subsets that need to be considered we make the following observation:

Observation 4.3. *For $S \subset B_d$, let $i, j \in S$. Then*

$$\frac{a_{j\theta}}{a_{j\gamma_d}} < \max_{\boldsymbol{\theta} \in \Omega, y_i \in \{0,1\}} \frac{w_i(\boldsymbol{\theta}, y_i) a_{i\theta} a_{i\gamma_d}}{w_i(\boldsymbol{\theta}, y_i) a_{i\gamma_d}^2 - \frac{1}{\sigma_{\gamma_d}^2}} \tag{8}$$

implies

$$\frac{a_{j\theta}}{a_{j\gamma_d}} < \max_{\boldsymbol{\theta} \in \Omega, \mathbf{y} \in \{0,1\}^N} \frac{\sum_{i \in S} a_{i\theta} a_{i\gamma_d} w_i(\boldsymbol{\theta}, y_i)}{\sum_{i \in S} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta}, y_i) - \frac{1}{\sigma_{\gamma_d}^2}}.$$

This states that for a subset S containing item j, there is a checkable bound on which other items may be included in S while ensuring that paradoxical results cannot occur. Note that

$$\operatorname*{argmax}_{\boldsymbol{\theta} \in \Omega, y_i \in \{0,1\}} \frac{w_i(\boldsymbol{\theta}, y_i) a_{i\theta} a_{i\gamma_d}}{w_i(\boldsymbol{\theta}, y_i) a_{i\gamma_d}^2 - \frac{1}{\sigma_{\gamma_d}^2}} = \operatorname*{argmin}_{\boldsymbol{\theta} \in \Omega, y_i \in \{0,1\}} w_i(\boldsymbol{\theta}, y_i)$$

where the minimum is due to the w_i being negative. The corresponding $w_i(\boldsymbol{\theta}, y_i)$ is usually a known value w^* that is constant across i, avoiding the need for further optimization. We emphasize that there is not necessarily a maximal subset: the use of an item with low relative discrimination on θ precludes the use of items with high relative discrimination and *vice versa.*

Observation 4.3 allows us to produce initial candidate subsets: for each j we set S_j to be all the items with greater relative weight on θ than item j and that do not violate (8). We now need to check condition (6) for this subset. If this fails, some item must be removed from the subset[1]. To decide which, we look for the item whose removal most decreases the right hand side of (6). This may be done explicitly. Alternatively, a fast approximation is to consider a Taylor expansion. Letting $\mathbf{w}$ be the maximizing weights, removing item i from S_j is equivalent to setting $w_i = 0$ and we label the resulting weight vector $\mathbf{w}_0$, then

$$F_{S_j}(\mathbf{w}_0) \approx F_{S_j}(\mathbf{w}) + w_i \frac{dF_{S_j}}{dw_i}$$

and we note that the numerator of

$$w_i \frac{dF_{S_j}}{dw_i} = \frac{\sum_{k \in S_j} w_i w_k a_{i\gamma_d} a_{k\gamma_d} (a_{i\theta} a_{k\gamma_d} - a_{k\theta} a_{i\gamma_d}) - \frac{w_i a_{i\theta} a_{i\gamma_d}}{\sigma_{\gamma_d}^2}}{\left(\sum_{k \in S_j} a_{k\gamma_d}^2 w_k(\boldsymbol{\theta}, y_k) - \frac{1}{\sigma_{\gamma_d}^2} \right)^2}$$

[1] We do not check all possible subsets of the current subset for the same computational reasons that we do not check all subsets in the first place.

is particularly cheap to calculate. The minimizer of this quantity could then be employed as an alternative. Either criterion may be usually expected to select one of the items with largest relative weight on θ unless some other item has high discrimination on both abilities.

Following the reasoning above, we can create an algorithm that produces a sequence of test subsets S_j, $j \in B_d$ for which estimates of θ are regular. These S_j will be referred to as the *regular subsets* of B_d.

Algorithm 4.1. *Within-Bundle Selection*

1. *Set*

$$r_i = \frac{a_{i\theta}}{a_{i\gamma_d}}, \; c_i = \max_{\theta \in \Omega, y_i \in \{0,1\}} \frac{w_i(\boldsymbol{\theta}, y_i) a_{i\theta} a_{i\gamma_d}}{w_i(\boldsymbol{\theta}, y_i) a_{i\gamma_d}^2 - \frac{1}{\sigma_{\gamma_d}^2}}, \; i \in B_d$$

2. *Loop over $j \in B_d$:*

    (a) *Initialize $S_j = \{i \in B_d : r_i \geq r_j, c_i < r_j\}$.*

    (b) *While*

$$r_j \leq \max_{\mathbf{w} \in \mathcal{W}} \frac{\sum_{i \in S_j} a_{i\theta} a_{i\gamma_d} w_i}{\sum_{i \in S_j} a_{i\gamma_d}^2 w_i - \frac{1}{\sigma_{\gamma_d}^2}} \tag{9}$$

    i. *Set $\{w_i^*\}_{i \in S_j}$ to be the optimal weights from (9).*

    ii. *Choose either*

$$k^* = \operatorname*{argmax}_{k \in S_j, k \neq j} \left[\frac{\sum_{i \in S_j} a_{i\theta} a_{i\gamma_d} w_i^*}{\sum_{i \in S_j} a_{i\gamma_d}^2 w_i^* - \frac{1}{\sigma_{\gamma_d}^2}} - \max_{\mathbf{w} \in \mathcal{W}} \frac{\sum_{i \in S_j, i \neq k} a_{i\theta} a_{i\gamma_d} w_i}{\sum_{i \in S_j, i \neq k} a_{i\gamma_d}^2 w_i - \frac{1}{\sigma_{\gamma_d}^2}} \right] \tag{10}$$

