Comparing Teachers across Subject Types and School Levels: Evaluating Teacher Effectiveness with Teacher Effectiveness Student Survey (TESS)

Xintong Li

University of Missouri

Abstract

Many K-12 school districts are required to have rigorous teacher evaluation systems that use multiple sources of data. One source of data, student surveys, are increasingly recognized as a cost-efficient and valid measure of teacher effectiveness. However, research has not addressed whether student surveys should be used to compare teachers in academic and non-academic subjects (e.g., math vs. choir). Similarly, research has not addressed whether student surveys can be used to compare teachers across school levels (e.g., elementary vs. high school). The current study addresses these questions using secondary data from 88 school districts in the Network for Educator Effectiveness (NEE) evaluation system during AY2017-2018. Four teaching practices (17 items) from the Teacher Effectiveness Student Survey (TESS) are analyzed using mean aggregation at the teacher level. Both Exploratory Factor Analysis and Confirmatory Factor Analysis support a two-dimensional structure (i.e., *cognitive press* and *social/emotional support*). Measurement Invariance was confirmed at scalar level across subject types and school levels before a two-way ANOVA was conducted. Results show that, in general, students rated academic teachers higher than non-academic teachers on both factors. Students gave similar ratings to academic teachers on both factors across school levels, but there was an increasing trend across school level for non-academic teachers. Results suggest that districts not compare student ratings of teacher effectiveness for academic and nonacademic teachers, nor across school levels.

Keywords: Teacher Evaluation, Student Survey, Subject, School levels

Introduction

Many K-12 school districts are now required to have rigorous teacher evaluation systems that utilize multiple sources of data to measure teacher effectiveness and make sure that ineffective teaching practices can be addressed. The Network for Educator Effectiveness (NEE) is one of such systems, which provides teacher evaluation resources to over 272 school districts throughout Missouri. NEE incorporates various data sources to measure teacher effectiveness, including classroom observations, unit of instruction (UOI), professional development plan (PDP), and teacher effectiveness student survey (TESS). Compared to other sources of data, student surveys are increasingly used by districts across the United States. Student survey is shown to be a cost-efficient form of measure that is reliable (e.g., Fauth, Decristan, Rieser, Klieme, & Büttner, 2014; Wagner, Göllner, Helmke, Trautwein, & Lüdtke, 2013) and predictive of student achievement (e.g., Downer, Stuhlman, Schweig, Martínez, & Ruzek, 2014; Fraser & McRobbie, 1995; Kane & Staiger, 2012; Roeser, Eccles, & Sameroff, 2000). To utilize the benefits of SET and better incorporate it into the current NEE evaluation system, TESS was created by the researchers and survey designers at the University of Missouri in partnership with school districts. Items are aligned with national InTASC (Interstate Teacher Assessment and Support Consortium) standards (Council of Chief State School Officers, 2011). Though other student surveys are also available across the country, e.g., Tripod Project Survey (MET Project, 2012), My Student Survey (Balch, 2012), and Panorama Student Survey (Panorama P. Education, 2015), none of them considered the diversified needs of districts nor provides formative feedback for teacher development. More importantly, none of the existing surveys were used to examine whether districts can compare teachers

in academic subjects, such as English Language Arts (ELA), Mathematics, Science and Social Studies, with teachers in non-academic subjects, such as Physical and Health Education (PHE), Arts, or other district specialized subjects. Therefore, it is unknown whether the existing evaluation system was built on a fair basis, especially when schools and districts use the assessment results for high-stakes decision makings. The same problem arises when school districts compare teachers across different school levels. The current study examines the factor structure of the most popular TESS components employed by the districts in the network, which indicate four different teaching practices. Further, the measurement invariant was examined across subject types (academic vs. non-academic subjects) and school levels. Finally, the effectiveness ratings based on student survey results were probed to see whether teachers can be compared on a fair basis across subjects and school levels.

Student Evaluation of Teacher Effectiveness

Student Evaluation of Teaching (SET) has been widely used for a long time at the post-secondary level and has gradually been introduced into K-12 schools as an accumulation of evidence shows that ratings from secondary and older primary students are as equally reliable and valid as adults to evaluate teacher effectiveness when using well-constructed and appropriately administered instruments (den Brok, Brekelmans, & Wubbels, 2004; Follman, 1992, 1995).

More recently, Wagner et al. (2016) investigated ratings of instructional quality from 5th-grade students at three time points and concluded that student ratings have moderate-to-high time consistency, which is similar with the time consistency of teacher ratings of instructional quality. Studies on many well-structured or widely used student

survey measures agree that student ratings of classroom climate and/or teaching quality are generally reliable, such as those on Tripod Project Survey (Kane & Staiger, 2012; Polikoff, 2015), My Student Survey (Balch, 2012; Voight & Hanson, 2012), the secondary version of Student Evaluation of Educational Quality (Marsh, Dicke, & Pfeiffer, 2019), and Panorama Student Survey (Delyser, Mascio, & Finkel, 2016; Panorama Education, 2015).

Though research generally supports the construct validity of these instruments as measures of teacher effectiveness (Kuhfeld, 2017; Wallace, Kelcey, & Ruzek, 2016), the definition of the construct itself is still controversial (Kane, McCaffrey, Miller, Staiger, & Foundation, 2013; Klassen & Kim, 2019). Researchers agree that teacher effectiveness is a multi-dimensional concept that involves a set of within-person attributes, including personality, motivation, beliefs, and disposition, which affect student outcomes while interacting with contextual factors (Klassen & Kim, 2019; Seidel & Shavelson, 2007). Though most student raters are not as knowledgeable as professionals at some aspect of teaching, such as curriculum design, content knowledge, etc.(Worrell & Kuterbach, 2001), they have extensive experience interacting with their teachers on a daily basis, which enables them to have a unique look at those attributes. Therefore, there is no agreement on what aspect of teaching should be included in student surveys and how those aspects should be structured.

