CHAPTER 6

TO WHAT EXTENT DO STUDENT PERCEPTIONS OF CLASSROOM QUALITY PREDICT TEACHER VALUE ADDED?

Stephen W. Raudenbush and Marshall Jean

Abstract

Surveys of student perceptions produce multiple measures of classroom quality. Our aim of this chapter is to decide which of these measures are most useful in predicting student learning. Conventional statistical methods can produce misleading results because the measures of classroom quality are quite highly correlated. We therefore introduce a new method, the Multilevel Variable Selection Model, and we apply this method to the Tripod survey of student perceptions, which provides seven indicators of classroom effectiveness based on twenty-eight items. We find that classrooms that are well-controlled and intellectually challenging produce comparatively large learning gains. Our new methods can readily be extended to study the combined contribution of student perception data, classroom observation data, and other measures to student learning gains.

Introduction

A key challenge in measuring teacher effectiveness is to clarify the predictive validity of student perceptions, classroom observations, and other indicators. We’d like to know which aspects of classroom life, as measured by these methods, are most useful in predicting student learning,
and we’d like to know just how predictive the entire ensemble of information might be. The problem is that the many indicators of classroom quality are highly correlated, and this makes it difficult to decide which kinds of data are most useful and how predictive the data are taken together. Unfortunately, standard methods of statistical prediction do not help much and can produce highly misleading results when the predictors of interest are highly correlated. In this chapter, we introduce a new method, the Multilevel Variable Selection Model (MVSM) to address this problem. We apply this method to the TRIPOD survey of student perceptions, which provides seven indicators of classroom effectiveness based on twenty-eight items. We find that classrooms that are well-controlled and intellectually challenging produce comparatively large learning gains. Controlling for control and challenge, teachers who “captivate” their students (as indicated by TRIPOD) actually do a little worse than those who do not. We consider possible explanations for the negative effect of captivate. These three indicators seem to capture nearly all of the information that TRIPOD provides about student learning. While past research suggests that student perceptions give important information about teacher effectiveness, our analytic approach suggests that TRIPOD as a whole explains more of the variance in teacher value added than is suggested using conventional analytic approaches. We check our results against an alternative method, “multilevel principal components analysis,” with strikingly convergent results. We discuss uses of these methods in revising applications of TRIPOD. Our new methods can readily be extended to study the combined contribution of student perception data, classroom observation data, and other measures to student learning gains.

**Why Measure Student Perceptions of Classroom Quality?**

We often hear fond stories about a particular teacher who made a difference in the life of a child, and it seems intuitively obvious that some teachers are more effective than others. Recent research supports this idea. Chetty, Friedman, Hilger, Saez, Schanzenbach, and Yagan (2010) computed for each of many kindergarten classrooms a “value-added” score—an estimate of how much children learned in their kindergarten year, on average, as measured by an achievement test. A key feature of this study is that the children had been assigned at random to their teachers. Remarkably, the researchers were able to obtain information on how these children fared as adults many years later. Children attending effective classrooms as indicated by value added attained more education and earned significantly more as adults, on average, than did children whose
classrooms had lower value-added scores. This result supports other studies that also used random assignment of children to classrooms (Kane & Staiger, 2008; Nye, Hedges, & Konstantopoulos, 2004), showing that teachers vary significantly in how effective they are at fostering learning, as measured by achievement tests.

Of course, teachers are not typically assigned at random to classrooms. A crucial methodological contribution of the work cited above and related work (c.f., Chetty et al., 2011) is that teacher value-added scores provide useful information even when it is not possible to assign teachers at random. The Measurement of Effective Teaching (MET) Project is the largest study to date to support this conclusion (Bill and Melinda Gates Foundation, 2010).

With these exciting results in mind, value-added scores are not a silver bullet. Their reliability in discriminating among teachers in any given year is modest (Glazerman, Goldhaber, Loeb, Staiger, Raudenbush, & Whitehurst, 2011; Kane & Staiger, 2008; McCaffrey et al., 2003). Moreover, it is difficult to collect such data frequently, and some teachers do not teach subjects that are covered by state achievement tests; for those teachers, we cannot compute value-added scores. Moreover, test scores provide little information that teachers might use to change their practices. The information tells us about global achievement, rather than pinpointing specific aspects of teaching that might need improvement. And value-added scores become available to teachers too late—after the tested children have moved on to another grade—to be used for improvement. Finally, achievement test scores capture only part of what makes a classroom effective for child and youth development.

Not surprisingly, then, policy-makers and researchers are intensely interested in collecting a wide array of information that can supplement achievement test scores as indicators of teaching effectiveness (Berk, 2005; Bill & Melinda Gates Foundation, 2013; Glazerman et al., 2011). Ideally, such measures of effectiveness would be reasonably reliable and would predict value-added scores while adding additional valuable information about the quality of a classroom experience.

Among the most promising approaches to assessing teaching quality is to obtain student perceptions via self-administered questionnaires. The TRIPOD Survey, developed by Ronald Ferguson of Harvard University, is one of the information sources for the MET study. Ferguson developed TRIPOD to reveal how teachers compare on what he called the “7Cs.” Thus, TRIPOD prompts each student to share perceptions of how much his teacher cares for him, how well the teacher controls the class; how effectively the teacher clarifies key concepts and assignments; how
effectively the teacher challenges students; how well the teacher captivates students by making schoolwork interesting; whether the teacher effectively confers with students to check their understanding of the schoolwork; and whether the teacher helps students consolidate understanding by summarizing key concepts and providing feedback on student work. The version of the survey we study in this chapter has twenty-eight questionnaire items that reveal information on these 7Cs (see Table 6.1). The TRIPOD survey can be administered frequently, analyzed rather quickly, with results fed back to teachers and administrators in time to support teacher improvement efforts. TRIPOD also potentially provides detailed information on multiple aspects of classroom life. Hence, the survey has potentially important utility in assessing and improving the quality of teaching practice.

Table 6.1. Design of the TRIPOD Survey

<table>
<thead>
<tr>
<th>C Construct</th>
<th>Item Text</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care</td>
<td>I like the way my teacher treats me when I need help.</td>
</tr>
<tr>
<td>Care</td>
<td>My teacher in this class makes me feel that he/she really cares about me.</td>
</tr>
<tr>
<td>Care</td>
<td>The teacher in this class encourages me to do my best.</td>
</tr>
<tr>
<td>Care</td>
<td>My teacher gives us time to explain our ideas.</td>
</tr>
<tr>
<td>Control</td>
<td>My classmates behave the way my teacher wants them to.</td>
</tr>
<tr>
<td>Control</td>
<td>Our class stays busy and does not waste time.</td>
</tr>
<tr>
<td>Control</td>
<td>Everybody knows what they should be doing and learning in this class.</td>
</tr>
<tr>
<td>Clarify</td>
<td>My teacher explains things in very orderly ways.</td>
</tr>
<tr>
<td>Clarify</td>
<td>In this class, we learn to correct our mistakes.</td>
</tr>
<tr>
<td>Clarify</td>
<td>My teacher explains difficult things clearly.</td>
</tr>
<tr>
<td>Clarify</td>
<td>My teacher has several good ways to explain each topic that we cover in this class.</td>
</tr>
<tr>
<td>Clarify</td>
<td>I understand what I am supposed to be learning in this class.</td>
</tr>
<tr>
<td>Clarify</td>
<td>My teacher knows when the class understands, and when we do not.</td>
</tr>
<tr>
<td>Clarify</td>
<td>This class is neat—everything has a place and things are easy to find.</td>
</tr>
<tr>
<td>Clarify</td>
<td>If you don’t understand something, my teacher explains it another way.</td>
</tr>
<tr>
<td>Challenge</td>
<td>My teacher pushes everybody to work hard.</td>
</tr>
<tr>
<td>Challenge</td>
<td>In this class, we have to think hard about the writing we do.</td>
</tr>
<tr>
<td>Challenge</td>
<td>In this class, my teacher accepts nothing less than our full effort.</td>
</tr>
<tr>
<td>Captivate</td>
<td>Schoolwork is interesting.</td>
</tr>
<tr>
<td>Captivate</td>
<td>We have interesting homework.</td>
</tr>
<tr>
<td>Confer</td>
<td>When he/she is teaching us, my teacher asks us whether we understand.</td>
</tr>
<tr>
<td>Confer</td>
<td>My teacher asks questions to be sure we are following along when he/she is teaching us.</td>
</tr>
</tbody>
</table>
Assessing the Validity of TRIPOD

A critical question involves predictive validity: To what extent do student perceptions predict student learning? Evidence from the first MET report (Measurement of Effective Teaching Project, 2010) suggests an affirmative answer. MET researchers administered the TRIPOD survey in over 2,500 classrooms. To assess the validity of student perception data, they correlated measures of the 7Cs with measures of student achievement. Interestingly, the student achievement measures were gathered not from the same students who responded to the survey, but rather from a different set of students taught by the same teacher. This guaranteed that student achievement could not be the cause of the student perceptions. A key finding was that various measures of student perceptions significantly predicted student learning.

