EFFICIENCY OF SPLIT EXAM DESIGNS

Linda M. Zeger, Educational Testing Service, Neal Thomas, University of North Carolina – Chapel Hill
Linda M. Zeger, MS–15T, Educational Testing Service, Princeton, NJ 08541

1 Introduction

It is often desirable to decrease the length of surveys, as there is typically a higher nonresponse rate and lower quality of response associated with lengthy surveys [1]. Furthermore, longer surveys may be more susceptible to nonignorable nonresponse [2]. Raghunathan and Grizzle [2] investigated decreases in the length of surveys through administration of split questionnaires in which each sampled individual receives only some of the survey items from a complete questionnaire.

Similarly, time constraints restrict the number of questions on educational tests such as the NAEP (National Assessment of Educational Progress) exams. Whereas in many surveys a characteristic of respondents is estimated from their responses to a single item, in the NAEP exams an estimate of students' abilities in one area is derived from responses to a number of test items. Hence, in the latter case the limited length of the exam produces measurement errors with variance that decreases with the number of items given.

The goal of the NAEP exams is to assess population characteristics of students' abilities, rather than to report individuals' scores. The estimator of a population mean of students' abilities has variance originating both from the finite sample size of the examinees and from the limited number of exam items given to each student. In a similar context Lord and others [3] have investigated the tradeoff between sample size and the number of items given to the sampled individuals.

Unlike the tests considered by Lord in which students' abilities in a single area were estimated, the NAEP exams assess students' abilities in several correlated skill areas, called subscales, within each subject. The question of efficient designs we consider is also different because the number of sampled students is fixed, and the issue is how to allocate items measuring different subscales. Measuring one subscale accurately in a portion of the sample could potentially provide information on another correlated subscale measured less accurately in that portion. We investigate this information borrowing for split exam designs in which the number of items devoted to a subscale varies across students. The accuracy of population estimates derived from split exams are compared to those of a balanced design, in which the allocation of items among subscales is the same for every student. This problem is a generalization of the design question for split survey questionnaires to nonzero measurement errors.

This paper is organized as follows: In Section 2 we use examples from recent NAEP exams to illustrate which types of designs result in information borrowing. In Section 3 we describe a model for the measurement errors similar to the NAEP models which enables tractable calculations of efficiencies for a wide range of designs. We discuss the maximum likelihood estimates derived from this model in Section 4. The efficiencies of a class of bivariate split designs are compared in Section 5. Finally, extensions to other types of designs, including those measuring a larger number of variables, are discussed in Section 6.

2 Information Borrowing

We consider split exam designs in which a constant total number of items is assigned to each student, but the number of items allocated to each subscale varies across the students. This uneven distribution of subscale items is sometimes imposed by the nature of exam items and the length constraint of the NAEP exams. For example, long reading passages consume a large portion of an exam, allowing students receiving these passages to be tested on only one Reading subscale measured by the items associated with the passage.

In 1992 NAEP administered a split design Reading test to 8416 sampled students with two subscales called Literature and Information. A test form consisting of only items on the Literary subscale was given to 38% of the examinees, while a second form of approximately the same length containing only items on the Information subscale was given to a second group of examinees comprising 37% of the sampled students. The remaining 25% of the examinees were given a test form of the same length with half the items devoted to each of the two subscales. A goal of our study is to investigate how such a design compares to the corresponding balanced design, in which all sampled students would be given the test form containing half its questions on each subscale.

In the split design the sample size of students receiving each subscale is smaller than in the corresponding balanced design. However, in the group receiving only one subscale, that subscale is measured more accurately than in the balanced design. Moreover, since the two subscales are highly correlated, one might expect that responses to one subscale may be used in the estimator of the mean of the other subscale, thus improving the accuracy of this estimator. In order to address the question of whether these effects offset the decrease in sample size in the split design, we first examine the borrowing of information between subscales for two NAEP exams.

