Does equating matter in value-added models?

Sedat Şen  
Harran University, Şanlıurfa, Turkey, sedatsen@harran.edu.tr  
_orcid.org/0000-0001-6962-4960

Ragıp Terzi  
Harran University, Şanlıurfa, Turkey, terziragip@harran.edu.tr  
_orcid.org/0000-0003-976-5054

İbrahim Yıldırım  
Harran University, Şanlıurfa, Turkey, iyildirim84@gmail.com  
_orcid.org/0000-0002-4137-2025

Allan S. Cohen  
University of Georgia, Athens, GA, USA, acohen@uga.edu  
_orcid.org/0000-0002-877-9378

ABSTRACT  
The purpose of this study was to examine the effect of equated and non-equated data on value-added assessment analyses. Several models have been proposed in the literature to apply the value-added assessment approach. This study compared two different value-added models: the unadjusted hierarchical linear model and the generalized persistence model. The former model assumes equated tests while the latter one relaxes this assumption. Two different data sets (equated and non-equated) were analyzed with both models. Value-added estimates for both models based on a statewide examination (equated) and a countrywide examination (non-equated) data were generally consistent. School rankings showed differences between the two models. The practical implication of this study is that although there were small differences in school rankings, a model requiring an equating assumption can be applied to a non-equated data set in a case when equating between test forms is not possible.

Keywords  
Value-added assessment, Hierarchical linear model, Generalized persistence model, Test equating.

Katma-değerli değerlendirme modellerinde eşitleme önemli mi?  
ÖZ  

Anahtar Kelimeler  
Katma-değerli değerlendirme, Hiyerarşik doğrusal model, Genelleştirilmiş süreklilik modeli, Test eşitleme.

Cite This  
INTRODUCTION

The effectiveness of a school or a teacher has been debated for decades. How to identify a qualified teacher or an effective school constitutes a major problem in education? The focus of our study which emerged from this problem on value-added analysis (VA) is a recent approach to determining school/teacher effectiveness. In this process, various approaches have been developed for the evaluation of the school effectiveness such as using school’s overall achievement scores and data envelopment analysis (Bessent & Bessent, 1980) that take into account teachers’ proficiency levels in teaching. Since the problem of “effectiveness” contains many variables (Marzano, 2003), only measurable benefits can be expressed in an unbiased manner. In this context, “student achievement” can be considered the most important indicator of school effectiveness (Balcı, 1988).

The assessment of student learning is a major policy issue in the field of education (Ercikan, 2006). As is well known, there are many factors that affect student achievement. Among a number of factors, teacher and school effects in students’ test score gains can be detected by a recent statistical approach to assessment of student learning (Sanders et al., 2002). Hanushek (1972) first used VAA in an accountability system. Sanders and his colleagues implemented VAA in the Tennessee Value Added Accountability System (TVAAS), a statewide testing system (Sanders & Horn, 1994; Sanders & Rivers, 1996; Sanders et al., 1997; Wright, Horn, & Sanders, 1997).

Sanders et al. (1997) defined a teacher value-added score as the differences between the predicted level of achievement and current achievement in a classroom taught by the teacher. In this definition, the magnitude of differences in the predicted and observed test scores is assumed to reflect teacher and school effectiveness. If the measured score is higher than the predicted score, it is interpreted to mean that the teacher and school “add” to student achievement, otherwise, they detract from student achievement.

A number of Value-Added Models (VAMs) have been developed to track individual students’ academic growth over years and in different subjects so that teachers’ contributions to that growth can be estimated (Braun, 2005). In this way, these models are intended to control for student-level socio-demographic variables (e.g., age, gender, and ethnicity) that may have effects on student achievement. The purpose of VAMs is to obtain accurate and reliable comparisons of student achievement across schools regardless of large demographic or ability differences in student populations. Some VAMs rely on fixed school effects while others rely on random school effects. For instance, early VAM applications (e.g., Hanushek, 1972; Murnane, 1975) assume fixed effects models, whereas, more recent applications (e.g., the TVAAS layer model) assume random effects models (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). The former methods are based on regression models (Tekwe et al., 2004), while the latter use more complex statistical models such as mixed models or hierarchical models (Aitkin & Longford, 1986; Raudenbush & Bryk, 1986) to assess school and teacher effects (Şen, Kim, & Cohen, 2017).