However, standard statistical procedures can’t tell us just how powerful TRIPOD measures are as predictors of achievement, nor can they tell us which aspects of student perception—that is, which of the 7Cs—are most important. We explain why in this chapter we present a new statistical method that enables us to answer these questions with some confidence. We shall use this method to summarize the overall power of TRIPOD to predict value added and to isolate those components of TRIPOD that are most predictive. The method is called the “multilevel variable selection model” or MVSM.

Conventional Analysis of TRIPOD

Using standard statistical methods, we find that each of the 7Cs measured in one year is positively associated with the achievement growth within a teacher’s classroom in a different year, as measured by the value-added score. However, when we use all seven to simultaneously predict value added, the key finding seems to be that some of these are extremely strongly positively associated with value added, while others are
significantly negatively associated with value added (controlling the others). These findings are implausible; it seems unlikely that any of the 7Cs is as powerful as the conventional results would indicate, nor does it seem reasonable that increasing some of the 7Cs has a really bad effect on children. The source of these implausible results is that the seven dimensions are highly correlated, particularly after we adjust for measurement error, and it is well known that a prediction model based on highly inter-correlated predictors becomes unstable: the coefficients associated with the seven predictors have very large standard errors and, as a result, can become very large in magnitude—positive or negative—even if the underlying true relationships are small. This is the problem of “collinearity” discussed in every textbook on multiple regression, the statistical method nearly universally used for predicting an outcome using multiple explanatory variables. When predictors are highly correlated, some predictors that truly have small positive effects can be estimated to have very harmful effects. Using standard methods, it is difficult to discern which of the 7Cs are important.

One consequence of obtaining unstable and therefore highly variable estimates of the importance of each prediction is that we will tend to overestimate the capability of the 7Cs together to predict the outcome. Thus, we are likely to overstate the predictive validity of the 7Cs.

Often, researchers will react to collinearity in ways that have the opposite fault: to understate the importance of the 7Cs. One way to do this is simply to look at one predictor at a time. Another is to create a single index, such as the mean of all of the predictors. These approaches eliminate the problem of dealing with highly inter-correlated predictors, but these approaches do not, in general, fully use all of the information in the predictors, and that is why they understate the utility of the predictors in accounting for the outcome. Moreover, these approaches don’t help us decide which information is most crucial to collect.

**Rationale for Using New Methods**

After applying our new MVSM, we will see that the implausibly large regression coefficients, both negative and positive, are “shrunk” to more plausible values, strengthening the idea that the 7Cs share a common source of variation with value added. Yet the model will enable us to read signals suggesting that some predictors are more important than others, and that some may undermine effective prediction. This more nuanced understanding, in principle, enables us to isolate a subset of predictors that maximize our ability to validly predict the outcome while eliminating redundant information in the predictors.
Our aim, then, is to arrive at the best possible summary of the information in TRIPOD. Our hope is that this will make TRIPOD more useful to teachers, while giving teachers and researchers an accurate account of just how helpful TRIPOD is in predicting student achievement gains. Finally, if we knew that a subset of TRIPOD items was particularly important as a predictor of student learning, we might be able to devise more efficient surveys of students.

This last point bears elaboration. Measures of student perceptions constitute only one source of information about classroom effectiveness. Educators are also interested in observing classrooms to rate the quality of instruction overall as well as in particular subjects such as reading and math. Supervisors may wish to look at artifacts of students’ work, such as essays or problem sets, and they may collect lesson plans and other information. The MET study itself has collected a wealth of information on each of many classrooms. It would be good to know which of these sources of information are most predictive of student learning.

So learning about which aspects of TRIPOD are most predictive of student learning is part of the larger problem of learning about how to tailor a whole system of data collection that provides useful information at reasonable cost and that avoids taking up teachers’ time with many pieces of information that might be redundant or uninformative. We believe the statistical procedures proposed here can help solve this larger problem, enabling us to combine many sources of information to devise a single model that best predicts student achievement gains.

In sum, we propose in this chapter a novel methodology for learning about which aspects of a data source like TRIPOD are most predictive of a valued outcome such as student learning. We call this the MVSM. It is “multilevel” because student perceptions vary at several levels: among students within a classroom, among classrooms within a school, and among schools. It is a “variable selection model” because it helps us select a subset of a large number of variables that is most useful in predicting a valued outcome. We test this model against an alternative approach that is a little more complicated but has a similar goal: a “multilevel principal components model.” The results are strikingly convergent: each provides similar results about how effective TRIPOD is overall in predicting student learning gains, as measured by value added, and which aspects of TRIPOD are more effective in doing so.

The MVSM appears to work well and is almost as easy to use as multiple regression, with which most analysts are already familiar. The method is applicable when we are faced with many correlated predictor variables. In this chapter, we apply this new methodology to a large-scale
data set from an urban U.S. school district. The data set combines survey data from TRIPOD with value-added statistics collected in a different year. As in MET, this design enables us to see whether student perceptions of teaching effectiveness collected one year predict how much a teacher’s students learn in a different year. The results may help inform the design and use of the TRIPOD survey and other surveys of students’ perceptions. We also anticipate that this methodology will be useful to researchers who are combining dense information from other sources, such as classroom observations, to obtain measures of teacher effectiveness validated as predicting student learning gains.

In the next section, we provide a conventional analysis of the TRIPOD data. We use standard methods to assess the reliability of TRIPOD measures, the correlations between them, and their predictive validity. The following section introduces our new method: We show how to use the MVSM, share the results of its application to TRIPOD, and contrast the findings with those based on conventional methods. To provide a check on these results, the last section of the paper prior to the Discussion section applies an alternative novel method known as “principal components regression.” This approach reproduces the results based on the MVSM, but uses a less direct way to get there. Our Discussion section summarizes our findings and makes recommendations for studies of classroom effectiveness.

A Conventional Analysis of the TRIPOD Data

We studied 405 fourth and fifth grade teachers working within ninety-five elementary schools in a large urban district. This district was one of those studied in MET, although the data were collected independently. Researchers administered the TRIPOD survey in each of these classrooms, yielding, on average, eighteen complete student surveys, consisting of twenty-eight responses to the items described in Table 6.1. To assess validity, we use value-added data collected on the same teachers during the year before collection of the TRIPOD data. Thus, the students who shared their perceptions of teacher effectiveness were independent of those whose achievement test scores supply the evidence of teacher impact on student learning.

A standard analysis will first assess the reliability of the TRIPOD measures. We’ll see that reliability is quite respectable. Next, we’d like to
know how strongly inter-correlated the 7Cs are. We find them to be very highly correlated, especially if we take into account measurement error. These high inter-correlations pose a challenge for the next step: assessment of validity. We’ll see that each of the 7Cs predicts value added, but that doesn’t tell us which of the 7Cs are most important or how strongly predictive the 7Cs are taken as a whole. A conventional multiple prediction exercise is the next logical step, but it produces highly misleading results, defining the challenge for our novel MVSM approach.