    or

$$k^* = \operatorname*{argmin}_{k \in S_j, k \neq j} \sum_{i \in S_j} w_k^* w_i^* a_{k\gamma_d} a_{i\gamma_d} (a_{k\theta} a_{i\gamma_d} - a_{i\theta} a_{k\gamma_d}) - \frac{w_k^* a_{k\theta} a_{k\gamma_d}}{\sigma_{\gamma_d}^2}. \tag{11}$$

    iii. *Remove k^* from S_j*

3. *Return S_j, $j \in B_d$.*

Note that we have elected not to remove item j in Step 2(b)ii should it provide optimal decrease. This retains the association of S_j with item j and prevents duplicate subsets being returned. Following the use of Algorithm 4.1, we may obtain a large number of regular subsets; some of these may be redundant in the sense that all their elements may be contained in some other subset. It is easy, as a post-processing step, to remove subset S_j from the collection of regular subsets if $S_j \subseteq S_i$ for some i. We label the remaining subsets the *maximal regular subsets* of the bundle.

In order to tighten the bounds implied by (6) when g_i is a member of the logistic family and MAP estimates are used, we replace the maximization over $\mathbf{w}$ in Steps 2b and 2(b)ii by setting $w_i(\boldsymbol{\theta}) = g_i(\mathbf{a}_i^T \boldsymbol{\theta})(1 - g_i(\mathbf{a}_i^T \boldsymbol{\theta}))$ and maximizing over θ and γ_d.

We have noted that only a narrow range of relative loadings on θ can be used, although where that range begins may vary. This does not necessarily compromise the discriminating ability of an overall test, however: each bundle may contain different relative loadings without compromising the regularity of $\hat{\theta}$. We explore modifications of selection algorithms that incorporate this algorithm in the next section.

4.2 Selecting Regular Tests

The observations above provide a means of checking whether a given test and ability estimate is capable of producing paradoxical results and we have developed algorithms to find subsets of items within bundles that will not do so. We now wish to combine these with a standard item selection algorithm to develop tests that produce regular ability estimates. We do not consider the problem selecting optimal tests here, but note that test construction algorithms typically maximize some criteria such as the trace or determinant of the expected Fisher information subject to practical constraints like content balance (see e.g. Veldkamp, 2002). Such algorithms can (in theory) readily be modified to consider *groups* of items as discussed below.

We assume that there are d bundles in an item pool with $N_1, \ldots, N_d$ items in each. We assume that the candidate pool of items and bundles is much larger than the prescribed number to be selected for the final test. A first method of applying *any* item selection criterion is to

1. Find all regular subsets of each bundle.
2. Consider all ways to create a sub-pool of items using one regular subset from each bundle. Perform item selection on each sub-pool.
3. Choose the best-performing test from the set of sub-pools.

This method requires performing item selection on all $\prod_{i=1}^{d} N_i$ ways of combining regular subsets of the bundles. This is clearly computationally infeasible, although some speedup may be attained by removing clearly sub-optimal subsets. As an alternative, we propose selecting regular subsets directly:

Algorithm 4.2. *Test Selection by Regular Subsets*

1. *For each bundle, form its regular subsets* $S_{i1}, \ldots, S_{iN_i}$*.*
2. *If bundles in the selected test must contain no more than* M *items and some* S_{ij} *has cardinality greater than* M*, perform item selection within* S_{ij}*.*
3. *Select* regular subsets *by whatever item selection criterion is being used. If a subset from bundle* j *is selected, remove all other subsets from that bundle from the pool of regular subsets.*
4. *If the resulting test is too long, perform item selection on the items in the test.*

This algorithm requires that an item selection method can be extended to a selection method for regular subsets, but we do not expect this to present problems. We note that the similar modifications could be made to computer-adaptive testing procedures that use item bundles.

The extent of the loss of efficiency due to a restriction to regular tests is not clear. It will only be severe when the optimal unrestricted test uses items with a wide range of relative weight on θ within each bundle. This may occur when there are relatively few potential items in each bundle pool and these have wide ranges of relative loadings. If this is the case, a decision must be made between the loss of efficiency and the production of a paradoxical result.

We also note that the item selection algorithms proposed in this section are conservative. They are based on conditions (2) and (5), neither of which is tight. There are (theoretical) tests with regular $\hat{\theta}$ for EAP, MAP or MR estimates that violate one or both conditions and this may include tests from the item pool being considered. Assessing the regularity of a proposed test directly would require checking $\hat{\theta}(\mathbf{y})$ for all possible response values, which is clearly infeasible. However, a Monte Carlo simulation would at least assess the probability of observing a paradoxical result; this may be considered if the tests produced by the algorithms above produce tests that are significantly less discriminatory than optimal.

5 Dependent Bundle Effects

Our discussion so far has been made solely in terms of population distributions for bundle effects in which the effects are taken as being all uncorrelated. This may be unrealistic in the real world. Hooker et al. (2009) observed that in two-dimensional models adding correlation between the dimensions means that (6) reduces the set of item parameters that can cause paradoxical results. In general, the analysis of regularity when bundle effects are allowed to be correlated becomes considerably more difficult analytically. We therefore restrict our attention to MAP estimates throughout this section and disregard the correction terms in A12.

While positive correlation between abilities improves regularity in two dimensions, in higher dimensions the situation is somewhat more complex, and can be counter-intuitive. In fact, adding small correlation between nuisance abilities makes (6) *more* restrictive:

Observation 5.1. *Under conditions A1 - A13, assume that*

$$\Sigma = \left[\begin{array}{cc} \sigma_{11}^2 & 0 \\ 0 & \Sigma_{\boldsymbol{\gamma}} \end{array} \right]$$

and let $\tilde{\Sigma}$ be diagonal with the same diagonal elements as Σ. If $\sigma_{\gamma_i,\gamma_j} < \epsilon$ for all i, j and $\epsilon > 0$ sufficently small, then

$$A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1}) A_{\boldsymbol{\gamma}} \left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1}) A_{\boldsymbol{\gamma}} - \Sigma_{\boldsymbol{\gamma}}^{-1} \right)^{-1} \mathbf{e}_{\gamma_d} \tag{12}$$

is bounded below by

$$A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1}) A_{\boldsymbol{\gamma}} \left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1}) A_{\boldsymbol{\gamma}} - \tilde{\Sigma}'^{-1} \right)^{-1} \mathbf{e}_{\gamma_d}.$$

Recall that we must have b_θ / b_{γ_d} greater than the left hand side of (12) in order to conclude that an item does not produce paradoxical results. This result is based on the derivative of (12) being negative and the calculations are given in Appendix C.