Estimating teacher effectiveness based on student achievement using VAMs requires longitudinal data in order to track the impacts of prior educational inputs on future achievement (Mariano et al., 2010). In order for test scores to be used to estimate growth, however, they need to be vertically equated and scaled to a common metric (Ballou et al., 2004; Briggs & Domingue, 2013; Doran & Cohen, 2005). Therefore, one challenging issue in modeling longitudinal data is the need for equating test scores as most standardized tests consist of different items varying in difficulty. This is a crucial point because inferences based on VAMs should be made with respect to the validity standards (Hill, 2009). Concerns have been raised about the limitations of vertical scaling and equating in part because latent constructs are subject to changes at each grade level. Furthermore, equating constructs that shift across grades can result in biased and distorted value-added teacher effectiveness (Braun & Wainer, 2007;
Martineau, 2006). Another issue that may affect the estimates of VAMs is the choice of vertical scaling methodology as this can affect the subsequent results from a given VAM (Briggs & Domingue, 2013; Briggs & Weeks, 2009).

The majority of VAMs require equating between consecutive grades, while some VAMs relax it by simply requiring scores across grades be linearly related (Shaw, 2012). In the absence of vertically scaled data sets, some possible alternative methods have been proposed for use with VAA. Mariano et al. (2010) and McCaffrey et al. (2003; 2004) proposed the variable persistence model for data with non-constant variances and covariances obtained from different developmental scales. It is also the case, however, that inferences about teacher effectiveness may be biased due to measurement error in previous test scores (Shaw, 2012). Reckase (2004) noted that comparing results across years does not provide unbiased estimates if different skills and academic domains are included in a VAA. A number of models have been introduced to deal with aforementioned issues (Broatch & Lohr, 2012; Lockwood et al., 2007; Mariano et al., 2010). Some of these VAMs ignore construct shift entirely and directly carry out analyses with vertically scaled test scores, while some VAMs model construct shift with vertically scaled test scores, and some other VAMs ignore vertically scaled test scores completely. There is no research reported, however, comparing the sensitivity of VAMs based on equated and non-equated data sets.

Two approaches to VAA were investigated in this study, the unadjusted hierarchical linear model (UHLMM; Tekwe, et al., 2004) and the generalized persistence model (GP model; Mariano et al., 2010). Results from these two models were compared to determine whether they were consistent with one another and the two models differed with respect to school rankings?

METHODOLOGY

In this study, two different data sets (equated and non-equated) were analyzed. Equated data for this study were taken from a vertically scaled statewide mathematics test administered to 8th graders from 2002 and 2003 in a large Southeastern state in USA. As these data were vertically scaled, they were consistent with models requiring equating. This test is a part of a criterion-referenced test that aims to assess student achievement in the high-order cognitive skills represented in the state standards in reading, mathematics, writing, and science. Three types of questions – multiple choice items, graded response items, and performance tasks – were used in this test. Non-equated data for this study were taken from the 2015 November and 2016 April administrations of a countrywide exam in Turkey, namely, the Exam for the Transition from Basic Education to Secondary Education (also known as TEOG; Ministry of Education). The exam scores from the Grade 8 mathematics section of this test were used for analyses under a non-equating condition. The exams consist of 20 multiple choice questions and data were collected from schools in a province located in southeastern Turkey. Twenty schools were randomly selected from each of the two data sets. The samples consisted of 9,811 students for the vertically equated statewide examination data set and 941 students for the countrywide examination data set.

VAMs Used in This Study

Unadjusted hierarchical linear model and generalized persistence model were used to examine the effect of vertical equating on value-added estimates. A brief explanation about these models is presented below.
Unadjusted hierarchical linear model

The UHLMM uses unadjusted change scores with a random intercept. This model consists of a two-level HLM described by the following equations;

Student-level model,

\[ d_{ijs} = \beta_{0is} + \varepsilon_{ijs} \]  \hspace{1cm} (1)

where \( d_{ijs} \) is the change score, \( \beta_{0is} \) is a random intercept associated with the school \( i \), and \( \varepsilon_{ijs} \) represents random error.

School-level model,

\[ \beta_{0is} = \gamma_{0s} + \xi_{is} \]  \hspace{1cm} (2)

where \( \gamma_{0s} \) is the mean of the random intercepts, \( \beta_{0is} \) and \( \xi_{is} \) are the random effect and random error for school \( i \) on the random intercept for subject area \( s \), respectively. \( \beta_{0is} \) and \( \xi_{is} \) are assumed to be independent. The single equation form can be written as

\[ d_{ijs} = \beta_{0s} + \xi_{is} + \varepsilon_{ijs}. \]  \hspace{1cm} (3)