One of the widely used models is a two-factor model consisting of *academic press* and *social support*, where *academic press* focuses more on the cognitive activation in academic learning while *social support* emphasizes the emotional and relational factor in the teacher-student interaction (Ferguson & Danielson, 2014; Lee & Smith, 1999).

Three-factor models that share similar concepts as those in the two-factor model were also proposed by some, with the addition of classroom management (Klieme, Pauli, & Reusser, 2009; Pianta & Hamre, 2009) or autonomy support (Schenke, Ruzek, Lam, Karabenick, & Eccles, 2017) as a third factor. Though models with seven or more factors was also proposed (Marsh et al., 2019; Ferguson & Danielson, 2014), there is conflicting empirical evidence that tends to support a two-factor structure instead (Kuhfeld, 2017; Phillips & Rowley, 2016; Wallace et al., 2016).

Further, the popular SET instruments show adequate criterion validity, predicting various outcome measures, such as teacher value-added scores, student academic achievement and academic self-concept (Scherer & Gustafsson, 2015; Wagner et al., 2016; Wallace et al., 2016; Worrell & Kuterbach, 2001). However, it should be noted that these outcomes are mainly based on tested academic subjects, and there is little literature on how student-rated teacher effectiveness is connected to outcomes in non-academic subjects. We do not even know whether we can compare the effectiveness of teachers in academic subjects and non-academic subjects on a fair basis.

Student Ratings of Teacher Effectiveness across Subjects and School Levels

The effectiveness of teachers in academic subjects is the major focus of current evaluation systems, while the effectiveness of those in non-academic subjects has long been ignored due to the fact that Physical and Health Education, Arts, or other district specialized subjects are not tested and therefore, not as many resources are allocated toward these subjects and the teachers (Bivona, 2012; Goe & Holdheide, 2011; Prince et al., 2009). That is to say, it may not be justified to use the student surveys designed mainly for academic teaching to evaluate teaching practices in non-academic subjects.

First of all, the way teachers interact with their students is different across different school subjects, which can be explained by differences in content knowledge, curriculum, and teaching objectives. Teachers in different subjects tend to use different instructional practices. For example, it is hard to imagine PHE teachers organizing group discussions as frequently as teachers in social studies. Besides, research shows that subject difference has a comparatively large effect on teachers' interpersonal behaviors when interacting with students (den Brok, Taconis, & Fisher, 2010). For example, science teachers are perceived as more pressing, less cooperative, and social science teachers are considered more uncertain (den Brok et al., 2010; Levy, Wubbels, den Brok, & Brekelmans, 2003; Telli, 2016). Since these differences are not relevant to teacher effectiveness, student ratings may not be valid as expected when comparing teachers across subjects.

Moreover, teachers in different subjects may be motivated and supported differently. Though motivation is typically considered as an important attribute of effective teachers, differences in teacher motivations across subjects and consequent differences in instructional quality may be due to the fact that system-school level support is subject specific (Klieme, 2013). For example, districts that use school-wide performance-based compensation systems take into limited or partial account the contribution of teachers in non-academic subjects. This makes those teachers "free-riders" in the system, and therefore, it is inevitable that the motivation of those teachers and the resources available to them within schools be compromised (Bivona, 2012; Goe & Holdheide, 2011). In other words, even if the measure of teacher effectiveness is

equally valid and comparable across subjects, it is still questionable whether teachers in different subject domains are evaluated on a fair basis.

Finally, since students may hold different beliefs and expectations for different school subjects, student ratings may be subject specific and measure slightly different constructs (Buehl, Alexander, & Murphy, 2002; Wagner et al., 2013). Moreover, the interpretation of the ratings may be different across subjects. For example, *press* is related to positive attitudes among students in academic subjects, but not in arts and sports (Telli, 2016).

As for school-level differences, since learning activities and classroom activities are different across grades, teachers at different school levels, i.e. elementary, middle and high schools, supposedly have different ways of interacting with students in the class (Gentry, Gable, & Rizza, 2002). Besides, students perceive worse relationships with their teacher as grade level increases (Breese, 2017). If we do not erroneously assume that teachers at higher school levels are inherently worse teachers, then it is questionable whether students at different school levels simply rate their teachers differently, and whether the difference is attributable at all to differences in teacher effectiveness.

Method

Sample

The current study involves anonymous students from 88 school districts in Missouri which employed NEE system in the 2017-2018 school year. Since students' anonymity and confidentiality in student ratings of teacher effectiveness is considered a necessary and practical element in teacher evaluation (Little, Goe, & Bell, 2009; Popham,

2013; Worrell & Kuterbach, 2001), the NEE system does not track any identifiable information of the students. Therefore, the exact number of participating students are unclear due to cases where one student may rate multiple teachers. In total, 29395 student ratings of 1398 teachers were included in the analysis , including 561 non-academic teachers, and 837 academic teachers, or 215 elementary teachers (Grade 4-6), 439 middle school teachers (Grade 7-9), and 744 high school teachers (Grade 10-12). Teacher subject type is defined either as academic or non-academic when at least 90 percent of the student ratings agree on the subject types, with ELA, mathematics, science, social studies and foreign languages considered as academic subjects while PHE, arts, or other district specialized subjects as non-academic subjects. The teachers were not further classified into specific subjects due to the fact that one teacher may teach multiple subjects within either subject type but rarely teach both academic and non-academic subjects at the same time in the data. The number of ratings for each teacher ranges from 3 to 169. Teachers in gifted programs and teachers in cross-level buildings were excluded from the sample.

