Observed Teacher Profiles in the Measures of Effective Teaching Study

Daniel L. Murphy mailto:dan.murphy@pearson.com
Katherine McKnight mailto:Kathy.mcknight@pearson.com

International Association for Educational Assessment Conference

Singapore, May, 2014

Abstract

In the past decade in the United States, teacher evaluation metrics have expanded rapidly as a means of promoting student learning and spurring improvement in the teaching profession. The use of multiple measures to measure teacher effectiveness is now commonplace. This paper describes in detail multi-dimensional profiles of teacher effectiveness using data from the Measures of Effective Teaching (MET) study and examines teacher and student characteristics associated with these profiles. Principal component and cluster analysis enabled the detection of patterns that fit profiles of emotional, instructional, and achievement support in which teachers of most skill levels exhibit relative strengths and weaknesses. The bottom 5% of teachers in terms of quality (i.e., *Below Standard*) exhibited low scores across all three dimensions. Teachers in the 6th-50th percentiles (i.e., *Developing*) exhibited low to below average scores on two dimensions with a third relatively stronger dimension in which scores were slightly above average. Teachers in the 51st-95th percentiles (i.e., *Proficient*) teachers exhibited above average to high scores across two of the dimensions with a third relatively weaker dimension in which scores were slightly below average. The top 5% of teachers (i.e., *Exemplary*) exhibited high scores on all three of the dimensions. Below Standard and Developing teachers tended to teach in middle schools with disproportionate representations of economically disadvantaged ethnic minority students, suggesting that the children who need the highest quality educational experiences have teachers who are struggling the most to provide it.

Keywords: teacher effectiveness, composite measures, cluster analysis

Observed Teacher Profiles in the Measures of Effective Teaching Study

Recent research indicates that teachers have a considerable effect on student learning ouctomes (e.g., Branch, Hanushek, & Rivkin, 2009; Chetty, Friedman, & Rockoff, 2011; Hanushek & Rivkin, 2010; Staiger & Rockoff, 2010). It follows that school systems which can reliably identify high- and low-performing teachers can use this information to inform personnel decisions, which in turn should improve student learning. To this end, many U.S. school systems are reconstructing their teacher evaluation systems to better differentiate high- from low-performing teachers.

A basic framework representative of evaluation systems currently implemented in U.S. schools (e.g., Denver, Hillsborough County, Florida, Washington, D.C.) uses four categories to discriminate high- from low-performing teachers based on their position in an overall distribution of teacher effectiveness scores (Hansen, Lemke, & Sorensen, 2013). The teacher effectiveness scores themselves are often composed of multiple measures; typically, one of the measures is based on teacher observation, one is based on student survey, and one is based on student achievement progress. Teachers are commonly categorized into performance levels based on the ranking of their composite score in the overall distribution of scores (Hansen, et al.).

A simple teacher effectiveness accountability system might, for example, categorize teachers into four performance levels, where ineffective teachers comprise the bottom 5 percent of composite scores, marginally effective teachers comprise the 6th through 50th percentiles, effective teachers comprise the 51st through 95th percentiles, and highly effective teachers comprise the top 5 percent of composite scores. The composite scores are supposed to provide teachers with meaningful data across several dimensions to inform instructional practices and improve student learning outcomes (Kane & Staiger, 2012). Furthermore, using evaluation results to inform professional development may empower teachers to self-direct their growth based on feedback from their evaluation (Nolan & Hoover, 2005). On the other hand, composite scores may capture several correlated dimensions in a single number, making it difficult to diagnose areas in need of improvement and inform professional development. Composites may also encourage simplistic or misguided policy conclusions if they are poorly constructed (Saisana & Tarantola, 2002).

The composite score approach can be useful for accountability and decision-making purposes in which a single summary statistic simplifies the evaluation of outcomes; however, composite scores also may mask useful information from higher order interactions among the individual measures. Principal component analysis and cluster analysis can capture and simplify such higher order interactions, revealing underlying structural patterns that may be obscured by the composite score approach (LoCasale-Crouch et al., 2007). In contrast with procedures that isolate variation one dimension at a time, cluster analysis displays multiple dimensions to empirically uncover core integrated profiles (Cronbach & Gleser, 1953). Examination of those integrated profiles promotes greater insight into the nature and complexity of the constructs being evaluated (Glutting, McDermott, & Konold, 1997).