There is an intuitive explanation for this phenomenon. We think of paradoxical results being caused by a subject getting an item correct that loads heavily onto a nuisance ability. This increases the estimate of the nuisance ability, forcing the estimate of the ability of interest to decrease in order to explain previously wrong answers. The use of a prior regularizes the nuisance ability estimate, reducing how far it can increase. This is how we are able to produce regular tests, even though maximum likelihood estimates always admit paradoxical results. When a prior models the nuisance abilities as being positively correlated, it provides less regulatory effect. Moreover, increasing the estimate of one nuisance ability will also increase the estimates of the other nuisance abilities, further depressing the estimate of the ability of interest.

The fact that increasing the estimate of the γ_d can result in the estimates for other nuisance abilities increasing complicates an attempt to provide a general method for selecting regular tests that does not simply brute-force check the condition (6) at each step. For specific prior covariance structures, however, bounds can be given that allow a modification of the algorithms in Sections 4.1 and 4.2. In particular, we consider an exchangeable covariance structure for the bundle abilities:

$$\sigma_{\gamma_i \gamma_j} = \left\{ \begin{array}{ll} \sigma^2 & i = j \\ \alpha & \text{otherwise.} \end{array} \right.$$

In other words $\Sigma_{\boldsymbol{\gamma}} = (\sigma^2 - \alpha) I + \alpha J$ where J is a matrix of all ones. For this prior structure, the following bound is derived in Appendix D:

Theorem 5.1. *Let $\Sigma = (\sigma^2 - \alpha)I + \alpha J$, and suppose that $w_i(\boldsymbol{\theta}, y_i) > w_i^+$. Then*

$$A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1}) A_{\boldsymbol{\gamma}} \left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1}) A_{\boldsymbol{\gamma}} - \Sigma_{\boldsymbol{\gamma}}^{-1} \right)^{-1} \mathbf{e}_{\gamma_d}$$

is bounded above by

$$\frac{\sum_{i \in B_d} a_{i\theta} a_{i\gamma_d} w_i(\boldsymbol{\theta}, y_i) - \frac{\alpha}{(\sigma^2-\alpha)^2} \sum_{j=1}^d \max \left(\frac{\sum_{i \in B_j} w_i a_{i\theta} a_{i\gamma_j}}{\sum_{i \in B_j} a_{i\gamma_j}^2 w_i(\boldsymbol{\theta}, y_i) - \frac{1}{\sigma^2 - \alpha}} \right)}{\sum_{i \in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta}, y_i) - \frac{1}{\sigma^2-\alpha}}. \tag{13}$$

This bound now provides us with an ability to restrict attention to item bundles, but with a correction to Step 9 in Algorithm 4.1:

$$r_j \leq \max_{\mathbf{w} \in \mathcal{W}} \frac{\sum_{i \in B_d} a_{i\theta} a_{i\gamma_d} w_i(\boldsymbol{\theta}, y_i) - \frac{\alpha}{(\sigma^2-\alpha)^2} \sum_{j=1}^d \max \left(\frac{\sum_{i \in B_j} w_i a_{i\theta} a_{i\gamma_j}}{\sum_{i \in B_j} a_{i\gamma_j}^2 w_i(\boldsymbol{\theta}, y_i) - \frac{1}{\sigma^2 - \alpha}} \right)}{\sum_{i \in B_d} a_{i\gamma_d}^2 w_i(\boldsymbol{\theta}, y_i) - \frac{1}{\sigma^2-\alpha}}.. \tag{14}$$

This correction will, of course, depend on the items selected in the previous bundles. This can be managed with an iterative algorithm:

Algorithm 5.1. *1. Initialize a set $S_{i1}, \ldots, S_{iN_i}$ of regular subsets for each bundle B_i.*

2. Select a test using the bundles and candidate regular subsets as in Algorithm 4.2.

3. Do until the test does not change:

(a) Loop over $i \in \{1, \ldots, d-1\}$:

i. Re-calculate the regular subsets of B_i.

ii. Remove the current subset of B_i from the test and replace it with the optimal regular subset.

iii. If B_i did not contribute to the current test, test whether replacing a subset from another bundle with the optimal subset from B_i improves the test. If so, do so.

Step 1 above is left undetermined. It could be initialized using Algorithm 4.1 as is, or modifying it by (14) applied using all the items not in the current bundle. We suspect that the latter will often produce too few, or too small, regular subsets, if any, to be useful. Further alternatives would include slowly increasing f to stabilize the selection procedure.

As an interesting observation, in the case of the Rasch models in Wang and Wilson (2005), we have that $a_{i\theta} = a_{i\gamma_d}$ and the (θ, γ_d)th and (γ_d, γ_d)th entries of $A^T W(\boldsymbol{\theta}, \mathbf{y}) A$ are both $\sum w_i(\boldsymbol{\theta}, y_i) a_{i\theta}^2$. Since $b_\theta = b_{\gamma_d}$, guarantees against paradoxical results require that (12) is less than 1 for all $w_i(\boldsymbol{\theta}, y_i)$. For an exchangeable covariance structure, substituting into (13) we observe that so long as $\alpha > 0$ our bound will always be violated if the number of bundles is sufficiently large.