Measures

The current study is based on the teacher effectiveness student survey (TESS), which is a modular survey with 110 items (including 3 screening items) nested in 25 independent components, or indicators, representing 25 different student-observable teaching practices that define effective teaching. The indicators are in alignment with InTASC teaching standards (Council of Chief State School Officers, 2011), which consists of the 10 most relevant evidence-based dimensions that define effective teaching. The Missouri Department of Elementary and Secondary Education adapted the InTASC

standards into *Missouri Model Teacher and Leader Standards* (Missouri Department of Elementary and Secondary Education, *2011*) which includes 9 standards with 39 teaching practices. TESS excluded 14 practices that are not observable by students (e.g., teacher self-assessment and improvement).

The initial item bank of TESS had been created by a team of experts in educational psychology and in survey administration, before cognitive interviews were conducted with students in 4th, 8th, 9th and 12th grades for further improvements. In the interviews, students were asked to explain their response and provide evidence to support themselves. Difficult words and unclear items were also properly adjusted based on the interviews.

The modular design of TESS addresses the multidimensionality of teacher effectiveness, and districts can select indicators that best represent their concerns and understandings of effective teaching. For the current study, the most popular 4 indicators (17 items) were selected, which were adopted by 44 percent of the schools within the network, i.e., cognitive engagement, problem-solving and critical thinking, teacher-student relationships, and instruction monitoring. For all the TESS items, students rate their teachers on a 4-point scale (0 = not true, 1 = sort of true, 2 = true, 3 = very true). The number of items also represents a typical length among surveys employed in the districts.

Cognitive engagement (CE). This indicator is one of those which examines how well a teacher uses content knowledge and perspectives aligned with appropriate instruction. Specifically, the items measure the degree to which a teacher cognitively engages students in the content in their teaching practices (e.g., “This teacher expects us

to think a lot and concentrate in this class"). 4 items are included (α = .875), and higher scores indicate more perceived practices in cognitively engaging students.

Problem-solving and critical thinking (PC). This indicator measures the extent to which a teacher uses instructional strategies that lead students to problem-solving and critical thinking in teaching practices (e.g., This teacher asks "how?" and "why?" questions to make us think more."). 4 items are included (α = .888), and higher scores indicate more noticeable teaching practices that induce problem-solving and critical thinking activities among students.

Teacher-student relationships (TSR). This indicator is one of those in TESS that examines whether a teacher successfully creates a positive classroom learning environment. Specifically, TSR measures the degree to which the students perceive secure relationships with a teacher (e.g., "This teacher knows me and cares about me."). 5 items are included in this indicator (α = .956), and higher scores represent more perceived secure teacher-student relationships.

Instruction monitoring (IM). This indicator involves whether teachers engage in formative feedback during a lesson to monitor learning at the individual and whole-class level and adjust their teaching (e.g., "This teacher explains the lesson in different ways if we don't get it at first"). 4 items are included (α = .954), and higher scores indicate better practices of a teacher in monitoring student learning.

Screening items. TESS includes 3 screening items that are evenly distributed among the survey items (e.g., "I am being totally honest on this survey."), which may

help to improve survey validity and identify inattentive responses. (Cornell, Klein, Konold, & Huang, 2012).

Procedures

TESS was delivered online at the end of the school term within an accessible time period specified by principals. Students entered an access code unique to each teacher to ensure they were evaluating the right teacher and at the same time prevent unauthorized access to the evaluation. Students remained anonymous throughout the entire evaluation procedures and their responses are confidential. An adult other than the evaluated teacher must administer the survey using standard administration scripts provided by NEE. The proctor read instructions to the students, informed them of the purpose of the survey, the anonymity of their responses, the voluntary nature of the evaluation, and how this evaluation is important for school improvement. Students were encouraged to ask questions, or request the explanation of difficult words, but the proctors were instructed not to interpret any survey items to avoid possible influences on students' responses. Further, responses that were finished within unrealistic time (three standard deviations from the mean) were flagged for manual review.

Though the online survey interface was designed to be intuitive and easy to use for students, the proctors also received simple training for potential technical issues. It was a common practice that all students in a building are included in the evaluation procedure; however, districts make independent decisions on the inclusion policy. That is to say, we are not sure how students are sampled to evaluate individual teachers due to the fact that students are anonymous and voluntary, the districts and principals make their

own inclusion policies, and the number of students nested within each teacher is not included in the evaluation system.

Analysis

Since SET is clustered in nature, where students are nested within teachers, it is appropriate to aggregate the mean scores of the items at the teacher level (Pituch & Stevens, 2015). The scorings of individual teachers are based on the latent factor scores generated with a validated measurement model. Though some suggest using multi-level models for clustered data to address measurement error and achieve more statistical power (Huang & Cornell, 2015; Marsh et al., 2012), it is most beneficial when the factors are at different levels, which is not the case of the current study. In TESS, individuals are anonymous raters and therefore, individual-level factors are not available. By using a latent measurement model on the basis of aggregated item scores, the teacher-level measurement error can be accounted and the student-level measurement error may be attenuated, though variability within teacher level is ignored (Dunn, Masyn, Jones, Subramanian, & Koenen, 2015; Richter & Brorsen, 2006). However, when aggregating student-level items to form teacher-level measures, we also introduce sampling error in addition to measurement error, which is a function of the average agreement among students in the same class and the number of sampled students in each class (Marsh et al., 2012). In addition, considering that the sampling and inclusion policy of individual schools are unknown, we decide to adjust the factor scores controlling the number of raters for individual teachers, the teacher level intra-class agreement and their polynomial terms.