In the present study, principal component and cluster analytic approaches were used to identify subgroups that display similar patterns of strengths and weaknesses across dimensions of instructional, emotional, and student achievement support. Although there is variability among classrooms within a profile, the expectation is that classrooms are more similar across dimensions within the same profile than they are across dimensions of the other profiles. Based on previous work examining teacher effectiveness and classroom performance, we expected to find clusters of teachers demonstrating differential patterns of performance (Stuhlman & Pianta, 2009; Kane & Staiger, 2012).

A secondary question addressed by this study is whether all students have the same probability of encountering effective teachers. Current United States educational policy explicitly states that all children should have equal access to high-quality schooling (NCLB, 2001). Previous research suggests, however, that children with fewer economic resources and members of racial minorities are less likely attend high-quality schools and encounter high-quality teaching (Lee & Burkham, 2002; Stuhlman & Pianta, 2009). Adding to this body of research, this study aims to examine similar quality and access issues with students in grades 4-8 who were randomly assigned to classrooms as part of the MET project. Our use of empirically based clustering procedures with data from the MET project enables the identification of patterns and correlates that may advance policy discussions and enhance teacher development and quality.

Using principal component and cluster analysis techniques to examine dimensions of teacher effectiveness in classrooms from 6 U.S. states, this study addresses the following questions: (1) What profiles describe variations in teacher effectiveness in a large sample of classrooms in grades 4-8 from large urban areas in the United States? (2) Do teacher profiles differ with regard to teacher characteristics (gender, education/certification, experience, school level) and specific classroom student compositions (poverty, ethnicity, and English learner status)?

Method

Data

The analysis in this report is based on data collected during the Bill & Melinda Gates Foundation's MET project. Specifically, the data include the teacher observation scores of 1,002 teachers from the following districts: Charlotte-Mecklenburg, N.C.; Dallas; Denver; Hillsborough Co., Fla.; New York City; and Memphis. This was the subset of MET project volunteers who taught math or English language arts in grades 4 through 8 and who agreed to participate in random assignment during the second year of the project. The sample was 84% female and 58% Caucasian. In addition, teachers in this sample had on average approximately 8 years of experience teaching in their current district. In terms of education and credentials, 35% held both a Bachelor's degree and at least a Master's degree. The average classroom was composed of 68% students who were racial or ethnic minorities, 57% who were economically disadvantaged, and 14% who were English learners.

Measures

Framework for Teaching

The Framework for Teaching is an instruction assessment tool developed by Charlotte Danielson as an extension of her work on Educational Testing Service's PRAXIS III framework for assessing new teachers (Danielson, 2007). The FFT consists of eight components that are scored on a four-point rubric (i.e., *unsatisfactory, basic, proficient, distinguished*). Videotaped lessons taught by MET project teachers were scored by 148 trained raters. The final FFT component scores were calculated by taking the average of multiple lessons rated by multiple raters. Those final component scores were then averaged and standardized to create one overall final FFT score.

Tripod Survey

The Tripod survey instrument, developed by Harvard researcher Ron Ferguson and the Tripod Project for School Improvement, measures the extent to which students experience the classroom environment as engaging, demanding, and supportive of their intellectual growth (Kane & Cantrell, 2010). Students report the degree to which they agree with statements on a five-point scale across seven constructs: *Care, Control, Clarify, Challenge, Captivate, Confer* and *Consolidate*. Each teacher’s final Tripod survey score was created by averaging scores across classrooms and scales and then standardizing the final average scores.

Value Added Scores

Student achievement was measured in two ways, with existing statewide standardized assessments and with two supplemental assessments, the Stanford 9 Open-Ended Reading assessment and the Balanced Assessment in Mathematics (Mihaly, McCaffrey, Staiger, & Lockwood, 2013). Teacher value added was estimated using a two-step procedure, which is described in detail in *Learning about Teaching: Initial Findings from the Measures of Effective Teaching Project* (Kane & Cantrell, 2010).

Composite Score

The standardized FFT, Tripod survey, and value-added scores were equally weighted and combined into a final composite variable score. To mimic teacher effectiveness ratings commonly seen in practice, teachers were then assigned to one of four teacher effectiveness categories based on the position of their composite score in the final distribution of scores. Teachers in the bottom 5% of the distribution were categorized as *Below Standard*; teachers in the 6th-50th percentiles were categorized as *Developing*; teachers in the 51st-95th percentiles were classified as *Proficient*; and the top 5% of teachers were classified as *Exemplary*.