6 Real-World Item Bundles

In order to examine the qualitative properties of the within-bundle selection algorithm above, we make use of the test parameters estimated in Li et al. (2006), Table 5. These parameters were estimated from the analytical reasoning section of the LSAT administered in 1992. We are not concerned with the estimation of item parameters here, so we treat the estimated parameters as fixed. Li et al. (2006) found that they obtained the best fit for observed data with their Model 1: placing no restrictions on bundle discriminations apart from positivity and setting Σ to be the identity matrix. They give estimated parameters for 24 items grouped into 4 bundles, using a probit link.

Our analysis was carried out in the `R` statistical programming language. The maximization (6) was undertaken using a Nelder-Meade simplex algorithm implemented in the `optim` routine when optimization was undertaken over $\boldsymbol{\theta}$. The optimization (7) was solved using the linear programming library `lpSolve`, converting (7) to the equivalent linear program (Craven, 1988):

$$(\mathbf{v}, t) = \text{argmax} \sum_{i \in B_d} a_{i\theta} a_{i\gamma_d} v_i \text{ subject to } \sum_{i \in B_d} \sigma^2_{\gamma_d} a^2_{i\gamma_d} v_i + t = 1,\ t \geq 0,\ 0 \leq \mathbf{v} \leq -tw^+$$

and converting $\mathbf{w} = -\mathbf{v}/t$ to find the optimizing weights. Code implementing Algorithm 4.1 and the analysis below is available from the first author.

In carrying out this analysis, we wanted to investigate both the real-world properties of the selection procedure and the relative performance of the various approximations used. As a first note, we found that the optimization (10) was never so computationally expensive that (11) proved appreciably faster, although that may prove to be the case for very large bundles. Our analysis below therefore solely makes use of (10). In general, we take the view that larger regular subsets are better and have therefore only considered maximal regular subsets, removing S_i if it is contained in some S_j. This is also the objective we use for comparing our various methods.

6.1 Small-Bundle Analysis

We applied our methods using (7) and the bounds $w_i \in [-1, 0]$, which are appropriate for the probit link function. When an independence prior was assumed, our algorithms suggested minor changes to the reported test. Both bundles 2 and 3 were left unchanged, while two subsets consisting of 5 out of 6 items were found for bundle 1, and subsets of size 5 and 6 out of 7 items were found for bundle 4. Testing all subsets within a bundle produced further regular subsets; however, these were not, in general, larger than those found by Algorithm 4.1. The items selected by both methods are detailed in Table 1.

In order to examine the effect of adding positively correlated items, we considered the same test but incorporated a correlation of 0.2 between each pair of bundle abilities and used the correction (13). This resulted in a correction factor $f = 0.31$. The maximal subsets found by Algorithm 4.1 were smaller and more numerous in all bundles than those found with zero correlation, although they were still among the largest of those found by testing all possible subsets. These are also reported in Table 1. As the correlation was increased, the size of the regular subsets reduced: at correlation 0.25, only singleton sets were given for bundle 4; by 0.5 no subsets were regular for bundle 1.

6.2 Larger Bundles

None of the bundles considered above had more than 7 items. In order to investigate within-bundle selection on a large bundle set we made use of the estimated parameters from Li et al. (2006), Table 9. These parameters had been estimated from data taken from reading comprehension subtests on a college-level EPT administered in 2000. In particular, we made use of the "Literal Meaning" bundle which contained 16 items, and undertook the same estimation procedure as above. From these 16 items, Algorithm 4.1 found

Bundle	All Subsets No Correlation	Algorithm 4.1 No Correlation	Algorithm 4.1 Correlation 0.05
1	2 3 4 5 6 1 2 4 5 6 1 2 3 4 6 1 3 4 5	2 3 4 5 6 1 2 4 5 6	1 2 5 3 4 6
2	7 8 9 10 11	7 8 9 10 11	7 8 9 11 7 8 10 11
3	12 13 14 15 16 17	12 13 14 15 16 17	12 13 16 17 13 14 16 17 13 15 16 17
4	18 19 21 22 23 24 18 19 20 22 23 18 20 21 22 23 24 19 20 21 22 23 24 18 19 20 21 23 18 19 20 21 22	18 19 21 22 23 24 18 19 20 22 23	18 19 18 20 23 24

Table 1: Maximal regular subsets of bundles in the analytical reasoning section of the LSAT administered in 1992, using parameters estimated in Li et al. (2006). First column: subsets chosen by exhaustive search. Second column: subsets chosen by Algorithm 4.1. Third column: subsets chosen by Algorithm 4.1 with correction (13) when an exchangeable covariance is assumed between bundle factors with correlation 0.05.

four maximal regular subsets of size 11, 13, 9 and 13 which are represented graphically in the left panel of Figure 1. An exhaustive search over all 2^{16} possible subsets in this instance found exactly the same collection of maximal regular subsets but required substantially more processing time (246.42 seconds as opposed to 0.75 seconds).

6.3 Checks of Conservativism

In order to investigate the effect of maximizing over $\mathbf{w}$ as opposed to $\boldsymbol{\theta}$ in Step 2b of Algorithm 4.1, we employed a logit model with the EPT subset above, making the usual correction of multiplying all item parameters by 1.7. Under a logit model using MAP estimates, we observe that

$$w_i(\boldsymbol{\theta}, y_i) = P_i(\mathbf{a}_i^T\boldsymbol{\theta})(1 - P_i(\mathbf{a}_i^T\boldsymbol{\theta})) \in [-1/4,\ 0]$$

where $P_i(t) = \text{logit}(t - d_i)$ and d_i is the difficulty parameter for the item. Difficulty parameters for this test were not provided in Li et al. (2006), and we instead created hypothetical parameters by making use of the first 16 difficulty parameters from the LSAT test above.