2.1 NAEP Model

The NAEP model used to estimate regression coefficients and their efficiencies is described in this section. We assume a random sample of n examinees is drawn from an infinite population of students. The i^{th} sampled student is assumed to have true abilities $\theta_i^T = [\theta_{i1}, \theta_{i2}, \cdots, \theta_{ip}]$ in p related subscales. The θ_i are assumed to be independent and normally distributed with means

$$E\left[\theta_{ij}\right] = \mathbf{x}_i \cdot \beta_j. \tag{1}$$

for $j = 1,2, \ . \ . \ p$, and $i = 1,2, \ . \ . \ n$. The $(1 \times l)$ vector of covariates for the i^{th} examinee, denoted by $\mathbf{x}_i$, includes background traits that are known for all examinees, and the $(l \times 1)$ vector β_j contains the regression coefficients for the j^{th} subscale. The variances and covariances, given $\mathbf{x}_i$, are

$$\begin{aligned} Var(\theta_{ij}) &= \sigma_{jj} \\ Cov(\theta_{ij}, \theta_{ik}) &= \sigma_{jk} \end{aligned} \tag{2}$$

for $i = 1,2, \ . \ . \ n$.

The true abilities θ_{ij} are not directly observed, but instead are measured through students' exam responses, where y_{ijk} denotes the response of the i^{th} student to the k^{th} question on the j^{th} subscale. It is assumed that each item on the exam pertains to only one subscale. NAEP models the measurement errors by specifying the probability of a correct response at ability θ_{ij}

$$\begin{aligned} Pr(y_{ijk} = 1 \mid \theta_{ij}) = \\ \lambda_{jk} + (1 - \lambda_{jk}) / \left[1 + exp\left(\alpha_{jk}(\theta_{ij} - \gamma_{jk})\right)\right] \end{aligned} \tag{3}$$

for dichotomous questions. The item specific parameters α_{jk} and γ_{jk} must be determined for each question, creating a complex likelihood function for each examinee. An analogous form is used to model the multiple possible responses to polytomous items.

2.2 NAEP Data

A likelihood function for the regression coefficients can be caluclated from the normal model for the θ_i and the measurement model specified in Section 2.1. The variance of the resulting maximum likelihood estimators (MLEs) for the β_j originates from the finite sample size as well as from the measurement errors of y_{ijk}. Here we examine how much the variances of the estimators of regression coefficients for one subscale are diminished by responses to items on a second subscale.

The univariate estimator $\tilde{\beta}_j$ for the j^{th} subscale is the maximum likelihood estimator that would be obtained if only the responses to items pertaining to the j^{th} subscale were used. A second estimator, which we call the bivariate estimator $\hat{\beta}_j$, is the maximum likelihood estimator obtained from responses to items on both the j^{th} subscale and on the additional subscale. The bivariate-univariate efficiency for a single regression coefficient, defined by

$$R = \frac{Var(\hat{\beta}_j)}{Var(\tilde{\beta}_j)}, \tag{4}$$

measures the relative information gain of the bivariate estimator over the univariate estimator. When R attains its maximum value of 1.0, the bivariate and univariate estimators have the same variance, and no information is gained from using responses to items on one subscale in estimation of a regression coefficient of the other subscale, whereas a small value of R indicates substantial information borrowing.

The responses to the NAEP exams and the model described in Section 2.1 can be used to calculate these efficiencies. We performed these calculations using a nine covariate regression model for a split Reading exam as well as a balanced Math exam. We report the efficiency of the estimated regression equation evaluated at the mean of the covariate values.

In the Reading exam described at the beginning of Section 2, the average value of R was 0.91 for the Literary subscale and 0.92 for the Information subscale. In the 1990 NAEP Math exam all 8790 sampled students were given roughly two thirds of their items on the Numbers and Operations subscale and the remainder on the Measurement subscale. This design is balanced because all students received similar allotments of items on the two subscales. The efficiency of this exam is 0.99 for the Numbers and Operations subscale and 0.98 for the Measurement subscale. These efficiencies are quite close to 1.0, indicating almost no information borrowing. Although the Reading exam had little information borrowing, it did have more than the Math exam, mainly because the Reading exam has a split design whereas the Math exam has a balanced design. We introduce a simplified measurement error model in Section 3 which allows us to clarify this relationship between R and exam design, as detailed in Section 4.