Data Analysis

In the first step, the Princomp procedure in SAS version 9.2 (SAS Institute Inc, Cary, North Carolina) was used to apply principal component analysis in order to examine the item-level relationships among the different measures. As input, we used the eight items from the FFT, the seven Tripod scale scores, and the two value added scores from the statewide and alternative assessments. The goal of this procedure was to reveal the internal structure of the composite measure in a way that best explained the variance in the scores. We plotted the eigenvalues associated with the principal components according to their size using a scree plot (Cattell, 1966), looking for a point in the plot where the slope flattened dramatically (i.e., the *elbow*), keeping only the components before the elbow. We also checked that the components before the elbow had eigenvalues larger than 1 (see, *e.g.,* Kaiser, 1961). After deciding on the number of principal components that best explained the variance in the scores, we created output data sets containing standardized principal component scores (i.e., *factor scores*).

The second step used the SAS Fastclus procedure (SAS Institute Inc, Cary, North Carolina) to perform a cluster analysis on the basis of distances computed from the factor scores outputted during the first step. The observations were divided into clusters such that every observation belonged to one and only one cluster. In addition, we specified a BY statement to obtain separate analysis on observations in groups defined by the final teacher effectiveness categories. Decisions about the number of clusters that best described the core teacher profiles were based on inspecting the cubic clustering criterion (CCC) statistic and *R square,* the proportion of variance accounted for by the clusters. CCC values greater than 2 or 3 indicate good clusters. Values between 0 and 2 indicate possible clusters; large negative values of the CCC can indicate outliers. Decisions about the final cluster solution were based on the fewest number of clusters that both produced acceptable CCC values and accounted for a substantial proportion of variance.

Results

Results from the PCA conducted in the first step of the data analysis indicated that three factors best represented the variance in the composite measure scores. As depicted in Figure 1, the elbow in the scree plot begins after the third component. The three components above the elbow account for 69% of the variance. In addition, the third component has an eigenvalue of 1.32, and the fourth component has an eigenvalue of 0.88, which further supports the three factor solution. The factor loadings for each component aligned with the individual measures, suggesting that each individual measure was capturing a unique construct.

We did not find satisfactory clustering solutions for the *Below Standard* and *Exemplary* teachers; therefore, we did not separate those two performance levels into clusters; instead, we considered each category to be its own cluster. The optimal solution for the *Developing* and *Proficient* teachers was three clusters. The two cluster and four cluster solutions both produced negative CCC values, suggesting that the three cluster solution was optimal. The CCC for the *Developing* teachers was 1.99 and the *R square* value was .46; for *Proficient* teachers the CCC was 1.24 and the *R square* value was .49.