We begin by noting that the maximal curvature of the logit model (-0.25) is substantially smaller than that of the probit (-1), and this difference is not entirely compensated for by the 1.7 correction factor. Because of this, the use of a logit model in the LSAT test found all four bundles to be regular without using subsets. The larger EPT test bundle, however, could not all be included. In this case both maximization over $\boldsymbol{\theta}$ and $\mathbf{w}$ returned the same subsets (sizes 12, 14 and 11).

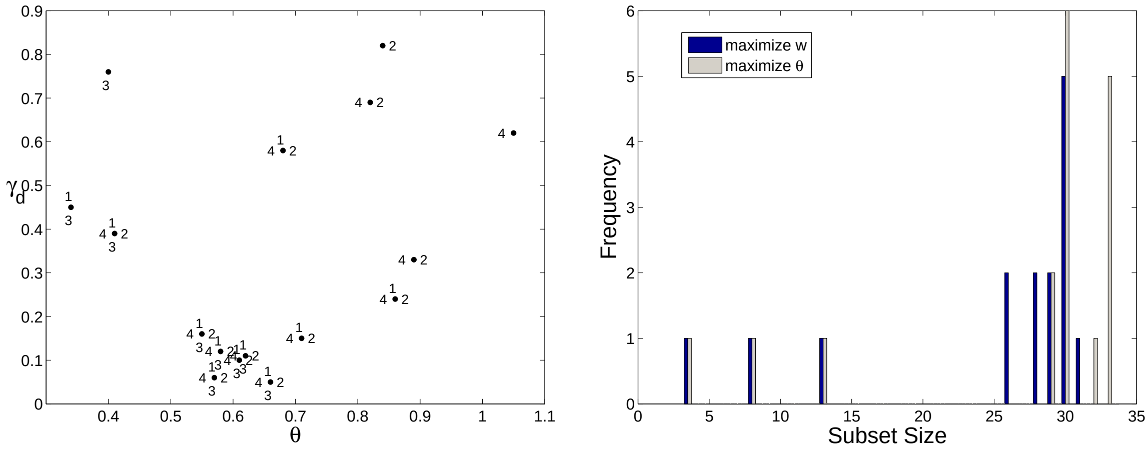

Figure 1: Left: A graphical representation of the maximal regular subsets found by Algorithm 4.1. Items are plotted on the (θ, γ_d) plane and labeled with the regular subset in which they appear. The bundles are approximately spread radially. Right: Sizes of maximal regular subsets of test of length 50 with difficulty parameters generated randomly on the unit square. Dark: selection maximizing over $\mathbf{w}$, light: maximizing over $\boldsymbol{\theta}$.

The difference between maximizing over $\boldsymbol{\theta}$ and over $\mathbf{w}$ will be largest when the difficulty parameters within the bundle are relatively spread out so that $W(\boldsymbol{\theta}, \mathbf{y}^{N-1})$ is not able to mimic the maximizing values found by (7). In order to test this difference further, we simulated a 50-item bundle by taking the discrimination parameters to be uniform on the unit square and the difficulty parameters to have a standard normal distribution. Qualitatively, we found that while the maximal regular subsets found by optimizing over $\boldsymbol{\theta}$ did tend to be larger than those found by optimizing over $\mathbf{w}$, the difference was typically only a few items. This is presented graphically in the right panel of Figure 1. We are therefore confident that the maximization over $\mathbf{w}$ does not represent an overly conservative approximation from a practical point of view.

7 One-Factor Models with Dependence

An alternative means of accounting for dependence within item bundles involves modeling it explicitly in a manner that is constant across subjects. This can be done by fixed dependency effects (Kelderman, 1984; Hoskens and de Boeck, 1997). Alternatively, a quasi-likelihood approach using generalized estimating equations (e.g. McCullagh and Nelder, 1989, ch. 9) could be used. In either case, only θ is estimated, and it is easy to find and verify conditions under which $\hat{\theta}$ is regular.

We begin by expressing the estimate for θ as a solution to a nonlinear equation:

$$\hat{\theta}(\mathbf{y}) = \{\theta : S(\theta, \mathbf{y}) = 0\} \tag{15}$$

where $S(\theta, \mathbf{y})$ is a score function that defines a number of estimates:

maximum likelihood: $S(\theta, \mathbf{y}) = dl(\theta, \mathbf{y})/d\theta$ where $l(\theta, \mathbf{y})$ is the log likelihood of θ given the observed response.

maximum *a posteriori*: $S(\theta, \mathbf{y}) = df(\theta, \mathbf{y})/d\theta$ where $f(\theta, \mathbf{y})$ is the log posterior of θ.

quasi-likelihood: does not require a probabilistic specification, but sets

$$S(\theta, \mathbf{y}) = \frac{d\mu(\theta)}{d\theta}^T V(\theta)^{-1}(Y - \mu(\theta))$$

where $\mu(\theta)$ is a model for the mean of the response vector Y given θ and $V(\theta)$ is a model for its covariance matrix.

Since these are all one-dimensional equations, we can now provide conditions under which these estimates are regular. We begin by defining a partial order for vectors of length N:

$$\mathbf{y}_0 \preceq \mathbf{y}_1 \text{ if } y_i^0 \leq y_i^1, \ i = 1, \dots, N,$$

with $\mathbf{y}_0 \prec \mathbf{y}_1$ implying at least one of the point-wise inequalities is strict.

Theorem 7.1. *Let $\hat{\theta}(\mathbf{y})$ be defined by (15) and assume that*

1. *$S(\theta, \mathbf{y})$ is decreasing in θ for all $\mathbf{y}$.*
2. *$\mathbf{y}_0 \preceq \mathbf{y}_1$ implies $S(\theta, \mathbf{y}_0) \leq S(\theta, \mathbf{y}_1)$ for all θ.*

Then $\mathbf{y}_0 \preceq \mathbf{y}_1$ implies $\hat{\theta}(\mathbf{y}_0) \leq \hat{\theta}(\mathbf{y}_1)$. That is $\hat{\theta}(\mathbf{y})$ is regular.