Generalized persistence model

The GP model is a general multivariate model for estimating teacher or school effects based on a longitudinal data set, which was developed by Mariano et al. (2010), it was intended to accommodate both school effect decay and scale changes. The GP model estimates are computed from a Bayesian framework for non-equated longitudinal data set. A student’s year \( t \) score depends on an overall year \( t \) mean for all students, plus a cumulative sum of the current year and past year schools’ effects, plus a random residual error term for the student in the current testing year. \( y_{it} \) is the achievement score of student \( i \) in year \( t \) and the GP model for this score is

\[ y_{it} = \mu_t + \left( \sum_{g=1}^{t} \sum_{j=1}^{J_g} \phi_{i,j|g} \theta_{g|j|t} \right) + \varepsilon_{it}, \]  \hspace{1cm} (4)

where \( \mu_t \) is the overall mean for the year, \( \phi_{i,j|g} \) equals 1 if student was taught by school \( j \) in year \( g \), and 0 otherwise. Therefore, the products of \( \phi_{i,j|g} \theta_{g|j|t} \) provide the school effects for the current and prior grades, and \( \varepsilon_{it} \) is the residual error term.

Results of the UHLMM and GP models based on the equated and non-equated data sets were compared in this study. As mentioned, the UHLMM requires equated test scores while the GP model relaxes this assumption. UHLMM analyses were conducted with SAS software using the code provided by Tekwe et al. (2004); GP model estimations were conducted using GPvam R package (Karl et al., 2012).
RESULTS

Value-added estimates from each VAM used in this study are shown in Table 1. The UHLMM provides value-added estimates as best linear unbiased predictors (BLUP), while the GP model provides empirical best linear unbiased predictors (EBLUP). As shown in Table 1, value-added estimates for the statewide test from both models appeared to be consistent in terms of sign for most of the schools except for Schools 3, 4, 6, 9, 13, and 20 (presented in bold). Similarly, value-added estimates for the countrywide test from both models appeared to be consistent in terms of sign for most of the schools except for Schools 8, 13, 15, 17, and 20 (presented in bold). However, the magnitude of the school estimates varied.

Table 1
Estimates of the school effects obtained from two VAMs based on grade 8 statewide and countrywide math test results