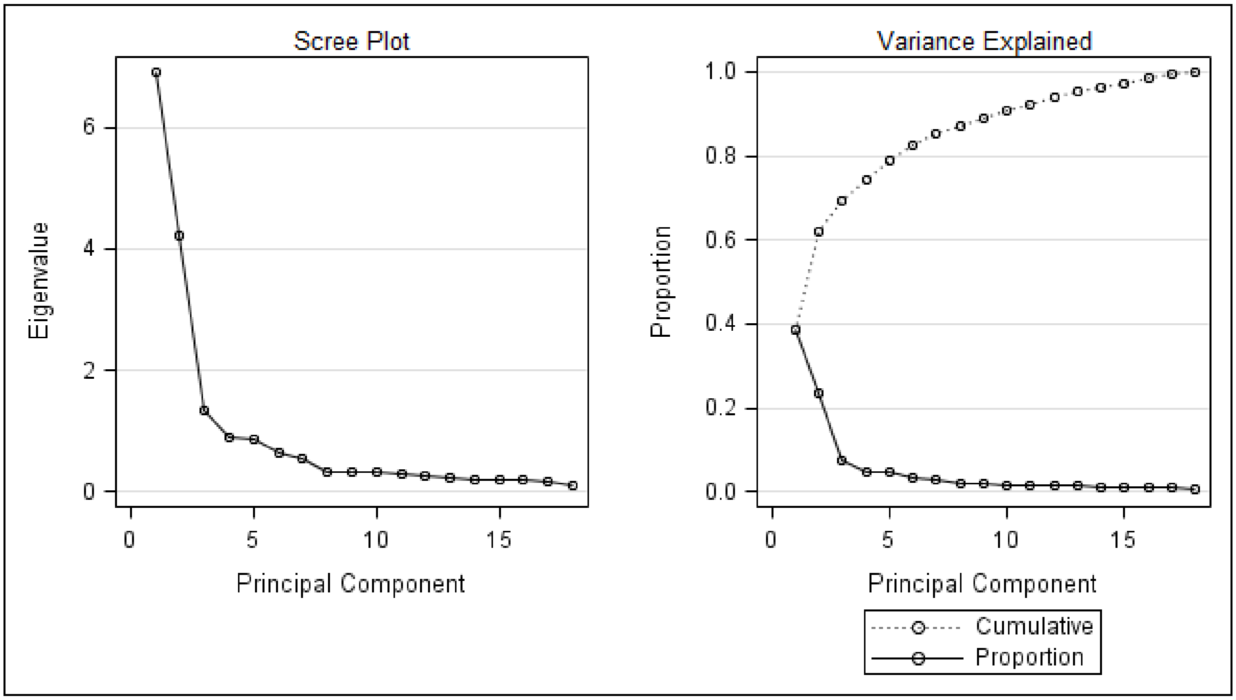

Figure 1. Scree plot produced from the principal component analysis of the teacher effectiveness composite scores.

Teacher Profiles

Table 1 displays descriptive statistics for each teacher profile cluster, as well as pairwise, statistically significant differences across profiles for each variable of interest. Differences in teacher effectiveness measures across profiles were analyzed with a one-way ANOVA ($p < .05$). To evaluate differences among clusters as a set when statistical significance was indicated, follow-up pairwise comparisons were made using Tukey's HSD test. Examination of Table 1 reveals statistically significant differences in individual measure scores across most teacher cluster profiles, which provides evidence that the teacher profiles are significantly different from one another.

Further, one teacher demographic variable was measured on a continuous scale (i.e., years experience within the current district) and was analyzed with a one-way ANOVA ($p <$ .05). The number of years experience a teacher had within their current district did not differ across profiles.

In addition, five categorical teacher demographic variables of gender, teacher ethnicity, location in elementary or middle school, and whether the teacher possessed an advanced degree were examined. Observed proportions within each cluster were compared to what would be expected if these characteristics were proportionately distributed across the eight profiles according to their representation in the total sample. Two-tailed standard errors of proportional difference tests were made for all possible pairwise comparisons within each profile. Significant proportional differences were found for location in elementary or middle school [Elementary, $\chi^2(7, N = 1002) = 60.82$, $p < .001$].

Finally, three demographic variables describing classroom composition of student ethnicity, proportion of English learners, and proportion of economically disadvantaged students were examined. Significant proportional differences were found for student ethnicity and proportion of economically disadvantaged students [Ethnic Minority, $\chi^2(7, N = 1002) = 60.82$, $p = .002$; Economic Disadvantage, $\chi^2(7, N = 1002) = 17.55$, $p = .014$].

In general, teachers in the lowest performance categories were more likely to teach in middle schools overrepresented by ethnic minority and economically disadvantaged students. By contrast teachers in higher categories were more likely to teach Caucasian students in elementary schools. Each teacher profile is examined in further detail below.

Exemplary Teachers

Table 2. Profile 1 Composite Measure Scores

FFT	Tripod	VAM
1.20	1.43	1.57

As shown in Table 2, teachers in this profile typically provide effective instruction, a positive classroom environment, and promote high student achievement growth. Students and adult classroom observers rated these teachers especially highly on their ability to communicate academic content to students, which suggests that these teachers may be effective mentor teachers or professional development leaders.

The teachers in this profile were not significantly different from their peers in terms of their demographic characteristics, their training, or the characteristics of the students that they taught, with the exception that they were comparatively more likely to be found in elementary schools. This suggests that all students in the sample had approximately the same likelihood of encountering an exemplary teacher regardless of the ethnic composition or socioeconomic status of their school, but elementary school students were more likely than middle school students to encounter them.