Proof. We have

$$S(\hat{\theta}(\mathbf{y}_1), \mathbf{y}_0) \leq S(\hat{\theta}(\mathbf{y}_1), \mathbf{y}_1) = 0.$$

Since $S(\hat{\theta}, \mathbf{y}_0)$ is decreasing in $\hat{\theta}$ the result follows. □

This result can also be extended to marginal inference. Using the terminology in Hooker et al. (2009), the next result follows directly from Lemma 3.5 in the same paper.

Theorem 7.2. *Let $T[f]$ be a monotone functional of the log posterior distribution. Under the conditions of Theorem 7.1, $T[f]$ is regular.*

One means of incorporating dependence among bundles is via including dependence terms that are independent of θ. Fixed dependency effects model this in terms of expanding a logistic model:

$$p(\mathbf{y}|\theta) = \frac{e^{\sum_{i=1}^N y_i a_i(\theta + b_i) + \sum_{i=1}^{N-1} \sum_{j=i+1}^N y_i y_j \beta_{i\gamma_j}}}{\sum_{\mathbf{z} \in \{0,1\}^N} e^{\sum_{i=1}^N z_i a_i \theta + \sum_{i=1}^{N-1} \sum_{j=i+1}^N z_i z_j \beta_{i\gamma_j}}}.$$

Here the $\beta_{i\gamma_j}$ are used to account for dependence between answers. They are zero if items i and j do not belong to the same bundle, and positive otherwise. We assume all the a_i are positive. In this case we can examine the likelihood score:

$$S(\theta, \mathbf{y}) = \sum_{i=1}^N y_i a_i - \frac{\sum_{\mathbf{z} \in \{0,1\}^N} z_i a_i e^{\sum_{i=1}^N z_i a_i \theta + \sum_{i=1}^{N-1} \sum_{j=i+1}^N z_i z_j \beta_{i\gamma_j}}}{\sum_{\mathbf{z} \in \{0,1\}^N} e^{\sum_{i=1}^N z_i a_i \theta + \sum_{i=1}^{N-1} \sum_{j=i+1}^N z_i z_j \beta_{i\gamma_j}}} \tag{16}$$

from which it is easy to see that Condition 2 of Theorem 7.1 holds. To show Condition 1 we observe that

$$\begin{aligned}\frac{d}{d\theta}S(\theta,\mathbf{y}) &= -\frac{\sum_{\mathbf{z}\in\{0,1\}^N} z_i a_i^2 e^{\sum_{i=1}^N z_i a_i \theta + \sum_{i=1}^{N-1}\sum_{j=i+1}^N z_i z_j \beta_{i\gamma_j}}}{\sum_{\mathbf{z}\in\{0,1\}^N} e^{\sum_{i=1}^N z_i a_i \theta + \sum_{i=1}^{N-1}\sum_{j=i+1}^N z_i z_j \beta_{i\gamma_j}}} \\ &\qquad + \left(\frac{\sum_{\mathbf{z}\in\{0,1\}^N} z_i a_i e^{\sum_{i=1}^N z_i a_i \theta + \sum_{i=1}^{N-1}\sum_{j=i+1}^N z_i z_j \beta_{i\gamma_j}}}{\sum_{\mathbf{z}\in\{0,1\}^N} e^{\sum_{i=1}^N z_i a_i \theta + \sum_{i=1}^{N-1}\sum_{j=i+1}^N z_i z_j \beta_{i\gamma_j}}}\right)^2 \\ &\leq 0\end{aligned}$$

from the same arguments as (3). To apply Theorem 7.1 for Bayesian methods we simply note that if we add a log-concave prior $\pi(\theta)$ to the log-likelihood (16), both conditions clearly continue to hold.

Quasi-likelihood methods suffer from the same complications as random-effects methods with correlated factors. It is easy to show that the estimate of θ is regular when $V(\theta)$ is diagonal. However, modeling positive correlations among the responses may yield paradoxical results.

8 Conclusions

Statistical estimates for multidimensional item response theory are prone to paradoxical results. These can have adverse real-world consequences. Prior information can be used to control this phenomenon but only in so far as the prior is not overwhelmed by the data. In real-world applications, practitioners may attempt to reduce these effects by the use of separable tests in which only one ability is measured by each item. Where item bundles are used to control for dependence between responses, separability is no longer possible. However, in this case we can think of tests comprised of item bundles as being collections of multiple short tests, each of which is individually regulated by the prior. In that sense it is possible to find such tests that do not produce paradoxical results.

We have investigated the extent to which this intuition can be applied to create methods to select regular tests. When a prior is used that represents all bundle effects as independent, it is possible to produce an algorithm that efficiently selects subsets from each bundle that can then be combined to form a test that cannot create paradoxical results. This algorithm can then be used within a test-selection algorithm. When the bundle effects are not independent, it is harder to produce such an analysis without placing particular structure on the prior covariance.

Our experience with item parameters estimated from real data demonstrates that our proposed routine for finding maximal regular subsets is computationally efficient and easy to implement and that the approximations we propose are not overly conservative. It also suggests that real-world tests based on item bundles require minor, but not drastic, alteration to be guaranteed to yield regular results when the bundle factors are modeled as being independent in the subject population. However, even mild dependence between bundle factors can substantially reduce the size of regular subsets when the bounds derived above are used.

It is important to note the special case of Rasch models in which discrimination parameters for the bundle ability and the ability of interest are the same. When these factors are modeled as independent, estimates of the ability of interest can be shown to be regular. However, when positive correlation between the bundle factors is allowed, paradoxical results remain possible.

The use of bundle abilities is not the only means of allowing for dependence between items. Fixed dependency effects are an alternative means of doing so that do not require individual bundle abilities to be modeled explicitly. These models are generally more computationally intensive to compute, but can be shown to always produce regular estimates of an ability of interest. There are yet further models, such as treating each bundle as a single unit with a polytomous response (Wilson and Adams, 1995), where conditions for regularity have not yet been ascertained.