**Reliability Analysis**

Our first step was to compute the reliability of measures of each of the 7Cs and their inter-correlations. For this purpose, we used the three-level multivariate measurement model described by Raudenbush, Rowan, and Kang (1991). At the first level, item responses for each child vary randomly around child-specific “true scores,” one for each of the 7Cs. At the second level, these child-specific “true scores” vary randomly around a teacher’s mean. At the third level, teacher-specific true scores vary randomly around a global, district-wide mean. This analysis tells us how much of the variance in each of the 7Cs lies between items, between children, and between teachers. The variance between teachers provides the “signal” we are interested in, that is, the information about how much teachers vary in their effectiveness, as indicated by each of the 7Cs. The variation between items may be regarded as item inconsistency, while the variability among students within a classroom may be regarded as “rater variance.” In effect, TRIPOD casts each student within a class as an informant, or rater, of the quality of the classroom; inconsistencies among student responses within a class are therefore regarded as “rater error” and are thus part of the measurement error.

Table 6.2 summarizes the results of this analysis. The intra-teacher correlation measures how similar student responses are and is equivalent to the fraction of variation in the “true” student perceptions between teachers. Table 6.2 thus shows that 24 percent of the variation in Clarify responses lies between teachers, while 35 percent of the variation in Control variation lies between teachers. All other intra-teacher correlations lie between these two numbers. Variables that have more items and larger fractions of variation between teachers will be more reliable. The reliabilities in Table 6.2 are very similar, ranging between .74 and .81.
Table 6.2. Reliability Analysis

<table>
<thead>
<tr>
<th>Variable</th>
<th>Number of Items</th>
<th>Student-level Variance</th>
<th>Teacher-Level Variance</th>
<th>Intra-Teacher Correlation</th>
<th>Teacher-Level Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care</td>
<td>4</td>
<td>.32</td>
<td>.13</td>
<td>.28</td>
<td>.79</td>
</tr>
<tr>
<td>Control</td>
<td>3</td>
<td>.27</td>
<td>.14</td>
<td>.35</td>
<td>.81</td>
</tr>
<tr>
<td>Clarify</td>
<td>8</td>
<td>.22</td>
<td>.07</td>
<td>.24</td>
<td>.78</td>
</tr>
<tr>
<td>Challenge</td>
<td>3</td>
<td>.27</td>
<td>.10</td>
<td>.26</td>
<td>.75</td>
</tr>
<tr>
<td>Captivate</td>
<td>2</td>
<td>.49</td>
<td>.16</td>
<td>.24</td>
<td>.74</td>
</tr>
<tr>
<td>Confer</td>
<td>6</td>
<td>.26</td>
<td>.09</td>
<td>.26</td>
<td>.79</td>
</tr>
<tr>
<td>Consolidate</td>
<td>2</td>
<td>.43</td>
<td>.17</td>
<td>.29</td>
<td>.77</td>
</tr>
</tbody>
</table>

Correlations Among the “True” 7Cs

The analysis also gives us estimates of the correlations among the 7Cs. These are adjusted for measurement error and displayed in Table 6.3. We see that the 7Cs are quite highly inter-correlated. The weakest correlation of .56 is between Challenge and Control, while the largest is .95 between Clarify and Confer. These results illustrate how difficult it will be to assess the independent contributions of each of the 7Cs to the prediction of value added. It will be nearly impossible to separate the contributions of Clarify and Confer, because they appear to provide nearly the same information about teachers. However, it may well be possible to separate the contributions of Challenge and Control. The latter are correlated, but not so highly, implying that some teachers challenge their students intellectually, but are not so skilled at creating well-controlled classrooms; other classrooms are well controlled, but apparently not too challenging.

Table 6.3. Correlations Among 7Cs After Correction for Measurement Error

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Clarify</th>
<th>Challenge</th>
<th>Captivate</th>
<th>Confer</th>
<th>Consolidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care</td>
<td>.73</td>
<td>.93</td>
<td>.63</td>
<td>.68</td>
<td>.89</td>
<td>.84</td>
</tr>
<tr>
<td>Control</td>
<td>.81</td>
<td>.56</td>
<td>.62</td>
<td>.75</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Clarify</td>
<td></td>
<td>.76</td>
<td>.74</td>
<td>.95</td>
<td>.85</td>
<td></td>
</tr>
<tr>
<td>Challenge</td>
<td></td>
<td></td>
<td>.57</td>
<td>.79</td>
<td>.69</td>
<td></td>
</tr>
<tr>
<td>Captivate</td>
<td></td>
<td></td>
<td></td>
<td>.73</td>
<td>.76</td>
<td></td>
</tr>
<tr>
<td>Confer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>.87</td>
</tr>
</tbody>
</table>

Assessing Validity

The next step in a standard analysis is to study validity. High reliability (see Table 6.2) is a necessary, but not sufficient, condition for
validity. To be valid, a measure must minimally predict a criterion, that is, a known measure of teacher effectiveness. We choose as our criterion a teacher value added to student math learning, derived from a sample of each teacher’s children in a different year from the year in which the TRIPOD data were collected.

There are three conventional ways of assessing validity. One is to assess the explanatory power of each of the 7Cs separately. We call these univariate model estimates. A second approach is to bite the bullet and include all of them in the regression, encountering the collinearity problem. A third is to combine all seven into a single scale or index—or possibly into two or three scales—using principal components or factor analysis. We’ll show how each of these approaches works, starting with univariate models.

**Univariate Prediction**

We estimate the simple linear prediction model

\[ Y_{ij} = \beta_p x_{pij} + u_j + r_{ij}. \]  

where \( Y_{ij} \) is the value-added score for teacher \( i \) in school \( j \); the predictor \( x_{pij} \) is one of the 7Cs \((x_{1ij} = \text{(care)}_y, x_{2ij} = \text{(control)}_y, \ldots x_{7ij} = \text{(consolidate)}_y)\); \( \beta_p \) is a regression coefficient—which indicates the strength of association between the predictor and the outcome; \( u_j \) is a school-level random effect having 0 mean and variance \( \tau^2 \); and \( r_{ij} \) is a teacher random effect having 0 mean and variance \( \sigma^2 \). \(^1\) The outcome and each predictor are standardized to have means of 0 and standard deviations of 1.0.

Table 6.4 provides the results. The first column omits all predictors and tells us how much of the variance in value added lies within and between schools. We see that 17 percent of the variance lies between schools and 83 percent lies within schools. The next column tells us that the first of the 7Cs (Care) significantly positively predicts value added, with a coefficient (equivalent here to a correlation coefficient) of 0.169 and a standard error of 0.049. After controlling for Care, the variance between and within schools is slightly reduced (to .17 and .81, respectively). The overall explanatory power of Care is \( R^2 = 1 - \tau^2 - \sigma^2 \) and is estimated thus to be 1-.17-.81 = .02 or 2.0 percent. The other columns are similarly interpreted. Note that the “C” with the largest coefficient is Challenge (coefficient of 0.256), while the second-largest is associated with Control (0.236). Captivate has the smallest coefficient (0.102). All achieve conventional levels of statistical significance (each is at least twice its
standard error). The maximum explained variance (for Control) is about .059 or 5.9 percent.

Table 6.4. Univariate Regressions Predicting Value Added

<table>
<thead>
<tr>
<th>Null Model</th>
<th>Care</th>
<th>Control</th>
<th>Clarify</th>
<th>Challenge</th>
<th>Captivate</th>
<th>Confer</th>
<th>Consolidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care</td>
<td>.169</td>
<td>.236</td>
<td>.225</td>
<td>.250</td>
<td>.102</td>
<td>.214</td>
<td>.143</td>
</tr>
<tr>
<td></td>
<td>(.049)</td>
<td>(.048)</td>
<td>(.049)</td>
<td>(.049)</td>
<td>(.051)</td>
<td></td>
<td>(.050)</td>
</tr>
<tr>
<td>Control</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Clarify</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Challenge</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Captivate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Confer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consolidate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\hat{\tau}^2$</td>
<td>.19</td>
<td>.17</td>
<td>.18</td>
<td>.17</td>
<td>.18</td>
<td>.20</td>
<td>.18</td>
</tr>
<tr>
<td>$\hat{\sigma}^2$</td>
<td>.81</td>
<td>.81</td>
<td>.78</td>
<td>.79</td>
<td>.78</td>
<td>.81</td>
<td>.79</td>
</tr>
</tbody>
</table>

These results are certainly interesting and corroborate the findings of the MET study (Bill & Melinda Gates Foundation, 2010), which also found that student perceptions predict value added. Indeed, each of the 7Cs taken separately significantly predicts value added. The explained variation is modest, but the coefficients are non-negligible in every case.