3 Measurement Model

The NAEP exams typically collect background information forming over a hundred covariates and often test more than two subscales, resulting in computationally intractable calculations of efficiencies using the model of Section 2.1. In order to calculate the variances of estimators for a large number of potential designs, we introduce a normal model to approximate the NAEP measurement error model.

Instead of modeling the measurement errors at the item level, we use only the aggregate score y_{ij} from all items given to the i^{th} student on the j^{th} subscale. The assumption of additive measurement errors, denoted by δ_{ij}, yields the observed score

$$y_{ij} = \theta_{ij} + \delta_{ij}. \tag{5}$$

The measurement errors satisty

$$E[\delta_{ij}] = 0. \tag{6}$$

We assume every item for each subscale provides equal information, and thus the measurement accuracy of y_{ij} is determined by n_{ij}, the number of items given to student i on the j^{th} subscale. The measurement error variance is then

$$Var(\delta_{ij}) = \frac{\tau_o}{n_{ij}} \tag{7}$$

where τ_o is a constant scale factor. The measurement errors are assumed to be independent of the θ_{ij}, and δ_{ij} is also assumed independent of δ_{rs} unless $i = r$ *and* $j = s$, resulting in independence of observed scores for different students.

The total variance, incorporating both sampling and measurement error variances, is thus

$$\Sigma^i_{jj} \equiv Var(y_{ij}) = \sigma_{jj} + \frac{\tau_o}{n_{ij}}. \tag{8}$$

This variance is not the same for all students on subscale j if the n_{ij} are not equal for all i. The covariances for all students are identical, and are denoted by

$$\Sigma^i_{jk} \equiv Cov(y_{ij}, y_{ik}) = \Sigma_{jk} = \sigma_{jk}, \quad j \neq k. \tag{9}$$

The additional assumption of normally distributed δ_{ij} results in a normal distribution for the y_{ij}, because the θ_{ij} are assumed normally distributed. The approximating normal measurement error model is equivalent to the generalized least squares (GLS) equation:

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_p \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & 0 & \cdots & 0 \\ 0 & \mathbf{X}_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \mathbf{X}_p \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_p \end{bmatrix}. \tag{10}$$

The outcome variable $\mathbf{y}_j$ in equation (10) is an $(n \times 1)$ column of the n observations y_{ij} for the j^{th} subscale. The vector β_j denotes the l regression coefficients for outcome variable $\mathbf{y}_j$, as in equation (1). Each covariate X_j is an $(n \times l)$ matrix containing the n rows of the examinees' l background variables $\mathbf{x}_{ij}$ used to predict the responses $\mathbf{y}_j$. In the NAEP exams the covariates for each subscale are identical $(\mathbf{x}_{ij} = \mathbf{x}_i)$:

$$\mathbf{X}_1 = \mathbf{X}_2 = \ldots = \mathbf{X}_p. \tag{11}$$

The normally distributed random errors, given by the $(n \times 1)$ column vector ϵ_j for the n observed examinees, are the sum of population and measurement errors of subscale score $\mathbf{y}_j$. Random error terms ϵ_{ij} and ϵ_{rk} for the j^{th} and k^{th} subscales are independent for different observations $(i \neq r)$, while for the same observation $(i = r)$, the covariances are given by (8) or (9).

Univariate MLEs arise from treating the p related equations of (10) as p separate equations. A separate MLE, $\tilde{\beta}_j$, is computed for the j^{th} equation for each j. In typical situations, multivariate MLEs are identical to the univariate MLEs, even when the outcome variables are highly correlated. However, Zellner [4] shows that when equation (11) is not satisfied, the multivariate MLEs are more efficient than the univariate MLEs, and information borrowing between variables occurs. In this case, multivariate MLEs for regression coefficients of one outcome variable depend on observations of the other outcome variables, in contrast to when (11) holds.