References

Ackerman, T. (1996). Graphical representation of multidimensional item response theory analyses. *Applied Psychological Measurement 20*(4), 311–329.

Bock, R., R. Gibbons, and E. Muraki (1988). Full-information item factor analysis. *Applied Psychological Measurement 12*, 261–280.

Craven, B. D. (1988). *Fractional Programming.* Berlin: Heldermann Verlag.

Douglas, J. A., L. A. Roussos, and W. Stout (1996). Item-bundle dif hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Jounral of Educational Measurement 33*, 465–484.

Finkelman, M., G. Hooker, and J. Wang (2007). Unidentifiability and lack of monotonicity in the multidimensional three-parameter logistic model. under review.

Hooker, G., M. Finkelman, and A. Schwartzman (2009). Paradoxical results in multidimensional item response theory. *Psychometrika to appear.*

Hoskens, M. and P. de Boeck (1997). A parametric model for local dependence among test items. *Psychological Methods 2*, 261–277.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika 49*, 223–245.

Li, Y., D. M. Bolt, and J. Fu (2006). A comparison of alternative models for testlets. *Applied Psychological Measurement 20*(1), 3–21.

McCullagh, P. and J. A. Nelder (1989). *Generalized Linear Models.* London: Chapman and Hall/CRC.

Reckase, M. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement 9*, 401–412.

Rijmen, F., F. Tuerlinckx, P. de Boeck, and P. Kuppens (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods 8*(2), 185–205.

Rosenbaum, P. R. (1988). Item bundles. *Psychometrika 53*, 349–359.

Veldkamp, B. P. (2002). Multidimensional constrained test assembly. *Applied Psychological Measurement 26*(2), 133–146.

Wang, W. and M. Wilson (2005). The Rasch testlet model. *Applied Psychological Measurement 29*(2), 126–149.

Wang, X., E. T. Bradlow, and H. Wainer (2002). A general Bayesian model for testlets: Theory and applications. *Applied Psychological Measurement 26*(1), 109–128.

Wilson, M. and R. J. Adams (1995). Rasch models for item bundles. *Psychometrika 60*, 181–198.

A Proof of Theorem 3.1

In order to extend the proof of Theoremns 6.1 and 6.2 in Hooker et al. (2009) to the case of EAP estimates and estimates with $b_{\boldsymbol{\gamma}}$ as a random effect, we follow a similar argument. Throughout, we refer to the posterior distribution for $\boldsymbol{\theta}$. Where a random-effects model is used, no prior is used for θ and we intend the log posterior to be read as being the distribution of $b_{\boldsymbol{\gamma}}$ *given* θ and the observed responses. Any (log) prior $\mu(\boldsymbol{\theta})$ will also be taken to be constant over θ in the random-effects case.

Let

$$l(\boldsymbol{\theta}, \mathbf{y}) = \sum_{i=1}^{N-1} l_i(\mathbf{a}_i^T \boldsymbol{\theta}, y_i) + \mu(\boldsymbol{\theta})$$

be the log posterior distribution (up to an additive constant) for θ given the first $N-1$ responses where $\mu(\boldsymbol{\theta})$ is the log prior. Without loss of generality, we assume $\mathbf{b}^T\mathbf{b} = 1$. We begin by defining

$$D^* = \begin{bmatrix} D \\ \mathbf{b} \end{bmatrix},$$

where D is a d-by-$(d+1)$ orthonormal matrix with rows orthogonal to $\mathbf{b}$. We re-parameterize our ability space $\boldsymbol{\psi} = D^*\boldsymbol{\theta}$ so that

$$l_i(\boldsymbol{\theta}^T\mathbf{a}_i, y_i) = l_i(\boldsymbol{\psi}^T D^*\mathbf{a}_i, y_i).$$

We note, importantly, that $l_N(\boldsymbol{\psi}^T D^*\mathbf{b}, y_n) = l_N(\psi_d, y_N)$, so the final item now only depends on ψ_d. We define a further transformation $\boldsymbol{\phi} = C\boldsymbol{\psi}$ to be an orthogonal projection into the d-dimensional space orthogonal to ψ_d such that

$$\boldsymbol{\phi}_1 = \boldsymbol{\theta}_1 - b_\theta \psi_d$$

and we will write $l^\phi(\phi_1, \boldsymbol{\phi}_{(1)}, \psi_d, \mathbf{y})$ to express the log posterior in terms of $\boldsymbol{\phi}$ and ψ_d where $\boldsymbol{\phi}_{(1)} = (\phi_2, \dots, \phi_{d-1})$. The log posterior for the marginal distribution of θ and ψ_d is now given by

$$\tilde{l}(\theta, \psi_d, \mathbf{y}) = \log \int e^{l^\phi(\theta - b_\theta\psi_d, \boldsymbol{\phi}_{(1)}, \psi_d, \mathbf{y})} d\boldsymbol{\phi}_{(1)}$$

and its cross derivatives with respect to θ and ψ_d are given by

$$\frac{d}{d\theta d\psi_d}\tilde{l}(\theta, \psi_d, \mathbf{y}) = \mathbf{c}_1^T D^T(A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A + K(\boldsymbol{\theta}))\mathbf{b} - b_\theta \mathbf{c}_1^T D^T(A^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A + K(\boldsymbol{\theta}))D\mathbf{c}_1$$

where $W(\boldsymbol{\theta}, \mathbf{y}^{N-1})$ and $K(\boldsymbol{\theta})$ are given as in A12. If this quantity is negative for all $\boldsymbol{\theta}$, the existence of a paradoxical result can be deduced from Theorem 5.2 in Hooker et al. (2009) for expected *a posteriori* estimates and by additionally applying Lemma 3.2 in the same paper for MR estimates. The regularity of the estimate if the quantity is positive follows from analogous arguments. This regularity condition is equivalent to (2). The calculations for this result are given in the proof of Theorem 6.1 in Hooker et al. (2009) and are not reproduced here. □