It should be noted that TESS indicators were created as separate stand-alone surveys instead of being based on latent factors. Therefore, the structure of the agglomerate measure was examined and validated with a two-step procedure, before the measurement invariance was examined across subject types and school levels. First, the number of the factor was examined with Exploratory Factor Analysis (EFA) using the combination of Kaiser's criterion (Kaiser, 1960) and parallel analysis (Horn, 1965). Second, A Confirmatory Factor Analysis (CFA) was used to validate the discovered factor structure by examining the fit indices of the measurement model. The likelihood ratio χ^2 was first examined, where $p > .05$ indicates a good model fit. However, since the χ^2 test was shown to be too restrictive (Hooper, Coughlan, & Mullen, 2008), other indices were more often used in accompany with χ^2, i.e., the comparative fit index (*CFI*), root mean square error of approximation (*RMSEA*) and root mean-square residual (*SRMR*). Good model fit is defined when the *CFI* > .95, *RMSEA* < .05 and *SRMR* < .05, and adequate model fit is defined as when the *CFI* > .90, *RMSEA* < .08 and *SRMR* < .08 (Browne & Cudeck, 1992; Hu & Bentler, 1999; MacCallum, Browne, & Sugawara, 1996).

We examined the measurement invariance following the procedures proposed by Schmitt and Kuljanin (2008). Accordingly, the configural invariance model was first examined, where only the factor structure was constrained across the focal grouping variables, i.e., subject types and school levels. The metric invariance model was tested, where factor loadings were constrained across the groups. Finally, the scalar invariance was examined where both the factor loadings and the intercepts were constrained. The measurement invariance was examined separately for subject types and school levels. To identify the model difference, we used the change in CFI ($\Delta CFI \geq -.01$) paired with the

change in RMSEA ($\Delta RMSEA \geq .015$) or the change in SRMR ($\Delta SRMR \geq .03$ for loadings and $\Delta SRMR \geq .01$for intercepts), as suggested by Chen (2007). It should be noted that since the group sizes are unequal in the current study, the model fit indices changes might be underestimated. However, the recommended cutoff points to reject measurement invariance given by Chen (2007) was on the basis of much smaller sample sizes, which might at the same time underestimate the appropriate cutoff points of the fit statistics, especially for $\Delta RMSEA$. Therefore, we keep using the recommended cutoffs.

Results

In general, the students were attentive in the rating process, as on average, all the ratings of individual teachers past at least two of the three screening tests. EFA results show that two factors are identifiable with the data, which is based on the agreement of Kaiser's criterion (Kaiser, 1960) and Horn's parallel analysis (Horn, 1965), $\lambda_1 = 11.11$, $\lambda_2 = 1.67$ and $\lambda_3 = 0.55$. In addition, the stringent cutoff value of .55 was used for the factor loadings (Tabachnick, Fidell, & Ullman, 2007). *Table 1* gives the psychometric properties of the agglomerate measure that consists of the four most popular indicators in TESS.

Items from cognitive engagement (CE) and Problem-solving and critical thinking (PC) form the first factor, while items from Teacher-student relationships (TSR) and Instruction monitoring (IM) form the second factor. Unsurprisingly, the first factor represents *cognitive press*, and the TSR items represent *social support* in the two-factor model. However, a closer look at the IM items reveals that from the students' perspective, monitoring instruction is typically realized by interaction with and showing care to students . For example, the item IM_3 ("This teacher checks often to make sure

we understand the lesson as we go along.") may be understood by students as how the teacher cares about whether they understand the lesson.

Table 1. Factor Structure and Loadings of Selected TESS Items

ID	Factor	Factor Loadings	
		CP	SS
CE_1	Cognitive Press (CP)	.709	
CE_2		.905	
CE_3		.903	
CE_4		.686	
PC_1		.778	
PC_2		.567	
PC_3		.690	
PC_4		.580	
TSR_1	Social/ Emotional Support (SS)		.834
TSR_2			.975
TSR_3			.925
TSR_4			.919
TSR_5			.955
IM_1			.625
IM_2			.599
IM_3			.633
IM_4			.679

Note: 1. Results are based on Principal Axis Factoring with Oblimin Rotation.
2. The Factor correlation $\rho_{cp\_ss} = .679$.

Initial CFA indicates that items TSR_2 ("Students enjoying being with this teacher.") and TSR_5 ("This teacher is friendly.") show unaccounted covariance with multiple items, probably because both are comparatively vague and general in concepts. Though they may perform well in the standalone TSR survey, both were excluded from the measurement model for more accurate factor conceptualization. Moreover, covariance estimates were added to the model among the error terms of TSR items and between item CE_2 and CE_3 due to highly similar concepts and wordings. *Figure 1* is

the simplified demonstration of the final measurement model with standardized parameter estimates.

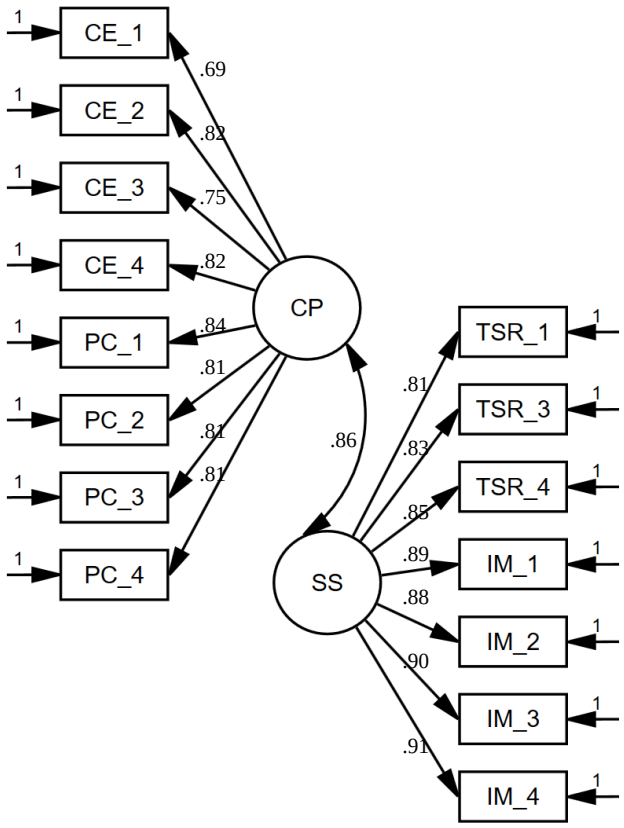


Figure 1. Measurement Model of Select TESS Indicators

CFA results showed that the measurement model provided an adequate to a good fit for the data, and all the factor loadings and factor variances were statistically significant. The model fit statistics are provided in *Table 2*. After the measurement model was first examined without grouping variables included (the total model), the measurement model was further probed with multiple group analysis with subject types as the grouping variable. The multiple group analysis starts with the configural model where only grouping variables were included and no constraints were applied. Then

constraints were applied across subject types on the measurement weights and intercepts consecutively to examine the metric and scalar models. A similar procedure was followed to examine the measurement invariance among school levels. Model fit statistics for the constrained models are also provided in the table.