In Section 4 we show a second case in which multivariate MLEs are more efficient than univariate MLEs occurs when the random error terms of ϵ_j are not identically distributed for all n observations, even when (11) holds. This situation arises in split exam designs in NAEP, where the scores of a sample of n students in each of p subscales are observed with variances that can differ across students, according to the number of items given.

We consider the model with no regressor variables so there is one coefficient for each subscale (l

$= 1$) in (10), which is the subscale mean: $\beta_j = \mu_j$. The corresponding covariate $\mathbf{X}_j$, a column containing n 1's for each j, obeys (11).

4 Estimators and their Efficiencies

In this section we show how the bivarate MLEs differ from the univariate MLEs, as well as how the variances of these estimators differ when the measurement error variance is not constant across examinees. For simplicity, we assume the Σ and τ_o are known throughout so that the effective variances of (8) and (9) are also known.

Using the normal measurement error model introduced in Section 3, we have derived expressions for the MLEs $\hat{\mu}_j$ of the subscale population means μ_j as well as Var($\hat{\mu}_j$).

Using the GLS formulation in the bivariate case ($p = 2$), we obtain the MLE [1]

$$\hat{\mu}_2 = \sum_{i=1}^{n} \frac{1}{v^i} \left[y_{i2} + \frac{\Sigma_{12}}{\Sigma_{11}^i}(\hat{\mu}_1 - y_{i1}) \right] \tag{12}$$

where $v^i = Var(y_{i2}|y_{i1}) / \sum_{m=1}^{n} Var(y_{m2}|y_{m1})$ and

$$Var(y_{i2}|y_{i1}) = \Sigma_{22}^i - \frac{(\Sigma_{21})^2}{\Sigma_{11}^i}. \tag{13}$$

Equation (12) shows that observations y_{i1} contribute to the estimator $\hat{\mu}_2$ by adjusting each term in the sum according to the "regression coefficients" $\frac{\Sigma_{12}}{\Sigma_{11}^i}$. For example, when $\frac{\Sigma_{12}}{\Sigma_{11}^i}$ is large and $y_{i1} < \hat{\mu}_1$, the observation y_{i2} is likely to be less than μ_2, so the estimator in (12) adjusts the i^{th} term in the summation upwards towards the mean. Thus each term in the sum of equation (12) is likely to be closer to the mean than y_{i2}, resulting in a more efficient estimator than the univariate estimator $\tilde{\mu}_2$, which is a weighted sum of the y_{i2} only. By eliminating $\hat{\mu}_1$ from (12) and expressing $\hat{\mu}_2$ directly in terms of y_{i1} and y_{i2}, it can be shown that if we assign the same number of items on the second subscale to all pupils, resulting in a single constant Σ_{22}^i for all i, then $\hat{\mu}_2$ is identical to the univariate estimator $\tilde{\mu}_2$. Thus for this allocation of items $\hat{\mu}_2$ does not depend on observations y_{i1}.

We obtain the variance of $\hat{\mu}_2$

$$\begin{aligned} Var(\hat{\mu}_2) \quad = \quad & \left[\sum_{i=1}^{n} \frac{1}{\Sigma_{22}^i} \left[1 + \left(\frac{1}{1 - \frac{(\Sigma_{12})^2}{\Sigma_{11}^i \Sigma_{22}^i}} \right) \times \right.\right. \\ & \left.\left. \left(\frac{\Sigma_{12}}{\Sigma_{11}^i} \right) \left(\frac{\Sigma_{12}}{\Sigma_{22}^i} - \frac{Cov(\hat{\mu}_1, \hat{\mu}_2)}{Var\hat{\mu}_2} \right) \right] \right]^{-1} , \end{aligned} \tag{14}$$