B Item Bundles and A12

The particular structure of item bundles means that when EAP or MR estimates are used the correction A12 only occurs in the $1/k'$ term of (4). To see this, we need only examine

$$A_\theta^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} \left(A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta}, \mathbf{y}^{N-1})A_{\boldsymbol{\gamma}} + K(\boldsymbol{\theta})_{\boldsymbol{\gamma},\boldsymbol{\gamma}} \right)^{-1} b_{\boldsymbol{\gamma}}.$$

Since $\mathbf{b}$ only has b_θ and b_{γ_d} as nonzero entries, the conditional expectations used in A12 can be replaced with expectations conditional on θ and γ_d. We observe that

$$\mathrm{Cov}\left(K\boldsymbol{\theta}|\theta,\gamma_d\right) = K\mathrm{Cov}\left(\boldsymbol{\theta}|\theta,\gamma_d\right)K^T$$

has zero rows and columns corresponding to θ and γ_d since K is diagonal. Since $A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta},\mathbf{y}^{N-1})A_{\boldsymbol{\gamma}}$ is also diagonal, the γ_dth row and columns in the matrix inverse are not changed by the correction factor. Since $b_{\boldsymbol{\gamma}}$'s only nonzero entry corresponds to the γ_dth entry, the correction A12 does not enter into (5).

C Proof of Observation 5.1

We demonstrate that for off-diagonal elements $\sigma_{\gamma_i\gamma_d}$ of Σ, the derivative of (12) at $\tilde{\Sigma}$ is positive. The observation then follows directly.

For the sake of notational simplicity, we let

$$\tilde{\mathbf{a}} = A_\theta^T W(\boldsymbol{\theta},\mathbf{y})A_{\boldsymbol{\gamma}},\ \tilde{A} = A_{\boldsymbol{\gamma}}^T W(\boldsymbol{\theta},\mathbf{y})A_{\boldsymbol{\gamma}}$$

since these will not change through the proof. Note that all the entries in $\tilde{\mathbf{a}}$ are negative and that $\tilde{A}$ is diagonal with non-positive entries. Letting $E_{\gamma_i\gamma_j}$ be the matrix of zeros with 1 in the (γ_i,γ_j)th position,

$$\begin{aligned}\frac{d}{d\sigma_{\gamma_i\gamma_d}}\tilde{\mathbf{a}}^T\left(\tilde{A}-\Sigma^{-1}\right)^{-1}\mathbf{e}_{\gamma_d} &= -\tilde{\mathbf{a}}^T\left(\tilde{A}-\tilde{\Sigma}^{-1}\right)^{-1}\tilde{\Sigma}^{-1}\left(E_{\gamma_i\gamma_d}+E_{\gamma_d\gamma_i}\right)\tilde{\Sigma}^{-1}\left(\tilde{A}-\tilde{\Sigma}^{-1}\right)^{-1}\mathbf{e}_{\gamma_d}\\ &= \frac{-\tilde{a}_{\gamma_i}}{\sigma_{\gamma_i\gamma_i}\sigma_{\gamma_d\gamma_d}\left(\tilde{A}_{\gamma_i\gamma_i}-\frac{1}{\sigma_{\gamma_i\gamma_i}}\right)\left(\tilde{A}_{\gamma_d\gamma_d}-\frac{1}{\sigma_{\gamma_d\gamma_d}}\right)}\\ &> 0.\end{aligned}$$

□

D Proof of Theorem 5.1

We make the same notational conventions as in Appendix C and observe that

$$\Sigma^{-1} = gI - eJ,\ g = \frac{1}{\sigma^2-\alpha},\ e = \frac{\alpha g}{(\sigma^2+(d-1)\alpha)}.$$

Making the further substitutions:

$$C = \tilde{A} - gI,\ \mathbf{c} = \mathrm{diag}(C^{-1})$$

we observe that by the Woodbury formula

$$(\tilde{A}-\Sigma^{-1})^{-1} = C^{-1} - e\mathbf{c}\left(I + e1^TC^{-1}1\right)\mathbf{c}^T$$

so that

$$
\begin{aligned}
\tilde{\mathbf{a}}^T(\tilde{A}-\Sigma^{-1})^{-1}\mathbf{e}_{\gamma_d} &= \frac{\tilde{a}_{\gamma_d}}{C_{\gamma_d\gamma_d}} - \frac{\frac{e}{C_{\gamma_d\gamma_d}}\tilde{\mathbf{a}}^T\mathbf{c}}{1+e\sum C^{-1}_{\gamma_d\gamma_d}} \\
&= \frac{\tilde{a}_{\gamma_d}}{\tilde{A}_{\gamma_d\gamma_d}-g} - \frac{\frac{e}{\tilde{A}_{\gamma_d\gamma_d}-g}\sum_{i=1}^d \frac{\tilde{a}_{\gamma_i}}{\tilde{A}_{\gamma_i\gamma_i}-g}}{1+\sum_{i=1}^d \frac{e}{(\tilde{A}_{\gamma_i\gamma_i}-g)}} \\
&\le \frac{\tilde{a}_{\gamma_d} - \frac{eg}{g-de}\sum_{i=1}^d \frac{\tilde{a}_{\gamma_i}}{\tilde{A}_{\gamma_i\gamma_i}-g}}{\tilde{A}_{\gamma_d\gamma_d}-g}. \qquad (17)
\end{aligned}
$$

The result now follows by substituting for e, g, $\tilde{\mathbf{a}}$ and $\tilde{A}$ and maximizing over $\sum_{i=2}^d \tilde{a}_{\gamma_i}/(\tilde{A}_{\gamma_d\gamma_d}-g)$. □