Though the goodness of fit index between the metric and scalar model seem large when comparing the measures across subject types, $\Delta CFI = .22$, none of the changes of the badness of fit indices cross the rejection threshold, $\Delta RMSEA = .001$ and $.009$, and $\Delta SRMR = .007$ and $.002$. Therefore, using the criterion outlined in the previous section, measurement invariance was verified between academic teachers and non-academic teachers. This means that we can compare the teacher effectiveness scores across different subject types using the agglomerate measure. As for the comparison among school levels, though measurement invariance seems to be verifiable, $\Delta CFI \leq .001$, the drop of CFI was remarkable and $RMSEA$ falls out of the acceptable range. A possible explanation is that the model complexity reduced the model fit. Therefore, a pairwise two group analysis was conducted among the school levels to further examine the measurement invariance with reduced model complexity.

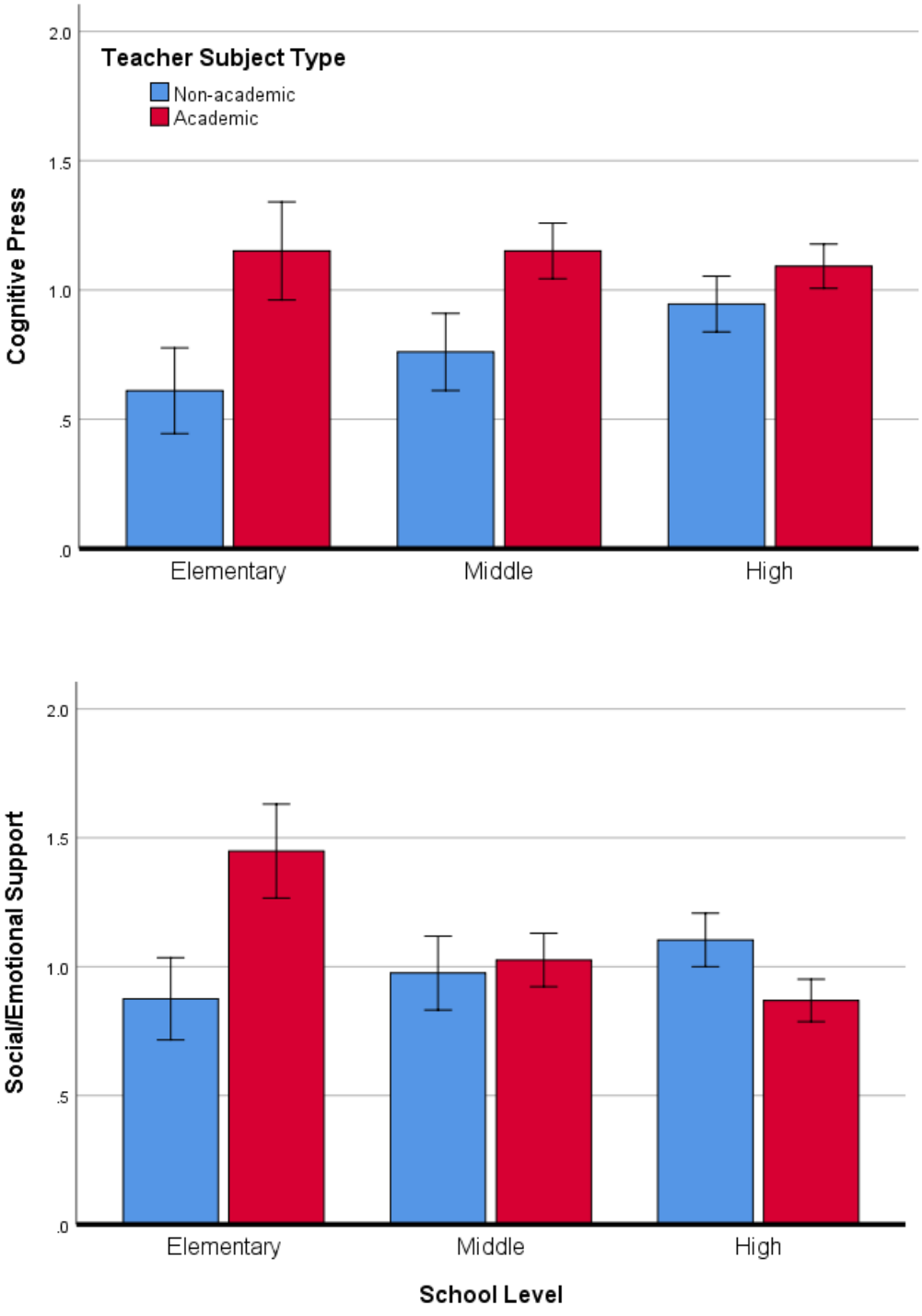
The results are also provided in Table 2, which shows that both *CFA* and *RMSEA* were remarkably recovered. Similar to the three group analysis, though the goodness of fit index difference is large between the metric and scale models when elementary teachers are compared with middle school teachers, $\Delta CFI = .009$ and $.021$, neither badness of fit indices is off the criterion range, $\Delta RMSEA = .001\ and\ .010$, $\Delta SRMR = .012$ and $.010$. A similar pattern was observed when comparing elementary and high-school teachers, $\Delta CFI = .001$ and $.026$, $\Delta RMSEA = .015$ and $\Delta SRMR \leq .01$. On the

other hand, the measurement invariance between middle and high-school levels was much clearly verifiable since the *CFA*s are very close. That is to say, by using the four indicators, the latent factor scores of teacher effectiveness are comparable across subject types and school levels.