where the term

$$\frac{Cov(\hat{\mu}_1, \hat{\mu}_2)}{Var\hat{\mu}_2} = \frac{\sum_{i=1}^{n} \frac{\Sigma_{12}}{\Sigma_{22}^i} \frac{1}{Var(y_{i1}|y_{i2})}}{\sum_{i=1}^{n} \frac{1}{Var(y_{i1}|y_{i2})}} \tag{15}$$

with $Var(y_{i1}|y_{i2})$ defined as in (13), is a weighted average of the $\frac{\Sigma_{12}}{\Sigma_{22}^i}$. When the Σ_{22}^i are very different from one another so as to produce sizable differences of $\frac{\Sigma_{12}}{\Sigma_{22}^i}$ from this average value, equation (14) will differ substantially from the univariate result

$$Var(\tilde{\mu}_2) = \left(\sum_{i=1}^{n} \frac{1}{\Sigma_{22}^i} \right)^{-1}. \tag{16}$$

It can be shown that in this case $Var(\hat{\mu}_2) < Var(\tilde{\mu}_2)$, which is consistent with the fact that $\hat{\mu}_2$ is the best linear unbiased estimator. When Σ_{22}^i is the same for all i, equation (15) yields $\frac{Cov(\hat{\mu}_1,\hat{\mu}_2)}{Var\hat{\mu}_2} = \frac{\Sigma_{12}}{\Sigma_{22}^i}$, and equation (14) reduces to the variance (16), as expected because the bivariate MLE reduces to the univariate estimator in this case. Therefore, unequal variances across examinees result in information borrowing from observations of one subscale to the bivariate estimator of the population mean of the other subscale, and make this estimator more efficient than the univariate estimator.

We use equation (14) to calculate the efficiencies R defined in equation (4) for estimates of subscale means corresponding to the designs of the NAEP Reading and Math exams specified in Section 2. The population variances (2) and the measurement error variances used in this calculation were derived from the NAEP data for the respective exams, subject to the assumed form of equation (7). The normal model predicts $R = 0.92$ for the Literary and Information subscales of the Reading exam. For both subscales on the Math exam, the normal measurement error model yields $R = 1.0$ since the variances Σ_{jj}^i are constant across examinees. These normal measurement model predictions are quite close to values of R calculated from NAEP measurement models and data that were presented in Section 2.2 for these designs. This comparison demonstrates the normal model's potential utility in predicting efficiencies.

We have obtained explicit expressions which appear as natural extensions to the bivariate case analogous to equations (12) and (14) for the trivariate problem, in which case $\hat{\mu}_2$ can depend on observations of both y_{i1} and y_{i3}. We have also extended the calculation of the variance of the estimators of the subscale means to the general multivariate case

[1] Note that (12) is similar to the formula obtained with a monotone missing data pattern in Little and Rubin [5].

with p subscales:[2]

$$Var \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \vdots \\ \hat{\mu}_p \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} \Sigma_i^{11} \cdots \sum_{i=1}^{n} \Sigma_i^{1p} \\ \sum_{i=1}^{n} \Sigma_i^{21} \cdots \sum_{i=1}^{n} \Sigma_i^{2p} \\ \vdots \quad \ddots \quad \vdots \\ \sum_{i=1}^{n} \Sigma_i^{p1} \cdots \sum_{i=1}^{n} \Sigma_i^{pp} \end{bmatrix}^{-1} . \quad (17)$$

Such cases are discussed in Section 6.

5 Split Designs

We have investigated a class of designs for an exam testing two subscales in a total of n sampled students with three exam forms. One exam form, given to m students, contains K items only on the first subscale, and has measurement error variance τ on this subscale. A second form, given to a second group of m students, consists of K items on the second subscale, and has a measurement error variance of τ on this subscale, as seen from equation (7). A third form, given to the remaining $n - 2m$ sampled students, has $K/2$ items on each of the two subscales and thus has measurement error variance 2τ on each subscale. The 1992 NAEP Reading exam design discussed in Section 2 can be approximated by the member of this class with $m = .375n$.