Table 1

	Profile 1	Profile 2	Profile 3	Profile 4	Profile 5	Profile 6	Profile 7	Profile 8	
	(N=50)	(N=120)	(N=158)	(N=173)	(N=140)	(N=163)	(N=148)	(N=50)	Total
Measures									
Composite	1.40^{a}	0.49^{b}	0.45^{b}	0.42^{b}	-0.37^{c}	-0.45^{cd}	-0.48^{d}	-1.47^{e}	0.00
	(0.23)	(0.31)	(0.30)	(0.30)	(0.31)	(0.31)	(0.31)	(0.27)	(0.69)
FFT (SD)	1.20^{a}	0.67^{b}	-0.20^{c}	1.02^{a}	0.22^{d}	-0.93^{e}	-0.59^{f}	-1.38^{a}	0.01
	(1.03)	(0.63)	(0.55)	(0.67)	(0.58)	(0.55)	(0.59)	(0.68)	(0.99)
Tripod (SD)	1.43^{a}	-0.16^{b}	0.86^{c}	0.55^{d}	-0.51^{e}	0.22^{f}	-1.12^{g}	-1.56^{h}	0.00
	(0.85)	(0.52)	(0.51)	(0.58)	(0.60)	(0.55)	(0.60)	(0.94)	(1.00)
VAM (SD)	1.57^{a}	0.95^{b}	0.70^{c}	-0.30^{d}	-0.81^{e}	-0.64^{e}	0.27^{f}	-1.45^{g}	0.00
	(0.84)	(0.60)	(0.62)	(0.52)	(0.66)	(0.66)	(0.58)	(0.85)	(1.00)
Teacher characteristics									
Female (%)	89	86	83	89	87	80	79	76	84
African American (%)	31	30	38	31	32	36	41	35	34
Caucasian (%)	60	63	54	64	61	52	51	64	58
Hispanic (%)	04	04	07	06	05	10	06	00	06
Elementary School (%)	60^{a}	68^{a}	44^{b}	64^{a}	74^{a}	40^{b}	50^{b}	44^{b}	56
Training									
Master’s degree plus (%)	38	37	38	29	34	34	39	38	35
Years district experience	7.6	7.5	7.5	7.2	7.8	6.8	9.4	7.7	08
Classroom characteristics									
Ethnic minority (%)	66	60^{a}	72	60^{a}	61^{a}	77^{b}	70	82^{b}	68
English learners (%)	08	12	15	13	14	17	13	14	14
Economic disadvantage (%)	50	51	63	51	46^{a}	56	61^{b}	64^{b}	56

N*ote*: ANOVA and chi-square pairwise significant differences are denoted by superscript letters.

Proficient Teachers

Table 3. Profile 2 Composite Measure Scores

FFT	Tripod	VAM
0.67	-0.16	0.95

Within profile 2, teachers provide effective instruction and a slightly negative classroom environment, yet they still promote high student achievement growth (see Table 3). Both students and adults rate these teachers as highly effective in managing classroom procedures and student behavior. Although adults perceive these teachers as establishing a positive culture for learning, students report that the teachers do not make school work interesting. The measures of student progress support the adults' perceptions.

As with profile 1 teachers, the teachers in this profile were not significantly different from their peers in terms of their demographic characteristics or their training, with the exception that they were comparatively more likely to be found in elementary schools. On the other hand, they were more likely to teach Caucasian students than teachers in profiles 6 and 8 (see Table 1).

Table 4. Profile 3 Composite Measure Scores

FFT	Tripod	VAM
-0.20	0.86	0.70

Teachers in profile 3 are below average effectiveness of the instructional support they provide, but they offer a positive classroom environment and promote high student achievement growth. Adult observers rate these teachers low on the quality of their questioning techniques and their ability to engage students in genuine discussion; however, students rate these teachers highly on their ability to clarify difficult concepts and captivate student attention. Student progress measures support the students' perceptions.

The teachers in this profile were not significantly different from their peers in terms of their demographic characteristics, their training, or the characteristics of the students that they taught, with the exception that they were comparatively more likely to be found in middle schools.

Table 5. Profile 4 Composite Measure Scores

FFT	Tripod	VAM
1.02	0.55	-0.30

Teachers in profile 4 provide effective instruction and a somewhat positive classroom environment, but their students grow slightly less than average students do. Adults rate these teachers highest on their use of questioning and discussion techniques, and students report that these teachers care a lot about them. It is curious that despite the agreement among students and adult observers that these are good teachers, measures of student achievement growth are below average.

Teachers in this profile were more likely to be found in elementary schools. In addition, they taught higher percentages of Caucasian students than teachers in the lower performing profiles of 6 and 8. This is the most frequent profile in the sample population.

Developing Teachers

Table 6. Profile 5 Composite Measure Scores

FFT	Tripod	VAM
0.22	-0.51	-0.81

Teachers in profile 5 provide instructional support that is more effective than average, but their classroom environment is somewhat negative, and their students lag well behind their peers an academic achievement growth. Although adult observers perceive these teachers as creating an environment of respect and rapport, students report that these same teachers do not challenge them or clarify concepts very well. The student progress measures support the students' perceptions.