Table 2. CFA Results of Model Fit and Measurement Invariance

Model		Constraints	N	*df*	χ^2 (*p*)	*CFI*	*SRMR*	*RMSEA*
Total		-	1398	85	838.327 (.000)	.963	.037	.079
Subject Types	Configural	Structure	1398	170	1080.413 (.000)	.955	.042	.062
	Metric	Loadings		185	1168.823 (.000)	.952	.049	.062
	Scalar	Intercepts		200	1624.950 (.000)	.930	.051	.071
All School levels	Configural	Structure	1395	305	2316.782 (.000)	.904	.104	.069
	Metric	Loadings		320	2360.080 (.000)	.903	.093	.068
	Scalar	Intercepts		335	2394.337 (.000)	.902	.093	.066
Elementary vs. Middle	Configural	Structure	654	170	588.420 (.000)	.949	.052	.061
	Metric	Loadings		185	615.175 (.000)	.940	.064	.060
	Scalar	Intercepts		200	829.261 (.000)	.919	.074	.070
Middle vs. High	Configural	Structure	1183	170	1033.995 (.000)	.954	.039	.066
	Metric	Loadings		185	1086.415 (.000)	.952	.056	.064
	Scalar	Intercepts		200	1185.542 (.000)	.947	.054	.065
Elementary vs. High	Configural	Structure	959	170	889.062 (.000)	.953	.052	.067
	Metric	Loadings		185	922.341 (.000)	.952	.060	.065
	Scalar	Intercepts		200	1320.161 (.000)	.926	.068	.077

Figure 2. Estimated Marginal Means of Teacher Effectiveness Scores



Note: The error bars represent 95% CI

A Two-way **ANOVA** was conducted to compare how teachers are different in *cognitive engagement* (CP) and *social/emotional support* (SS) across subject types and school levels. The effectiveness scores are the latent factor scores based on the measurement model while controlling for the number of ratings and within-teacher intro-class correlations (ICCs; two-way mixed absolute agreement). The results show that in general teachers in academic subjects were rated significantly higher on *cognitive press* than their non-academic counterparts, $F(1, 1392) = 33.48, p < .001$. However, no significant difference was observed across school levels, $F(2, 1392) = 1.93, ns$. *Figure 2* provides a visual demonstration of the mean teacher CP and SS scores across subject types and school levels with 95% Cis. The interaction between school level and subject type is also statistically significant on CP scores, $F(2, 1392) = 4.59, p = .010$, which is illustrated in *Figure 2*. An increasing trend of CP scores for non-academic teachers was observed across school levels while the CP scores for academic teachers are similar. A supplementary one way **ANOVA** indicates that non-academic teachers have significant CP score difference across school levels, $F(2, 558) = 5.12, p = .006$, though an *HSD* post hoc analysis shows that significant different only exist between non-academic teachers in elementary and high schools, $MD = -.34, p = .006$. On the other hand, academic teachers shows no difference in CP scores across school levels $F(2, 834) = .48, ns$. That is to say, the gap of the mean CP scores between academic and non-academic teachers was the largest in elementary schools and was gradually closed as school level increases.

As for teachers' SS scores, subject types, $F(1, 1392) = 5.41, p = .020$, school levels, $F(2, 1392) = 3.23, p = .04$, and their interactions, $F(2, 1392) = 16.90, p <$

.001 , all have significant effects on the student ratings of their teachers' SS scores. A supplementary one way **ANOVA** shows that SS scores are different across school levels for both academic teachers, $F(1, 834) = 15.39, p < .001$, and non-academic teachers, $F(1, 558) = 3.40, p = .034$. An *HSD* post hoc analysis further reveals that for non-academic teachers, SS difference exists only between elementary and high school level, $MD = -.23, p = .034$. On the other hand, SS score differences among academic teachers were only significant between elementary and middle schools, $MD = .42, p < .001$, and between elementary and high schools $MD = .58, p < .001$, but not between middle and high schools, $MD = .16, p = .064$.

Discussion

The current study examines the validity and factor structure of an agglomerate student survey measure of teacher effectiveness consisting of 4 most popular modules, or indicators, from TESS. The results indicate that the involved 17 items can form a valid measure of teacher effectiveness, which shows good psychometric properties and construct validity with a two-factor structure that conforms to relevant theories. It is not surprising to see that teachers in academic subjects tend to receive higher scores in *cognitive press* across school levels, which also indicates the adequate construct validity of the current measurement model.

The measurement invariance was also verified across subject types and school levels in the current study; however, it is not suggested that the factor scores be used to compare teacher effectiveness for high-states decisions across subject types and school levels because being theoretically comparable does not mean a fair comparison is

justified, especially considering that the less popular TESS items may bring in more differences across subject types and school levels. The measurement invariance only suggests that students across subject types and school levels tend to possess comparable conceptualization and standards for the latent factors. As mentioned in the previous section, the differences in latent scores may simply reflect the variations in available resources and consequent motivation, diverse content knowledge and curricula, different teaching objectives, or the combination of any of the factors that are not fair to be attributed to differences in teacher effectiveness.

CP scores were constantly higher for teachers in academic subjects simply because the learning activities are supposed to be more cognitively intense than non-academic ones, and therefore, teaching practices that aim to engage students in cognitive activities may be more frequently observed by students.