The problem is that the univariate analysis tells us nothing about how well the 7Cs taken together predict value added, nor does it supply any information about the relative importance of the 7Cs. To answer these questions, it is standard to formulate a multiple prediction model, the topic to which we now turn.

**Multivariate Prediction**

We now expand Equation (1) to include all 7Cs as predictors of the outcome. The model is thus:
\[ Y_{ij} = \beta_1(\text{care})_{ij} + \beta_2(\text{control})_{ij} + \beta_3(\text{clarify})_{ij} + \beta_4(\text{challenge})_{ij} \\
+ \beta_5(\text{captivate})_{ij} + \beta_6(\text{confer})_{ij} + \beta_7(\text{consolidate})_{ij} + u_j + r_{ij} \]

\[
= \sum_{p=1}^{7} \beta_p x_{prij} + u_j + r_{ij},
\]

where the symbols take on the same meaning as in Equation 1.²

Results are in Table 6.5 (Column 1). One notable feature of these results is the extraordinarily large standard errors, as compared to those in Table 6.4. In the worst case, for Clarify, the standard error of .372 is 76 times larger than the standard error in the univariate model (Table 6.4)! Closely related, the coefficients estimates in Table 6.5 are highly variable, ranging from a minimum of about –.23 (for Care and Captivate) to a maximum of .55 for Clarify. Yet only the coefficients for Challenge and Captivate achieve a nominal level of statistical significance. Finally, if we were to believe these results, we would conclude that the model explains about 10.8 percent of the variance in value added, almost double the explained variance of 5.9 percent in the most predictive model in Table 6.4. All of these anomalies are attributable to the collinearity problem: standard regression methods provide extremely imprecise estimates of coefficients when predictors are highly correlated (please recall the high inter-correlations noted in Table 6.3). These nasty features of conventional prediction often lead analysts to abandon multiple prediction and apply the third conventional approach: combining all measures into a single composite score.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Care</td>
<td>−.230 (.203)</td>
<td>−.049 (.078)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>.146 (.105)</td>
<td>.179 (.072)</td>
<td>.244 (.069)</td>
<td>.219 (.069)</td>
</tr>
<tr>
<td>Clarify</td>
<td>.550 (.372)</td>
<td>.131 (.062)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Challenge</td>
<td>.202 (.100)</td>
<td>.197 (.071)</td>
<td>.251 (.067)</td>
<td>.228 (.068)</td>
</tr>
<tr>
<td>Captivate</td>
<td>−.227 (.092)</td>
<td>−.160 (.076)</td>
<td>−.247 (.074)</td>
<td>−.198 (.073)</td>
</tr>
<tr>
<td>Confer</td>
<td>−.131 (.257)</td>
<td>.039 (.078)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Consolidate</td>
<td>-.100 (,.134)</td>
<td>-.106 (.084)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>------------</td>
<td>-------------</td>
<td>-------------</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>$\hat{\tau}^2$</td>
<td>.14</td>
<td>.15</td>
<td>.15</td>
<td>.15</td>
</tr>
<tr>
<td>$\hat{\sigma}^2$</td>
<td>.77</td>
<td>.76</td>
<td>.77</td>
<td>.76</td>
</tr>
<tr>
<td>$\hat{\Delta}^2$</td>
<td></td>
<td>.03</td>
<td></td>
<td>.04</td>
</tr>
</tbody>
</table>

**Composite Score**

The third approach would combine the 7Cs into one single index. The problem with this approach is a potential loss of information. Despite the high correlations among the 7Cs, it might be the case that certain Cs carry uniquely important information, and that teachers would especially benefit from improving the corresponding aspects of their teaching. However, we defer discussion of the composite approach until we come to the section of our chapter on principal components analysis.

**The Multilevel Variable Selection Model (MVSM)**

Recall that the univariate regressions (Table 6.4) give sensible estimates of the predictive validity of each of the 7Cs, taken one at a time. However, because each of these regressions uses only one piece of information generated by TRIPOD, we suspect that these models underestimate the predictive power of TRIPOD as a whole. Therefore, the maximum explanatory power of these models ($R^2 = 5.9$ percent) is a lower bound on the predictive power of TRIPOD as a whole. Moreover, the seven regressions in Table 6.4 give us no information on how best to combine information from TRIPOD to explain student growth in achievement.

In contrast, the regression with seven predictors (Column 1, Table 6.5) yields coefficient estimates that appear greatly exaggerated in absolute value. It seems that the very large standard errors caused by collinearity create large chance differences in the coefficient estimates. One result is that our interpretation of the relative importance of each of the 7Cs is distorted. A second result is that the overall explanatory power of the model, which increases with the absolute value of the regression coefficients, is exaggerated. Therefore, we must regard the explanatory power of this model ($R^2 = 10.8$ percent) to be an upper bound of the explanatory power of TRIPOD for these value-added outcomes.
How, then, can we obtain estimates of the predictive power of the 7Cs that are plausible and that produce a realistic assessment of the explanatory power of TRIPOD, which must lie somewhere between $R^2 = 5.9$ percent and $R^2 = 10.8$ percent?

For this purpose, we adopt a special case of “ridge regression” (Hoerl & Kennard, 1970). Technically, ridge regression adds a small positive constant, $k$, to the sum of squares of each explanatory variable in the computation of the regression coefficients (see the Appendices for details). This stabilizes estimation and “shrinks” unreliable coefficient estimates toward 0. An interesting feature of ridge regression is that it induces a small bias in regression coefficients, while significantly reducing the sampling variance of the estimates. As a result, we can readily prove that the ridge-based coefficient estimates will be more accurate (have smaller expected mean squared error) than will the coefficients estimated by least squares (Lindley & Smith, 1972). The question naturally arises: How do we select the size of $k$? Lindley and Smith proposed a solution using a method known as “empirical Bayes” that is now increasingly used in the social sciences. This approach is quite elegant statistically and is based on a rationale that makes considerable intuitive sense in the context of our example.

As mentioned, we have strong reason to suspect that the magnitudes (absolute values) of the seven regression coefficients generated by the conventional multiple regression (Column 1, Table 6.5) are exaggerated, and we know why: collinearity among the predictors has induced great uncertainty about these coefficients, as reflected in the huge standard errors we see in the table. Lindley and Smith reasoned that, although some large part of the variation in these estimated coefficients represents random error (“noise”), some part of the variation in these estimates reflects variation in the true regression coefficients. If we can estimate the amount of noise in these estimates, we might then “back out” the variance of the true coefficients, call it $\Delta^2$. This provided Lindley and Smith with an answer to the question, “How large should the value of $k$ be in ridge regression?” Lindley and Smith’s reasoning led them to derive the optimal value of $k$ as

$$ k = \frac{\sigma^2}{\Delta^2} \, . \ (3) $$

Equation 3 says that the $k$ should be inversely proportional to $\Delta^2$, the variance of the true values of the coefficients, $\beta_p$, $p = 1, \ldots, 7$. (The teacher-level variance $\sigma^2$ is just a scaling factor; think of it as the constant of proportionality that simply depends on the units of the outcome.) If the coefficients in the seven-predictor model vary mostly because of noise—that is, the true coefficients are nearly equal—$\Delta^2$ will be very small, $k$ will be large, and all of the coefficients in Equation 2 will be “shrunk” toward a
common value. That value should presumably be the value we would obtain if we created a single index of the 7Cs—an average—and used that average as a predictor. (This is essentially what we obtain if we use the first principal component of the 7Cs as a predictor; see below). In contrast, if the coefficients in the seven-predictor model vary in part because the true coefficients are, in fact, highly variable, \( \hat{\Delta}^2 \) will be large, \( k \) will be small, and the solution will look very much like that given by the conventional regression model. We develop this idea mathematically in the Appendices; our contribution is original in one way: We have integrated Lindley and Smith’s approach into a hierarchical linear model that represents the fact that teachers are nested within schools. This enables us to simultaneously estimate how much variation lies within and between schools; and we have the option, not employed here for simplicity, of allowing the regression coefficients to randomly vary over schools, based on a theory that schools may be heterogeneous in the extent to which value added predicts learning.