Normal measurement model predictions for the variance of estimators of subscale means, given by equation (14) or (17), enable comparisons for this class of designs. The range of the class is covered by holding n fixed and varying m from 0 to n/2. When m = 0, the design is balanced with all examinees receiving the third form. Because the exam design for this class is symmetric with respect to interchanging the two subscales, and $\sigma_{11} \cong \sigma_{22}$ for the exams considered by NAEP, it follows from (14) that Var($\hat{\mu}_1$) $\cong$ Var($\hat{\mu}_2$). [3] Thus the various split designs can be compared to the balanced design ($m = 0$) through an efficiency defined for a subscale mean:

$$E = \frac{Var(\hat{\mu}_2^{spt})}{Var(\hat{\mu}_2^{bal})} \quad (18)$$

where $\hat{\mu}_2^{spt}$ refers to the MLE of μ_2 derived from the split design, and $\hat{\mu}_2^{bal}$ is the corresponding MLE for the balanced design.

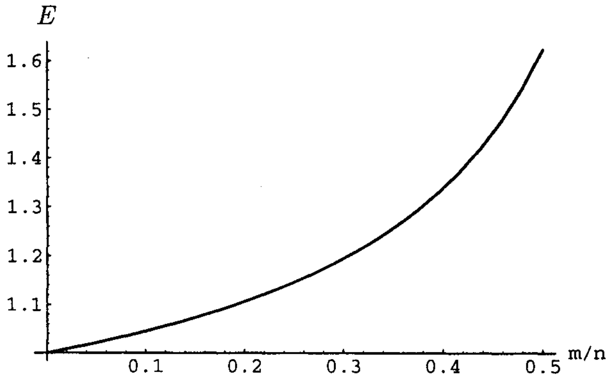

Figure 1: The efficiency E of the balanced design relative to the split designs for the symmetric bivariate class with $\rho = 0.88$ and $\tau/\sigma_{11} = 0.3$.

The efficiency E is plotted as a function of m/n in Figure 1 for a correlation between θ_{i1} and θ_{i2} of $\rho = 0.88$ and a variance ratio of $\tau/\sigma_{11} = 0.3$, the values of these parameters in the 1992 NAEP Reading exam. It is seen that Var($\hat{\mu}_2^{spt}$) increases with m/n, and we have proven that this is true for any $0 \leq \rho < 1$ and $\tau/\sigma_{11} \geq 0$ for $0 \leq \frac{m}{n} \leq 0.5$ [6], demonstrating that the balanced design is optimal.

Figure 1 illustrates that at $m/n = 0.375$, which is the design corresponding to the 1992 NAEP Reading exam, the normal measurement model predicts E = 1.31. This value is quite close to the value of E = 1.28 calculated directly from the NAEP data with the measurement model discussed in Section 2.1. This comparison, lending support to the normal model for efficiency predictions, confirms that the split design used by NAEP entails a loss in efficiency in estimating subscale means of approximately 30% relative to the balanced design.

The multivariate MLE $\hat{\mu}_2$ for split designs with discrepant Σ_{22}^{i} is more efficient than the univariate MLE $\tilde{\mu}_2$, as shown in Section 4. However, despite the occurrence of information borrowing in $\hat{\mu}_2$ for split designs and its absence in balanced designs, we have shown that any split design of the class considered here is less efficient than the balanced design. Thus information borrowing and the smaller measurement error variance in part of a split design are not great enough to compensate for the smaller sample size observed for each subscale.

Survey responses can be modeled as having zero measurement errors, a limiting case of the above class of designs. The i^{th} respondent's answer to the j^{th} survey question is given by $y_{ij} = \theta_{ij}$ when the j^{th} question appears on this respondent's questionnaire. Setting $\tau = 0$ in this class of designs corresponds to a split survey assessing two questions where only (1 - $2m/n$) of the respondents receive both questions. In this case when $\rho = 0.88$, equation (14) yields $E = 1.19$ for $m/n = 0.375$. The

[2] Here Σ_i^{jk} denotes the $(j,k)^{th}$ matrix element of the inverse of the matrix given by equations (8) and (9). For $p = 2$ the (2,2) matrix element of (17) can be expressed as (14).