The observation that both mean CP and SS scores of academic teachers, in general, are constantly higher across school levels when compared to their non-academic counterparts may be explained by the fact that the academic subjects are typically tested subjects and therefore more resources and supports were given to those teachers, and at the same time they have more opportunities interacting with their students. A second possible explanation is that teaching standards and best practices are developed for core academic courses, and likewise the popularity of the indictors is a reflection of the systematic bias in favor of academic teachers. One exception of the pattern is that the mean SS score of high school non-academic teachers was higher than the academic teachers, which may be due to intense learning activities in academic classes and consequent lack of social/emotional support perceived by the students. Moreover, an

decreasing trend of SS scores is observable among academic teachers, which echoes a similar finding by Breese (2017).

Another interesting finding is the consistent increasing trend of both SS and CP scores among non-Academic teachers. It may be due to the fact that an increasing number of non-academic classes are electives across school levels, which means students gave higher ratings to their non-academic teachers simply because they like the teachers or subjects. This also explains why high school students tend to have better relations with their non-academic teachers. On the other hand, non-academic teachers in higher grades might also have higher motivation and better interactions with students, when student take their classes out of interest. This in turn may help the non-academic teachers have better teaching practices (Thoonen, Sleegers, Oort, Peetsma, & Geijsel, 2011).

In general, districts and schools should be very careful if they conclude based on a student survey that the non-academic teachers are not as responsible and effective as their academic counterparts. Even when using student survey for formative assessment, districts and schools should also be aware of the systematic differences exits across school levels and subjects. For example, the interpretation may be different in the same school, when a non-academic teacher is not cognitively engaging, from when an academic teacher shows a similar problem. Moreover, decision makers should be cautious when distributing resources using student survey as evidence because the difference may be the results of the unequal distribution of resources and, therefore, the further disparity of resources may further enlarge the gap.

Limitations and Future Studies

Since the study was based on mean aggregation at the teacher level, the individual-level measurement error may compromise the validity of the teacher evaluation scores, even when it may be attenuated in the mean aggregation. For example, student with better grades may tend to give higher ratings as observed in post-secondary education (Spooren & Mortelmans, 2006). Moreover, even when having considered the potential biases caused by sampling error, and having adjusted the factor scores in an effort to partition the sampling error (Schenke et al., 2017; Schenke, Ruzek, Lam, Karabenick, & Eccles, 2018), we cannot guarantee that the sampling error is fully partitioned due to the lack of knowledge how within-group agreement and group sample size affect the sampling error. For future studies, when outcome variables are available, such as teacher efficacy or teacher effectiveness based on classroom observation, teacher effectiveness based on student surveys can be examined to see how its connections to those outcomes are different across subject types and school levels.

Reference

Balch, R. T. (2012). *The validation of a student survey on teacher practice* (Doctoral Dissertation). Vanderbilt University Retrieved from Vanderbilt University Library Electronic Thesis and Dissertations. (Accession No.07182012-153440)

Bivona, L. (2012). Options for Including Teachers of Nontested Grades and Subjects in Performance-Based Compensation Systems. *American Institutes for Research (AIR) , Washington, DC*.

Breese, L. (2017). *How Students Perceive Their Relationships with Teachers*. Retrieved from https://blog.panoramaed.com/understanding-teacher-student-relationships-through-data/

Browne, M. W., & Cudeck, R. (1992). Alternative Ways of Assessing Model Fit. *Sociological Methods & Research, 21*(2), 230-258. doi:10.1177/0049124192021002005

Buehl, M. M., Alexander, P. A., & Murphy, P. K. (2002). Beliefs about Schooled Knowledge: Domain Specific or Domain General? *Contemporary Educational Psychology, 27*(3), 415-449. doi:https://doi.org/10.1006/ceps.2001.1103

Chen, F. F. (2007). Sensitivity of Goodness of Fit Indexes to Lack of Measurement Invariance. *Structural Equation Modeling: A Multidisciplinary Journal, 14*(3), 464-504. doi:10.1080/10705510701301834

Cornell, D., Klein, J., Konold, T., & Huang, F. (2012). Effects of validity screening items on adolescent survey data. *Psychological Assessment, 24*(1), 21-35. doi:10.1037/a0024824

Council of Chief State School Officers. (2011). InTASC model core teaching standards: A resource for state dialogue. *Washington, DC: Council of Chief State School Officers*.

Delyser, L. A., Mascio, B., & Finkel, K. (2016). *Introducing student assessments with evidence of validity for NYC's CS4All.* Paper presented at the ACM International Conference Proceeding Series.

den Brok, P., Brekelmans, M., & Wubbels, T. (2004). Interpersonal Teacher Behaviour and Student Outcomes. *School Effectiveness and School Improvement, 15*(3-4), 407-442. doi:10.1080/09243450512331383262

den Brok, P., Taconis, R., & Fisher, D. (2010). How well do science teachers do? Differences in teacher-student interpersonal behavior between science teachers of other (school) subjects. *The Open Education Journal, 3*, 44-53.

Downer, J. T., Stuhlman, M., Schweig, J., Martínez, J. F., & Ruzek, E. (2014). Measuring Effective Teacher-Student Interactions From a Student Perspective A Multi-Level Analysis. *The Journal of Early Adolescence, 35*(5-6). doi:doi.org/10.1177/0272431614564059

Dunn, E. C., Masyn, K. E., Jones, S. M., Subramanian, S. V., & Koenen, K. C. (2015). Measuring psychosocial environments using individual responses: an application of multilevel factor analysis to examining students in schools. *Prev Sci, 16*(5), 718-733. doi:10.1007/s11121-014-0523-x

Missouri Department of Elementary and Secondary Education. (2011). *Missouri Model Teacher and Leader Standards: A Resource for State Dialogue*. Retrieved from https://dese.mo.gov/sites/default/files/StandardsInformationDocument.pdf

Fauth, B., Decristan, J., Rieser, S., Klieme, E., & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1-9.