The results of our MVSM analysis are given in Column 2 of Table 6.5. The evidence suggests that the true regression coefficients vary modestly, \( \hat{\Delta}^2 = 0.03 \), meaning that most of the variation in the estimates shown in Column 1 is noise. As a result, the coefficient estimates based on MVSM are much closer to 0 than those based on the conventional analysis of Column 1. In particular, the coefficient for Clarify is shrunk from .550 to .131. Moreover, the standard errors based on MSVM have “calmed down” considerably. None exceeds 0.084—still larger than in the univariate regressions, but much smaller than in the conventional seven-variable regression.

Finally, the explanatory power of the model is now estimated to be \( R^2 = 7.8 \) percent—about midway between the lower bound of 5.9 percent, based on the univariate models and the upper bound of 10.8 percent, based on the conventional seven-variable regression. Interestingly according to the MVSM, one source of unexplained variation is the variability \( \hat{\Delta}^2 = 0.03 \) in the unknown true regression coefficients. We can think of this component as reflecting uncertainty that arises from collinearity in the 7Cs.

These results would suggest that an academically challenging environment within a well-controlled classroom significantly positively predicts value added (see coefficients for Challenge and Control). Holding these constant, attempts to “captivate” students by making schoolwork and homework “interesting” negatively predict value added.

---

**Multilevel Principal Components**
Regression

We had mentioned that a strategy for eliminating collinearity among multiple predictors is to combine them into a single index. We can use a simple average. A more sophisticated version of this approach is to transform the original 7Cs into seven uncorrelated “principal components,” which we shall call the “7 Ss,” where each S is a weighted average of the 7Cs. The standard approach is to rank order the principal components in order of their variances, known as Eigenvalues in the language of this methodology. We display the results in Table 6.6. The table shows the 7 Ss, each a linear combination of the 7Cs. The set of coefficients that transform each of the 7Cs into each S is called an “Eigenvector.” Notice that the Eigenvector for the first S contains the values (.39, .35, .40, .35, .36, .40, .39). These numbers are very similar, so we see that the first S is, in essence, proportional to the mean of the 7Cs.

Table 6.6. Principal Components Transformation of the 7Cs

<table>
<thead>
<tr>
<th>Transformation of original 7Cs (Eigenvector * 7C vector)</th>
<th>Variance (Eigenvalue)</th>
<th>Coefficient, $\hat{\beta}_p$, (se)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1 = .39 * \text{care} + .35 * \text{control} + .40 * \text{clarify} + .35 * \text{challenge} + .36 * \text{captivate} + .40 * \text{captivate} + .39 * \text{consolidate}$</td>
<td>5.88</td>
<td>.082 (.021)</td>
</tr>
<tr>
<td>$S_2 = -.17 * \text{care} -.61 * \text{control} -.07 * \text{clarify} + .76 * \text{challenge} -.07 * \text{captivate} + .11 * \text{captivate} -.06 * \text{consolidate}$</td>
<td>.38</td>
<td>.057 (.079)</td>
</tr>
<tr>
<td>$S_3 = -.04 * \text{care} + .50 * \text{control} + .15 * \text{clarify} + .35 * \text{challenge} -.70 * \text{captivate} + .08 * \text{captivate} -.32 * \text{consolidate}$</td>
<td>.34</td>
<td>.415 (085)</td>
</tr>
<tr>
<td>$S_4 = -.56 * \text{care} + .43 * \text{control} -.16 * \text{clarify} + .30 * \text{challenge} + .54 * \text{captivate} -.19 * \text{captivate} -.23 * \text{consolidate}$</td>
<td>.23</td>
<td>.091 (.095)</td>
</tr>
<tr>
<td>$S_5 = -.35 * \text{care} + .20 * \text{control} -.28 * \text{clarify} -.02 * \text{challenge} -.29 * \text{captivate} -.06 * \text{captivate} + .82 * \text{consolidate}$</td>
<td>.11</td>
<td>-.057 (.145)</td>
</tr>
<tr>
<td>$S_6 = .51 * \text{care} + .10 * \text{control} -.16 * \text{clarify} + .29 * \text{challenge} + .02 * \text{captivate} -.78 * \text{captivate} + .09 * \text{consolidate}$</td>
<td>.04</td>
<td>-.043 (.229)</td>
</tr>
<tr>
<td>$S_7 = .33 * \text{care} + .14 * \text{control} -.82 * \text{clarify} + .06 * \text{challenge} + .05 * \text{captivate} + .41 * \text{captivate} -.13 * \text{consolidate}$</td>
<td>.01</td>
<td>-.548 (.397)</td>
</tr>
</tbody>
</table>

The second S assigns a large positive weight to Challenge and a large negative weight to Control. In our earlier discussion of the correlation matrix in Table 6.3, we noted that some teachers were high on Control but low on Challenge. The second S will assign a large value to such teachers. The third S appears to distinguish teachers who are high on Control and...
Challenge, but low on Captivate. Note that the first S has a much larger Eigenvalue (variance) than does any other S. Moreover, the Eigenvalues diminish as we look down the table from S1 to S7.

Having clarified the definition of each principal component, let’s now use the 7 Ss to predict value added using the model

$$Y_j = \sum_{p=1}^{7} \theta_p s_{pij}^* + u_j + r_y, \quad (4)$$

where $s_{pij}^*$, $p = 1,...,7$ is principal component $p$ associated with the empirical Bayes estimates $x_{pij}^*$, $p = 1,...,7$. The results are in the last column of Table 6.7. Note that the first S (recall this is essentially the mean of the 7 Ss) is highly statistically significantly predictive of value added, with a t-ratio of $0.082/0.021=3.90$. The very small standard error associated with this first S results from its large Eigenvalue. Predictors that have large variance provide more precise estimates of regression coefficients than do those with small variance. The second S, which distinguishes teachers whose classrooms have high Control and low Challenge, is not different from 0. This is consistent with our earlier results suggesting that Control and Challenge are both important in predicting value added. The third S, representing teachers who are high on Challenge and Control but low on Captivate, is positive and statistically significant, again consistent with our earlier results, which assigned positive coefficients to Challenge and Control and a negative coefficient to Captivate net the contribution of the other two (see Table 6.5, Columns 2 and 3). None of the other Ss contribute significantly to the prediction of value added.

Note the dramatic inflation of standard errors in Table 6.7 for Ss that have small Eigenvalues. In particular, the standard error for the coefficient for S7 is .397, similar to the largest standard error we obtained using conventional regression (Table 6.5, Column 1)! This result helps us to see how principal components regression works. Using principal components, we have solved the problem of collinearity—correlations among all of the 7 Ss are zero. However, we have “traded off” the problem of collinearity for the problem of small variance in some of the predictors. The data provide no leverage for estimating the contribution of Ss that have small Eigenvalues, as indicated by their large standard errors.