<table>
<thead>
<tr>
<th>School ID</th>
<th>UHLMM BLUP</th>
<th>SE</th>
<th>GP EBLUP</th>
<th>SE</th>
<th>School ID</th>
<th>UHLMM BLUP</th>
<th>SE</th>
<th>GP EBLUP</th>
<th>SE</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>45.58</td>
<td>8.44</td>
<td>20.71</td>
<td>8.72</td>
<td>12</td>
<td>2.47</td>
<td>1.79</td>
<td>1.12</td>
<td>1.23</td>
</tr>
<tr>
<td>2</td>
<td>23.65</td>
<td>7.86</td>
<td>17.57</td>
<td>8.47</td>
<td>4</td>
<td>1.79</td>
<td>1.86</td>
<td>0.67</td>
<td>1.26</td>
</tr>
<tr>
<td>10</td>
<td>21.38</td>
<td>9.32</td>
<td>26.53</td>
<td>9.10</td>
<td>3</td>
<td>0.82</td>
<td>1.78</td>
<td>0.59</td>
<td>1.23</td>
</tr>
<tr>
<td>9</td>
<td>19.24</td>
<td>8.85</td>
<td>-2.79</td>
<td>8.89</td>
<td>6</td>
<td>1.39</td>
<td>1.97</td>
<td>0.52</td>
<td>1.29</td>
</tr>
<tr>
<td>3</td>
<td>10.66</td>
<td>8.82</td>
<td>-4.46</td>
<td>8.88</td>
<td>1</td>
<td>0.98</td>
<td>1.81</td>
<td>0.43</td>
<td>1.24</td>
</tr>
<tr>
<td>12</td>
<td>8.84</td>
<td>7.98</td>
<td>9.59</td>
<td>8.52</td>
<td>7</td>
<td>1.07</td>
<td>1.79</td>
<td>0.29</td>
<td>1.23</td>
</tr>
<tr>
<td>8</td>
<td>6.99</td>
<td>7.57</td>
<td>7.96</td>
<td>8.36</td>
<td>17</td>
<td>-0.26</td>
<td>1.81</td>
<td>0.25</td>
<td>1.24</td>
</tr>
<tr>
<td>4</td>
<td>2.11</td>
<td>9.00</td>
<td>-6.89</td>
<td>8.96</td>
<td>15</td>
<td>-0.78</td>
<td>1.19</td>
<td>0.14</td>
<td>1.08</td>
</tr>
<tr>
<td>20</td>
<td>0.46</td>
<td>8.34</td>
<td>-5.50</td>
<td>8.68</td>
<td>20</td>
<td>-0.13</td>
<td>1.83</td>
<td>0.09</td>
<td>1.25</td>
</tr>
<tr>
<td>6</td>
<td>-0.09</td>
<td>8.88</td>
<td>4.34</td>
<td>8.91</td>
<td>5</td>
<td>0.39</td>
<td>1.65</td>
<td>0.08</td>
<td>1.19</td>
</tr>
<tr>
<td>19</td>
<td>-0.85</td>
<td>8.37</td>
<td>-2.55</td>
<td>8.69</td>
<td>2</td>
<td>0.21</td>
<td>1.67</td>
<td>0.00</td>
<td>1.20</td>
</tr>
<tr>
<td>7</td>
<td>-4.83</td>
<td>8.53</td>
<td>-0.73</td>
<td>8.75</td>
<td>8</td>
<td>0.25</td>
<td>1.93</td>
<td>-0.02</td>
<td>1.28</td>
</tr>
<tr>
<td>13</td>
<td>-5.79</td>
<td>7.82</td>
<td>1.65</td>
<td>8.46</td>
<td>13</td>
<td>0.35</td>
<td>1.79</td>
<td>-0.09</td>
<td>1.23</td>
</tr>
<tr>
<td>15</td>
<td>-8.79</td>
<td>9.09</td>
<td>-8.54</td>
<td>9.00</td>
<td>11</td>
<td>-1.43</td>
<td>1.85</td>
<td>-0.33</td>
<td>1.25</td>
</tr>
<tr>
<td>18</td>
<td>-13.06</td>
<td>7.68</td>
<td>-10.34</td>
<td>8.39</td>
<td>18</td>
<td>-0.59</td>
<td>1.87</td>
<td>-0.38</td>
<td>1.26</td>
</tr>
<tr>
<td>11</td>
<td>-14.58</td>
<td>7.62</td>
<td>-2.93</td>
<td>8.37</td>
<td>10</td>
<td>-0.66</td>
<td>1.87</td>
<td>-0.47</td>
<td>1.26</td>
</tr>
<tr>
<td>14</td>
<td>-14.76</td>
<td>7.42</td>
<td>-2.27</td>
<td>8.29</td>
<td>16</td>
<td>-1.12</td>
<td>1.85</td>
<td>-0.58</td>
<td>1.25</td>
</tr>
<tr>
<td>17</td>
<td>-16.93</td>
<td>9.06</td>
<td>-17.09</td>
<td>8.98</td>
<td>19</td>
<td>-1.16</td>
<td>1.64</td>
<td>-0.66</td>
<td>1.19</td>
</tr>
<tr>
<td>16</td>
<td>-19.55</td>
<td>7.48</td>
<td>-6.04</td>
<td>8.32</td>
<td>9</td>
<td>-1.76</td>
<td>1.08</td>
<td>-0.75</td>
<td>1.06</td>
</tr>
<tr>
<td>1</td>
<td>-39.67</td>
<td>7.66</td>
<td>-20.49</td>
<td>8.39</td>
<td>14</td>
<td>-1.83</td>
<td>1.79</td>
<td>-0.93</td>
<td>1.23</td>
</tr>
</tbody>
</table>

School IDs are sorted by UHLMM; inconsistent results (in terms of signs) between two models are presented in bold.

Value-added scores are typically used for school ranking. The schools were ranked based on the value-added scores obtained from two different models. The school ranks based on the UHLMM and GP model are presented in Table 2. As shown in Table 2, school rankings based on statewide test data showed differences between UHLMM and GP models. Only the least successful school (i.e., School 1) was found to be the same in both models. The three most successful schools were the same but they were in different orders for the two models. School rankings for the countrywide test data also showed differences between UHLMM and GP models (see Table 2). As shown in Table 2, twenty-five percent of the schools (Schools 12, 4, 1, 9, and 14) were ranked the same by both models (presented in bold). Although other schools’ ranks did differ, the school rankings appeared to be close to each other from both models. Spearman rank correlations were calculated between the ranks obtained from both models. The correlations were .699 and .878 for the statewide (equated) data and the countrywide (non-equated) data, respectively.
Table 2
*School ranks obtained from the two VAMs based on grade 8 statewide and countrywide math test results*