The teachers were comparatively more likely to be elementary school teachers. They were also more likely than teachers in profiles 6 and 8 to teach in schools that are disproportionately composed of Caucasian students, and less likely to teach economically disadvantaged students than teachers in profiles 7 and 8.

Table 7. Profile 6 Composite Measure Scores

FFT	Tripod	VAM
-0.93	0.22	-0.64

Teachers in profile 6 provide ineffective instructional support, but the classroom environment is somewhat positive. The students, however, lag behind their peers in achievement growth. Both students and adults rate these teachers low on measures of classroom behavior management. Students rate the teachers as being above average in their ability to captivate attention and make the class interesting. Adults rate these teachers low on establishing a culture of learning. Student progress measures support the adults' perceptions.

The teachers in this profile were comparatively more likely to be found in middle schools. They were also more likely than teachers in profiles 2, 4, and 5 to teach in schools with disproportionately large proportions of ethnic minority students.

Table 8. Profile 7 Composite Measure Scores

FFT	Tripod	VAM
-0.59	-1.12	0.27

Teachers in profile 8 are below average in the instructional support they provide, and their classroom environment is negative. Somewhat surprisingly, students show academic progress in spite of the apparently limited emotional and instructional support. Both students and adults report that teachers in this profile struggle to captivate student attention and engage students in learning, yet students manage to grow more than the cohort average.

The teachers in this profile also were comparatively more likely to be found in middle-schools. Their classrooms were comparatively more likely to be composed of economically disadvantaged students than teachers in profile 5.

` **Below Standard Teachers**

Table 9. Profile 8 Composite Measure Scores

FFT	Tripod	VAM
-1.36	-1.56	-1.45

Teachers in profile 8 typically provide ineffective instructional support, a negative classroom environment, and their students lag behind their peers in academic achievement growth. Students and adult classroom observers score these teachers especially low on items measuring classroom behavior management. Put simply, the teachers in this profile are struggling and in need of intensive professional development.

The teachers in this profile were comparatively more likely to be found in middle schools. They were more likely than teachers in profile 5 to teach economically disadvantaged students; they were also more likely than teachers in profiles 2, 4, and 5 to teach in schools with disproportionately large proportions of ethnic minority students.

Discussion

In this paper, eight profiles were observed empirically through multi-stage clustering procedures on three dimensions of teacher effectiveness. The internal and external validity of these profiles was acceptable, suggesting that teachers within each profile were similar to each other and different from those in the other profiles. In short, the groups of teachers profiled in this paper are distinct in terms of their teaching effectiveness, which ultimately affects the students who encounter them.

Overall in this study, approximately 5% of the teachers were rated as showing high levels of instructional and emotional support of children while also promoting high student achievement growth. At the other extreme, nearly 5% of this sample was rated, on average, as the least effective teachers on all three dimensions (poorest quality profile). The majority of teachers collectively fit six types characterized by ratings falling between the most and least effective profiles: Developing teachers exhibited two relatively weak dimensions and one relatively strong dimension; whereas proficient teachers exhibited two relatively strong dimensions and one relatively weak dimension.

Although distinct profiles of teacher effectiveness emerged, there seemed to be little relationship between teacher effectiveness and features of teacher characteristics. For example, the most effective teachers (Profile 1) and least effective teachers (Profile 8) did not differ from one another on teacher characteristics like teacher training, experience, ethnicity, or gender.

By contrast, perhaps of most concern are the findings that the profile characterized by the least effective teachers (Profile 8), about 5% of the sample, are more likely to be found in classrooms with higher proportions of ethnic minority students and students in poverty, which are considered to be risk factors for school difficulties. In other words, it appears that the least effective teachers were also those with the heaviest concentration of the most disadvantaged students. This finding is consistent with previous work concerned about the inequality in the U.S. education system (Kozol, 1991; Lee & Burkham, 2002), and that children who need the highest quality educational experience to be successful appear to be getting the poorest (Pianta et al., 2005).

It is important to note, however, that the association is only correlational. We do not know if risk concentration drives teacher effectiveness or the other way around. Given the challenges faced by these high-risk students and teachers, perhaps the best utilization of resources involves concentrated professional development support that directly focuses on raising the quality of socio-emotional and instructional interactions.