This is a perfect tradeoff, as shown in the first two columns of Table 6.7. It is straightforward to translate the $\hat{\theta}_p$ coefficients from Equation (5) (and Table 6.6) back into the corresponding $\hat{\beta}_p$ values. As shown in the first two columns of Table 6.7, the translation perfectly reproduces the results of the conventional regression using the 7 Cs.
Table 6.7. Principal Components Regression and “Back-Translation” to 7Cs

<table>
<thead>
<tr>
<th></th>
<th>Conventional Least Squares</th>
<th>Principal Components: Back Translation to 7Cs</th>
<th>Trimmed model: Ridge Regression (Empirical Bayes)</th>
<th>Trimmed Model: Principal Components: Back Translation to Cs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care</td>
<td>-.230</td>
<td>-.230</td>
<td>.005</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>.146</td>
<td>.146</td>
<td>.219</td>
<td>.202</td>
</tr>
<tr>
<td>Clarify</td>
<td>.550</td>
<td>.550</td>
<td>.090</td>
<td></td>
</tr>
<tr>
<td>Captivate</td>
<td>-.227</td>
<td>-.227</td>
<td>-.198</td>
<td>-.266</td>
</tr>
<tr>
<td>Confer</td>
<td>-.131</td>
<td>-.131</td>
<td>-.074</td>
<td></td>
</tr>
<tr>
<td>Consolidate</td>
<td>-.100</td>
<td>-.100</td>
<td>-.098</td>
<td></td>
</tr>
</tbody>
</table>

While the principal component method trades off the collinearity problem for a small variance problem, the small variance problem is easy to solve: simply eliminate from the model those components that have small variance and, correspondingly, large standard errors! We therefore recomputed the model (Equation 4) using only the first three Ss. We then translated the results back and compared them to what we obtained using our new MVSM. These are shown in the last two columns of Table 6.7. We see that the results are extremely similar: large positive contributions for Control and Challenge; controlling for these, we see a negative contribution of Captivate.

Discussion

A key challenge in assessing classrooms is to clarify the predictive validity of a large number of explanatory variables. This arises because researchers increasingly have the option of collecting data on student perceptions, classroom observations, and other sources. There is potential to collect a great deal of information, some of which is redundant or otherwise uninformative. We’d like to know which aspects of classroom life as measured by these methods is most useful in predicting student learning, and we’d like to know just how predictive the entire ensemble of information might be. We have provided two methods for achieving these aims: a multilevel variable selection model and a multilevel principal components regression model.

When we applied these two methods to the TRIPOD data, we found that the two methods gave us very similar results. TRIPOD, taken as a
whole, explains about 7.8 percent of the variation in learning gains achieved by students in a different year from the year in which the student perceptions were collected. This is equivalent to a correlation between predicted and observed value added of \( r = .29 \). (This underestimates the correlation between TRIPOD and the “true” value added because the latter is measured with error.)

We found that teachers whose classrooms are well controlled and intellectually challenging produce comparatively large learning gains. Controlling for control and challenge, teachers who “captivate” their students, as indicated by TRIPOD, actually do a little worse than those who do not. This latter result seems a little puzzling. Captivate is measured by two items: “Schoolwork is interesting” and “We have interesting homework.” Why would affirmative answers to these questions produce negative associations with value added—after adjusting for Control and Challenge? One possibility is that student perceptions that school and homework are interesting is a consequence of Control and Challenge. Teachers who challenge their students intellectually in well-controlled classrooms may make school interesting. We generally do not wish to control for consequences of causes in regression. Regression tells us what happens to an outcome when we increase the value of a predictor, holding constant other predictors.

The idea of increasing Control and Challenge while holding constant Captivate may not make sense. Alternatively, it may be that teachers who are good at making things interesting without increasing Control and Challenge are comparatively ineffective. We do not intend to adjudicate such explanations here, but rather to suggest that careful probing of predictive validity using these methods may trigger a useful re-examination of certain details of the classroom assessment procedure. More generally, we expect that this methodology will be useful in assessing the predictive validity of a wide range of indicators of classroom quality, and the results may lead to improvements in how these indicators are conceived and used.

**NOTES**

1. Our method accounts for measurement error of each of the 7Cs (see Note 2).
2. Of course, we do not observe the true values \( x_{pij}, \quad p = 1, \ldots, 7 \) of the 7 Cs, so we cannot estimate the theoretical model (1) from the data. A commonly used option is to simply substitute the observed values, which we denote \( X_{pij}, \quad p = 1, \ldots, 7 \) into Equation 1, then to estimate the model. These observed values are the simple mean responses of the students in each classroom. We know, however, that when observed values \( X \) measure true values \( x \) with error, such a procedure will give us
biased estimates of $\beta_1, \beta_2, \ldots, \beta_7$. Hence, our estimate of $R^2$, the explanatory power of the model, will also be biased toward 0.

To solve this problem, we substitute a multivariate version of Truman Kelley’s (1925) “estimated true score” for the unobserved true values, $x_{ijpq}$, $p = 1, \ldots, 7$.

Conditional on the values $X_{ijpq}$, $p = 1, \ldots, 7$ that we actually do observe, the expected value of our outcome based on (1) becomes

$$E(Y_{ij} \mid X_{ij}, \ldots, X_{pq}) = \sum_{p=1}^{7} \beta_p E(x_{ijpq} \mid X_{ij}, \ldots, X_{pq})$$

where $x_{ijpq} = E(x_{ijpq} \mid X_{ij}, \ldots, X_{pq})$ is the conditional mean of the true score $x_{ijpq}$ given the observed scores $X_{ij}, \ldots, X_{pq}$. These can be estimated via the empirical Bayes method (Morris, 1983), but now using the multivariate approach (Raudenbush, Rowan, & Kang, 1991), sometimes known as “multivariate shrinkage.” Substituting the empirical Bayes estimates $x_{ijpq}^*$ for $x_{ijpq}$ in Equation 1 “identifies” our model—that is, equates the parameters of (2), which we can, in principal, estimate, with the parameters of (1), which is a theoretical model because we do not observe $x_{ijpq}$. Actually, our solution to the problem of measurement error—namely, the substitution of $x_{ijpq}^*$ for $x_{ijpq}$, actually makes the collinear problem a little worse. The reason is that the correlations among the empirical Bayes estimates are even a little higher than are the correlations among the true values of the 7Cs. This is a well-known result (Raudenbush, 1999): each of the observed values of the 7Cs carries information about the other values of the 7Cs; multivariate shrinkage fully exploits this information, inducing, however, more dependence among the empirical Bayes estimates. We see this in Table 6.4, which displays correlations among the empirical Bayes estimates $x_{ijpq}^*$, $p = 1, \ldots, 7$.

These correlations now range from .63 to .97, a bit higher than in Table 6.3, meaning that the collinearity problem is now slightly more severe than indicated in Table 6.3. The reason Tables 6.2 and 6.3 are not more discrepant is that the 7Cs are measured fairly reliably. Those using our methods can expect $x_{ijpq}^*$ to be much more collinear than $x_{ijpq}$ when the reliability is lower. (See the chart below.)

<table>
<thead>
<tr>
<th></th>
<th>Control</th>
<th>Clarify</th>
<th>Challenge</th>
<th>Captivate</th>
<th>Confer</th>
<th>Consolidate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Care</td>
<td>.77</td>
<td>.96</td>
<td>.71</td>
<td>.77</td>
<td>.93</td>
<td>.88</td>
</tr>
<tr>
<td>Control</td>
<td>.86</td>
<td>.63</td>
<td>.68</td>
<td>.80</td>
<td>.73</td>
<td></td>
</tr>
<tr>
<td>Clarify</td>
<td></td>
<td>.81</td>
<td>.81</td>
<td></td>
<td>.97</td>
<td>.89</td>
</tr>
<tr>
<td>Challenge</td>
<td>.67</td>
<td></td>
<td>.85</td>
<td>.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Captivate</td>
<td></td>
<td></td>
<td></td>
<td>.80</td>
<td>.83</td>
<td></td>
</tr>
</tbody>
</table>
References


Appendix A: Multivariate Shrinkage to Remove Measurement Error in Predictors

Equation 1 can be written as

\[ Y = \beta_0 + \beta^T \pi + \varepsilon \quad (2) \]

where \( \beta \) is a 7 by 1 vector of regression coefficients, \( \pi \) is a 7 by vector of “true scores” (that is, the latent, true values of each of the 7 Cs), \( \beta_0 \) is a fixed intercept, and \( \varepsilon \) is a random disturbance composed of a school-level and a classroom-level component. We do not observe the true values \( \pi \), but instead observe estimates \( \hat{\pi} \) based on a survey using twenty-eight items. Following Raudenbush and Sadoff ((2008) we can estimate \( \beta \) without bias by conditioning (2) on the estimates \( \hat{\pi} \):

\[ E(Y \mid \hat{\pi}) = \beta_0 + \beta^T E(\pi \mid \hat{\pi}) + E(\varepsilon \mid \hat{\pi}) \]

\[ = \beta_0 + \beta^T \pi^* \quad (3) \]

where \( \pi^* = E(\pi \mid \hat{\pi}) \) can be computed by estimated the joint distribution of \( \pi, \hat{\pi} \) under multivariate normality and \( E(\varepsilon \mid \hat{\pi}) = E(\varepsilon) = 0 \).
under the assumption that the model error $\varepsilon$ is independent of the 7Cs. We can relax the latter assumption by including fixed effects of schools and adding student covariates.