<table>
<thead>
<tr>
<th>Ranking</th>
<th>Statewide Test School ID</th>
<th>UHLMM</th>
<th>GP</th>
<th>Countrywide Test School ID</th>
<th>UHLMM</th>
<th>GP</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>10</td>
<td></td>
<td>12</td>
<td>12</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>5</td>
<td></td>
<td>4</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>10</td>
<td>2</td>
<td></td>
<td>6</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9</td>
<td>12</td>
<td></td>
<td>7</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>8</td>
<td></td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>6</td>
<td></td>
<td>3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>8</td>
<td>13</td>
<td></td>
<td>5</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>7</td>
<td></td>
<td>13</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>14</td>
<td></td>
<td>8</td>
<td>20</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>6</td>
<td>19</td>
<td></td>
<td>2</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>19</td>
<td>9</td>
<td></td>
<td>20</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>11</td>
<td></td>
<td>17</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>3</td>
<td></td>
<td>18</td>
<td>13</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>15</td>
<td>20</td>
<td></td>
<td>10</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>18</td>
<td>16</td>
<td></td>
<td>15</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>4</td>
<td></td>
<td>16</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>14</td>
<td>15</td>
<td></td>
<td>19</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>17</td>
<td>18</td>
<td></td>
<td>11</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>16</td>
<td>17</td>
<td></td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>1</td>
<td>1</td>
<td></td>
<td>14</td>
<td>14</td>
<td></td>
</tr>
</tbody>
</table>

Consistent rankings between two models are presented in bold.

**DISCUSSION and CONCLUSION**

In this study, equated statewide test data and non-equated countrywide test data were analyzed with both UHLMM and GP models. In general, the estimated effects of most of the schools are compatible for both data sets. On the other hand, school rankings showed differences between the UHLMM and GP model for both data sets. The school rankings based on the two VAMs were closer for the non-equated data set than for the equated data set. The correlation of school effects across models also appeared to be stronger in the non-equated data (i.e., the countrywide test data) than in the non-equated data (i.e., the statewide examination test data). Thus, the differences between results of the UHLMM and GP model appeared to be smaller for the non-equated data set. One possible explanation for the differences between two data sets may be due to the equating effect. As a result, it can be concluded that practitioners should be careful about the model choice. When test scores are equated, then the data should be analyzed with the UHLMM. When data are not equated, then tests can be estimated by either the UHLMM or GP model. As the test score equating is a tedious process and is not always possible in the real testing applications (e.g., the lack of anchor items), practitioners may prefer either the UHLMM or GP model. These results are consistent with Yıldırım and Şen’s (2018) study. Yıldırım and Şen (2018) have compared the GP model to the UHLMM under non-equated data set and they found that tests can be estimated by either the UHLMM or GP model for non-equated data set.

Test equating is an important process if one wants to compare results from different forms of the same test. This is likewise important when the test scores from multiple years are to be compared. However, this issue has not been studied sufficiently for comparison of value-added assessments. A relatively small number of studies have been reported examining scaling effects on value-added estimates (e.g., Briggs & Domingue, 2013; Briggs & Weeks, 2009; Briggs, Weeks, & Wiley, 2008). Briggs and Weeks (2009), for example, examined the effect of different scaling methods on school level
estimates, and they found that scaling did have an effect on the estimates. Similarly, in the present study showed, there were differences observed in terms of school-level value-added estimates and in school rankings between equated and non-equated data sets. Briggs and Domingue (2013) note that choices in vertical scaling may also have an effect on teacher and school effects. Although only one data set with vertical scaling was examined in this study, results provide evidence that test equating may have an effect on model selection and school estimates. Although vertical scaling appears to be important for growth models, vertical equating using IRT does not guarantee an equal interval scale in value-added assessment applications (Ballou, 2009).

Several VAMs have been developed for determining teacher and school effectiveness. Each model has some strengths and weaknesses. Persistence models are different from gain score models in that they incorporate persistence of school effects. Another possible explanation for the difference between the two data sets in this study may be due to school effect estimates that are sensitive to different modeling specifications, such as the persistence of school effects. Although VAMs appear to provide an objective tool for use in educational accountability systems, these models should be used cautiously along with other tools to determine effective and ineffective schools (Beardsley, 2008).

Thinking of a different perspective, some countries such as Turkey and the US give a key role to private schools and tutoring centers for high-stakes tests. It could be particularly helpful for parents to decide which school or tutoring center they should send their children. In this regard, ranking of these schools and centers based on its effectiveness could be considered as an alternative procedure because tracking students’ academic growth can be explained by teachers’ contributions using VAMs. Furthermore, considering the performance salaries of teachers in these private institutions (Boran, Atalmis, & Sagir, 2015), the importance of using VAMs cannot be ignored.

In this study, models were compared using empirical data sets for two consecutive years for a single subject (i.e., mathematics). VAM applications are also potentially biased if school- and student-related covariates are excluded, although some VAMs can statistically control school- and student-related variables. Research on the effects of equating on value-added scores might benefit by inclusion of covariates for school- and student-related variables.

REFERENCES


Katma-değerli değerlendirme modellerinde eşitleme önemli mi?