There are several limitations in this study that need to be considered. Although cluster analysis provides a way to reduce variability within a composite measure of teacher effectiveness to more manageable and interpretable group comparisons, characteristics of individual teachers can get lost in the analysis. Put simply, there may be individual teachers within any particular teacher profile group for whom the overarching profile is a poor fit. In addition, effect sizes are small for several of the pairwise comparisons, and therefore should be interpreted with caution.

References

Branch, G., Hanushek, E., & Rivkin, S. (2009). Estimating principal effectiveness (CALDER Working Paper 32). Washington, DC: National Center for Analysis of Longitudinal Data in Educational Research. Retrieved from http://www.urban.org/uploadedpdf/1001439-Estimating-Principal-Effectiveness.pdf

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate behavioral research*, *1*(2), 245-276.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood*. NBER Working Papers, December 2011. Retrieved from http://www.nber.org/papers/w17699

Cronbach, L., & Gleser, G. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50*, 456–473. Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching*. ASCD.

Glutting, J. J., McDermott, P. A., & Konold, T. R. (1997). Ontology, structure and diagnostic benefits of a normative subtest taxonomy from the WISC-III standardization sample. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 349–372). New York: Guilford.

Hansen, M., Lemke, M., & Sorensen, N. (2013). *Combining multiple performance measures: Do common approaches undermine districts' personnel evaluation systems?* Washington, DC: American Institutes for Research. Retrieved from http://www.air.org/files/Combining_Multiple_Performance_Measures.pdf

Hanushek, E. A., & Rivkin (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review*, 100(2), 267-271.

Kaiser, H. F. (1961). A NOTE ON GUTTMAN'S LOWER BOUND FOR THE NUMBER OF COMMON FACTORS1. *British Journal of Statistical Psychology*,*14*(1), 1-2.

Kane, T., & Cantrell, S. (2010). Learning about teaching: Initial findings from the measures of effective teaching project. *MET Project Research Paper, Bill & Melinda Gates Foundation*, 9.

Kane, T. J., & Staiger, D. O. (2012). Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains. Policy and Practice Brief. MET Project. *Bill & Melinda Gates Foundation*. Retreived from http://files.eric.ed.gov/fulltext/ED540962.pdf

Kozol, J. (1991). *Savage inequalities: Children in America's schools*. New York: Crown.

Lee, V. E., & Burkham, D. T. (2002). *Inequality at the starting gate: Social background differences in achievement as children begin school*. Washington, DC: Economic Policy Institute.

LoCasale-Crouch, J., Konold, T., Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D. & Barbarin, O. (2007). Observed classroom quality profiles in state-funded pre-kindergarten programs and associations with teacher, program, and classroom characteristics. *Early Childhood Research Quarterly*, *22*(1), 3-17.

Mihaly, K., McCaffrey, D. F., Staiger, D. O., & Lockwood, J. R. (2013). A composite estimator of effective teaching. *Seattle, WA: Bill & Melinda Gates Foundation*.

Morris, R., Blashfield, R., & Satz, P. (1981). Neuropsychology and cluster analysis: Potentials and problems. *Journal of Clinical Neuropsychology, 3*, 79–99.

No Child Left Behind Act of 2001 [NCLB] (2001). Retrieved May 29, 2002. Available at: www.ed.gov/legislation/ESEA02/.

Nolan, J., Jr., & Hoover, L. A. (2005). Teacher supervision and evaluation: Theory into practice. New York: Wiley.

Pianta, R., Howes, C., Burchinal, M., Bryant, D., Clifford, R., Early, D., & Barbarin, O. (2005). Features of pre-kindergarten programs, classrooms, and teachers: Do they predict observed classroom quality and child-teacher interactions?. *Applied developmental science*, *9*(3), 144-159.

Saisana, M., & Tarantola, S. (2002). State-of-the-art Report on Current Methodologies and Practices for Composite Indicator Development. *EUR 20408 EN Report*.

Staiger, D. O., & Rockoff, J. E. (2010). Searching for effective teachers with imperfect information. *The Journal of Economic Perspectives*, 24(3), 97-117.

Stuhlman, M. W., & Pianta, R. C. (2009). Profiles of educational quality in first grade. *The Elementary School Journal*, *109*(4), 323-342.

Version, S. A. S. 9.2 SAS Institute Inc. *SAS Campus Drives*, *Cary*, *NC*, *27513*.