---

**Appendix B. Ridge Regression with Empirical Bayes for Two-Level Data**

**B.1. Exchangeability Within Regressions**

**The Model**

Let us begin with the standard OLS regression model

$$Y_i = \beta_0 + \sum_{p=1}^{P} \beta_p x_{pi} + e_i \equiv \beta_0 + \mathbf{x}_i^T \beta + e_i, \quad e_i \sim iid \ N(0, \sigma^2), \quad (1)$$

for $i = 1, \ldots, n$ where $Y_i$ is a continuous outcome, in our case the value-added score for teacher classroom $i$, hypothesized to be a linear function of $P$ known covariates $x_{i1}, \ldots, x_{iP}$, elements of the $P$ by 1 vector $\mathbf{x}_i$, plus an additive random disturbance term $e_i$ assumed independently and identically distributed with mean 0 and variance $\sigma^2$. We can also stack these equations to represent the model in matrix notation, yielding

$$Y = \mathbf{1}_n \beta_0 + \mathbf{X} \beta + e, \quad e \sim N_n(0, \sigma^2 \mathbf{I}_n), \quad (2)$$

where $Y$ is an $n$ by 1 vector of outcomes, $\mathbf{X}$ is the $n$ by $P$ matrix of predictors, $\mathbf{1}_n$ is an $n$ by 1 vector having elements equal to unity, and $e$ is a random disturbance term. Hence $\beta_0$ is a scalar intercept and $\beta$ is a vector of coefficients to be estimated.

We know that the OLS estimator (given the intercept $\beta_0$)

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T (Y - \mathbf{1}_n \beta_0) \quad (3)$$

is “best linear unbiased,” and, in the case where $e$ is multivariate normal, OLS provides the unique, minimum variance unbiased estimator. However, the problem we face is that $P$ may be large, so large that $\mathbf{X}^T \mathbf{X}$ may be ill-conditioned; even if $\mathbf{X}^T \mathbf{X}$ is non-singular, the OLS estimates may be so noisy that they become uninterpretable and non-
replicable in new samples. For example, considering the TRIPOD data alone, there are at least twenty-eight items in some versions of the survey, each of which could become a regressor, and they are positively inter-correlated. Even with a large $n$, OLS estimates, while computable, will likely be unstable across samples, so that any attempt to rank these OLS estimates in importance may be futile. In the illustrative example here, we used seven scales, rather than twenty-eight item responses, as predictors, but the method allows a very large number of predictors, and we develop that idea here.

Rather than specify, say, $P = 28$ regressors, one for each item in the TRIPOD survey, for example, we might impose a radical simplification: combine these twenty-eight items into a single mean of all $P$ items and regress the outcome on this mean, yielding a single regression coefficient, call it $\delta_0$. This strategy would ensure stable estimation, but would prohibit us from learning anything about the relative importance of the twenty-eight items in predicting the value-added outcome.

**A Compromise**

So far we are faced with a choice between conceiving the predictor of value added to have twenty-eight dimensions versus one. A principled compromise is to impose an exchangeable prior distribution on $\beta$ of the form

$$\beta = \mathbf{1}_p \delta_0 + \nu \sim N(\mathbf{1}_p \delta_0, \Delta^2 I_p). \quad (4)$$

We can think, as Bayesians, that $\delta_0$ is our best *a priori* guess about the value of any specific coefficient $\beta_p$ in the $P$ by 1 vector $\beta$, while $\Delta^2$ represents the degree of uncertainty we have about the proposition that $\beta_p$ is near $\delta_0$. A frequentist interpretation is that, in the case of TRIPOD, the twenty-eight items represent a sample from a large universe of items that measure the quality of the classroom climate; $\delta_0$ is the population mean of the coefficients associated with these items; and $\Delta^2$ is the population variance of those coefficients.

So our linear model now follows from substituting (4) into (3)

$$Y = \mathbf{1}_p \beta_0 + \mathbf{X} \mathbf{1}_p \delta_0 + \mathbf{X} \nu + e \sim N_p(\mathbf{1}_p \beta_0 + \mathbf{X} \mathbf{1}_p \delta_0, V), \quad V = \Delta^2 \mathbf{X} \mathbf{X}^T + \sigma^2 I_p. \quad (5)$$
The New Estimator

The maximum likelihood estimator of the parameters of (5), given the variance-covariance parameters and \( \beta_0 \), is

\[
\hat{\delta}_0 = \left( I_p X^T V^{-1} X I_p \right)^T \left( I_p X^T V^{-1} Y - I_p \beta_0 \right), \tag{6}
\]

while the empirical Bayes posterior mean of the exchangeable coefficients, given the MLE is

\[
\beta^* = I_p \hat{\delta}_0 + \left( X^T X + \sigma^2 \Delta^2 I_p \right)^{-1} X^T Y - I_p \beta_0 \). \tag{7}
\]

This new estimator (7) lies on a continuum between the two extremes we discussed above. Suppose, for example, that the twenty-eight TRIPOD items each contribute uniquely to the prediction of value added. Then the heterogeneity among them would be large, so that \( \Delta^2 \) will be very small and (6) will converge to the OLS estimator (3) using the full complement of twenty-eight predictors. In contrast, suppose that, after controlling for the mean of the twenty-eight items, the individual items make no additional contribution. In this case \( \Delta^2 \to \infty \), \( V = \sigma^2 I_p \) and (7) converges to

\[
\beta^* = I_p \hat{\delta}_0 + \left( X^T X + \sigma^2 \Delta^2 I_p \right)^{-1} X^T (Y - I_p \beta_0) / P \), \tag{8}
\]

where \( \bar{X} \) is the mean of the \( P = 28 \) items. So every element of \( \beta^* \) is now a constant and is proportional to what we would obtain by simply combining the twenty-eight items into a single mean. We therefore see how \( \beta^* \) locates our inference on a continuum between the most elaborate and the most parsimonious model.

An important feature of \( \beta^* \) is that it will exist, even when \( P \) is so large that \( X^T X \) becomes singular. This results from the augmentation of \( X^T X \) by the prior variance ratio \( \sigma^2 \Delta^2 I_p \) in (7). However, if \( X^T X \) is non-singular, the OLS estimate will exist, and we can gain insight by rewriting (7) as

\[
\beta^* = I_p \hat{\delta}_0 + \left( X^T X + \sigma^2 \Delta^2 I_p \right)^{-1} X^T \hat{X} \hat{\beta} \\
= \Lambda \hat{\beta} + (I_p - \Lambda) I_p \hat{\delta}_0 \tag{9}
\]

where \( \Lambda \) is a multivariate “reliability matrix” converging to \( I_p \) when the OLS estimate is estimated precisely (e.g., the \( P = 28 \) coefficients are heterogeneous and/or the sample size is large).
It is straightforward to elaborate the structure imposed on the regression coefficients using the model

$$\beta = z\delta + v \sim N(\delta_0, \Delta^2 I_p),$$

(10)

where \(z\) is a matrix of predictors. Equation (5) is a special case in which \(z = 1_p\). For example, we might assume a priori that the twenty-eight items in TRIPOD measure two constructs, so that \(\delta\) would be a 2 by 1 vector. The general model thus extends (5) to become

$$Y = 1_n \beta_0 + Xz\delta + Xv + e \sim N_p(1_n \beta_0 + Xz\delta, V), \quad V = \Delta^2 XX^T + \sigma^2 I_p.$$  

(11)

**B.2. Clustering by School: Exchangeability Between Regressions**

We now confront the fact that the MET involves clustering by school. We may want to represent school differences by means of fixed effects or random effects, with or without randomly varying regression coefficients (Raudenbush, 2009). The random coefficient model allows the association between school quality indicators in \(X\) and the outcome \(Y\) to vary by school. Such variation may be partly (or entirely) predictable on the basis of school characteristics or partly (or entirely) random. For example, the associations between student perceptions and outcomes may vary by the level of the school (elementary versus secondary) or by the overall level climate of the school. Given randomization of teachers within schools, as in the MET study, an interesting model has fixed intercepts and random coefficients. To allow a flexible and comprehensive range of options, we shall allow the regressions to be partly predictable and partly exchangeable between schools. This involves an elaboration of our basic model (Equation 11).