[3] We defer until Section 6 discussion of nonsymmetric designs used when subscale population variances are unequal.

somewhat smaller efficiency loss here than in the NAEP Reading exam is due to greater information borrowing resulting from the increased correlation between y_{i1} and y_{i2}. When $\rho = 0.95$ the zero measurement error model produces $E = 1.09$ at $m/n = 0.375$, indicating only a 9% loss in efficiency when only 25% of the respondents are given both survey questions relative to when all respondents are given both questions. These results are consistent with those of Raghunathan and Grizzle, who found that when the correlation between items is high, a split survey design is almost as efficient as the complete questionnaire [2].

6 Discussion

We have extended our study to designs outside those in the class discussed in Section 5. For bivariate exams with $\sigma_{11} \neq \sigma_{22}$ split designs can be compared using an efficiency based on the largest eigenvalue λ_L of the Var$(\hat{\mu})$ matrix. We examined a class of designs testing n students with three exam forms each containing K items, given to m_1, m_2, and $n - m_1 - m_2$ students respectively. The first form has items only on subscale 1, while the second form has items only on subscale 2, and a third form contains k items on subscale 1 and the remaining $K - k$ items on subscale 2. Allowing m_1, m_2, and k to vary through much of their allowable values, we found that for the values of ρ and σ_{11}/σ_{22} which we explored λ_L is minimized for a balanced design ($m_1 = m_2 = 0$), with k depending on σ_{11}/σ_{22} so that more items are allocated to the subscale with greater σ_{jj}.

We have examined numerous additional split designs measuring from two to six subscales, with different numbers of forms and various allocations of subscale items. In all cases we investigated no split designs were more efficient than balanced designs consisting of the same length form. However, when population variances are equal, symmetric split designs with two or more subscales per form are not as inefficient relative to the corresponding balanced design as those with only one subscale per form, due to increased information borrowing. Thus in all cases explored in our investigation a balanced design is most efficient. Split designs can decrease respondent burden and perform almost as well as a balanced design when there are high correlations, small measurement errors, and properly chosen item allocations.

References

[1] Adams, L.M., and Gale, D. (1982), "Solving the Quandry Between Questionaire Length and Response Rate in Educational Research", *Research in Higher Education* **17**,231-240; Herzog, A.R., and Bachman, J.G. (1980), "Effects of Questionnaire Length on Response Quality", *Public Opinion Quarterly* **45**, 549-559; and subsequent studies.

[2] Raghunathan, T.E., and Grizzle, J.E. (1995), "A Split Questionnaire Survey Design", *Journal of the American Statistical Association* **90**, 54-63.

[3] Lord, F.M. (1962), "Estimating Norms by Item Sampling", *Educational and Psychological Measurement* **22**, 259-267; Pandey, T. and Carlson, D. (1976), "Assessing payoffs in the estimation of the mean using multiple matrix sampling designs", *Advances in Psychological and Educational Measurement*, edited by de Gruijter, D., and Van der Kamp, L., New York: Wiley; Shoemaker, D. (1970), "Item examinee sampling procedures and associated standard errors in estimating test parameters", *Journal of Educational Measurement* **4**, p. 255-262; Sirotnik, K., and Wellington, R., "Incidence Sampling: An Integrated Theory for Matrix Sampling", *Journal of Educational Measurement* **14**, 343-399.

[4] Zellner, A. (1962), "An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias", *Journal of the American Statistical Association*, **57**, p. 348-368.

[5] Little, R.J.A., and Rubin, D.R. (1987), *Statistical Analysis with Missing Data*, New York: John Wiley & Sons, 98-102.

[6] Zeger, L.M., and Thomas, N., (1995) "Increase of Variance in Split Designs", *Technical Report*, Educational Testing Service.