Denote the value added outcome for classroom \(i\) in school \(j\) as \(Y_{ij}\) for teachers \(i = 1,...,n_j\) within schools \(j = 1,...,J\). Stack these outcomes within school \(j\) to define the outcome vector \(Y_j\). The predictor variables are elements measured by classroom observations, student perceptions, and other sources, represented in the \(n_j\) by \(P\) matrix \(X_j\). Within school \(j\), we represent the outcome as a standard linear regression

$$Y_j = 1_{n_j} \beta_{0j} + X_j \beta_j + e_j, \quad e_j \sim N_{n_j}(0, \sigma^2 I_{n_j})$$

(12)

where \(\beta_j\) is a \(P\) by 1 vector of school-specific coefficients.
The problems we face are that \( P \) may be very large and \( n_j \) may be very small. Thus, Equation (1), which specifies \( JP \) coefficients, will not be estimable without some restrictions on the parameters. To reduce the dimensionality of \( \beta_j \) within each school, we adopt exchangeability within regressions, as above:

\[
\beta_j = z\delta_j + \nu, \quad \nu \sim N(0, \Delta^2 I_p) \quad (13)
\]

where \( \delta_j \) is a \( Q \)-dimensional coefficient vector, \( Q < P \). Substituting (13) into (12), our school-specific model becomes

\[
Y_j = \mathbf{1}_{n_j} \beta_{0j} + X_j z \delta_j + X_j \nu + e_j, \quad (14)
\]

an obvious extension of (11).

Now the coefficient vector \( \delta = \delta_1, \ldots, \delta_J \) has dimension \( JQ \), smaller than \( JP \), but still a large number, with new parameters added for every additional school added to the sample. This problem of “proliferating parameters” will lead to inconsistent estimates. To address this problem, we adopt exchangeability between schools, thereby allowing key parameters to be fixed over schools, while others are exchangeable. We therefore have

\[
\beta_{0j} = W_{0j}^\top \gamma_0 + u_{0j}, \\
\delta_j = W_j \gamma + u_j, \\
\left( u_{0j} \quad u_j \right) \sim N(0, \tau). \quad (15)
\]

Here \( W_{0j} \) is a vector of school characteristics that predict variation in school-specific intercepts \( \beta_{0j} \), and the associated regression coefficients are the elements of the vector \( \gamma_0 \). Similarly, \( W_j \) is a school-specific matrix of explanatory variables that account for between-school heterogeneity in the school-specific regression coefficients, \( \delta_j \), and the associated coefficients are \( \gamma \). Now combining (15) into (14), we have the mixed linear regression model

\[
Y_j = W_{0j}^\top \gamma_0 + X_j z W_j \gamma + \mathbf{1}_{n_j} u_{0j} + X_j z u_j + X_j \nu + e_j \\
= \left( \mathbf{1}_{n_j} W_{0j}^\top \quad X_j z W_j \right) \begin{pmatrix} \gamma_0 \\ \gamma \end{pmatrix} + \left( \mathbf{1}_{n_j} \quad X_j \right) \begin{pmatrix} u_{0j} \\ u_j \end{pmatrix} + X_j \nu + e_j. \quad (16)
\]

We can represent (16) succinctly as a mixed linear model:

\[
Y_j = A_{\beta j} \theta_j + A_{\delta j} \theta_{\delta j} + X_j \nu + e_j, \\
\theta_j \sim N_R(0, \tau), \quad \nu \sim N_p(0, \Delta^2 I_p), \quad e_j \sim N_{n_j}(0, \sigma^2 I_{n_j}) \quad (17)
\]
where $A_{\beta} = \left(1_n, W_{\delta_j}^T, X_j^Tz_jW_j \right)$ is the $n_j$ by $F$ design matrix for the fixed coefficient vector $\theta_{\beta} = (\gamma, \gamma)^T$ and $A_{\theta} = \left(1_n, X_j \right)$ is the $n_j$ by $R$ design matrix for the between-school random effects vector $\theta_{ij}$. 

**B.3. Estimation of Fixed Coefficients**

Given the variance-covariance components, the maximum likelihood estimator of the fixed coefficient vector in (18) is

$$
\hat{\theta}_j = \left( \sum_{j=1}^J A_{\beta}^T H_j A_{\beta} \right)^{-1} \sum_{j=1}^J A_{\beta}^T H_j Y_j,
$$

$H_j = M_j - M_j X_j C_j^{-1} M_j$

$M_j = I_n - A_{\theta} C_j^{-1} A_{\theta}^T$

$C_j = A_{\theta}^T A_{\theta} + \sigma^2 \tau^{-1}$

$C_{ij} = \sum_{j=1}^J X_j^T M_j X_j + \Delta^2 I_p.$

**B.4. Estimation of Variance-Covariance Components**

To estimate the variance-covariance components, we adopt the EM algorithm (Dempster, Laird, & Rubin, 1977). This requires expressions for the conditional distributions of the variance components, given the data and given current estimates of the unknown parameters $\psi = (\theta, \tau, \Delta, \sigma^2)$. We have (i) $v | Y, \psi \sim N(v^*, V_{vv})$, where

$$
v^* = V_{vv} \sum_{j=1}^J X_j^T M_j (Y_j - A_{\theta_j} \hat{\theta}_j)
$$

$$
V_{vv} = \left( \sum_{j=1}^J X_j^T M_j X_j + \sigma^2 \Delta^2 I_p \right)^{-1}
$$

(ii) $u_j | Y, \psi \sim N(u_j^*, V_{uj})$ where

$$
\theta_{ij}^* = C_j^{-1} A_{\theta}^T (Y_j - A_{\theta} \hat{\theta}_j - X_j v^*)
$$

$$
V_{ujj} = C_j^{-1} A_{\theta}^T X_j C_j^{-1} X_j^T C_j^{-1}
$$

(19)
**M-Step**

If the school random effects $\theta_j, j = 2, \ldots, J$, the within-regression random effects $v$, and the fixed coefficients $\theta_j$ were known, the maximum likelihood (subscripted “CD” for “complete-data”) estimators the variance-covariance components could be computed in a single step:

$$
\hat{\tau}_{CD}^2 = J^{-1} \sum_{j=1}^J \theta_j \theta_j^T \quad \hat{\Delta}_{CD} = P^{-1}v^Tv \quad \hat{\sigma}_{CD}^2 = N^{-1} \sum_{j=1}^J e_j^T e_j. \quad (21)
$$

**E-Step**

These “complete-data” MLEs cannot be computed, because the sufficient statistics required for them are unknown. The idea behind EM is to substitute for these sufficient statistics their conditional expectations, given the observed data and current estimates of the parameters:

$$
E \left( \sum_{j=1}^J \theta_j \theta_j^T \mid Y, \psi \right) = \sum_{j=1}^J (\theta_j^T \theta_j + V_{rj})
$$

$$
E(v^Tv \mid Y, \psi) = v^Tv^* + tr(V_{vv})
$$

$$
E \left( \sum_{j=1}^J e_j^T e_j \mid Y, \psi \right) = \sum_{j=1}^J e_j^T e_j^* + J\Delta^2 + \sigma^2 \tau^{-1} tr \left( \sum_{j=1}^J A_j^T A_j \right) + \sigma^2 \Delta^2 tr \left( \sum_{j=1}^J X_j^T X_j \right).
$$

(22)

Each iteration of the algorithm increases the log-likelihood until the achievement of a desired rate of convergence.