Ferguson, R., & Danielson, C. (2014). How Framework for Teaching and Tripod 7Cs Evidence Distinguish Key Components of Effective Teaching. In T. Kane, K. Kerr, & R. Pianta (Eds.), *Designing teacher evaluation systems: New Guidance from the measures of effective teaching project* (pp. 98-143). San Francisco, CA: Jossey-Bass.

Follman, J. (1992). Secondary school students' ratings of teacher effectiveness. *The High School Journal, 75*(3), 168-178.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal, 25*(1), 57-78.

Fraser, B., & McRobbie, C. (1995). Science Laboratory Classroom Environments at Schools and Universities: A Cross-National Study. *Educational Research and Evaluation, 1*(4), 289-317. doi:doi.org/10.1080/1380361950010401

Gentry, M., Gable, R. K., & Rizza, M. G. (2002). Students' perceptions of classroom activities: are there grade-level and gender differences? *Journal of Educational Psychology, 94*(3), 539. doi:dx.doi.org/10.1037/0022-0663.94.3.539

Goe, L., & Holdheide, L. (2011). Measuring Teachers' Contributions to Student Learning Growth for Nontested Grades and Subjects. Research & Policy Brief. *National Comprehensive Center for Teacher Quality, Washington, D.C.*

Hooper, D., Coughlan, J., & Mullen, M. R. (2008). Structural equation modelling: Guidelines for determining model fit. *Electronic Journal of Business Research Methods, 6*(1), 53-60.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179-185. doi:10.1007/BF02289447

Hu, L. t., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55. doi:10.1080/10705519909540118

Huang, F. L., & Cornell, D. G. (2015). Using Multilevel Factor Analysis With Clustered Data: Investigating the Factor Structure of the Positive Values Scale. *Journal of Psychoeducational Assessment, 34*(1), 3-14. doi:10.1177/0734282915570278

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement, 20*(1), 141-151. doi:10.1177/001316446002000116

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have We Identified Effective Teachers? Validating Measures of Effective Teaching Using Random Assignment.* Retrieved from https://files.eric.ed.gov/fulltext/ED540959.pdf

Kane, T. J., & Staiger, D. O. (2012). *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains.* . Retrieved from http://k12education.gatesfoundation.org/download/?Num=2678&filename=MET_Gathering_Feedback_Practioner_Brief.pdf

Klassen, R. M., & Kim, L. E. (2019). Selecting teachers and prospective teachers: A meta-analysis. *Educational Research Review, 26*, 32-51. doi:https://doi.org/10.1016/j.edurev.2018.12.003

Klieme, E. (2013). The Role of Large Scale Assessments in Research on Educational Effectiveness and School Development. In v. D. M., G. E., K. I., & Y. K. (Eds.), *he Role of International Large-Scale Assessments: Perspectives from Technology, Economy, and Educational Research*. Dordrecht: Springer.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras Study. Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janik (Ed.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137-160). Münster: Waxmann.

Kuhfeld, M. (2017). When Students Grade Their Teachers: A Validity Analysis of the Tripod Student Survey. *Educational Assessment, 22*(4), 253-274. doi:10.1080/10627197.2017.1381555

Lee, V. E., & Smith, J. B. (1999). Social Support and Achievement for Young Adolescents in Chicago: The Role of School Academic Press. *American Educational Research Journal, 36*(4), 907-945. doi:10.2307/1163524

Levy, J., Wubbels, T., den Brok, P., & Brekelmans, M. (2003). Students' Perceptions of Interpersonal Aspects of the Learning Environment. *Learning Environments Research, 6*(1), 5-36. doi:10.1023/A:1022967927037

Little, O., Goe, L., & Bell, C. (2009). *A practical guide to evaluating teacher effectiveness*. Retrieved from https://files.eric.ed.gov/fulltext/ED543776.pdf

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. doi:10.1037/1082-989X.1.2.130

Marsh, H. W., Dicke, T., & Pfeiffer, M. (2019). A tale of two quests: The (almost) non-overlapping research literatures on students' evaluations of secondary-school and university teachers. *Contemporary Educational Psychology, 58*, 1-18. doi:https://doi.org/10.1016/j.cedpsych.2019.01.011

Marsh, H. W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A. J. S., Abduljabbar, A. S., & Köller, O. (2012). Classroom Climate and Contextual Effects: Conceptual and Methodological Issues in the Evaluation of Group-Level Effects AU - Marsh, Herbert W. *Educational Psychologist, 47*(2), 106-124. doi:10.1080/00461520.2012.670488

MET Project (2012). *Asking students about teaching: Student perception surveys and their implementation*. Retrieved from http://k12education.gatesfoundation.org/download/?Num=2504&filename=Asking_Students_Practitioner_Brief.pdf

Panorama Education. (2015). *Validity brief: Panorama student survey*. Retrieved from https://go.panoramaed.com/hubfs/Panorama_January2019%20/Docs/validity-brief.pdf

Phillips, S. F., & Rowley, J. F. S. (2016). The Tripod School Climate Index: An Invariant Measure of School Safety and Relationships. *Social work research, 40*(1), 31-39. doi:10.1093/swr/svv036

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, Measurement, and Improvement of Classroom Processes: Standardized Observation Can Leverage Capacity. *Educational Researcher, 38*(2), 109-119. doi:10.3102/0013189X09332374

Pituch, K. A., & Stevens, J. P. (2015). *Applied multivariate statistics for the social sciences: Analyses with SAS and IBM's SPSS.* New York: Routledge.

Polikoff, M. S. (2015). The stability of observational and student survey measures of teaching effectiveness. *American Journal of Education, 121*(2), 183-212. doi:10.1086/679390

Popham, W. J. (2013). *Evaluating America's teachers: Mission possible?* Thousand Oaks, California: Corwin Press.

Prince, C. D., Schuermann, P. J., Guthrie, J. W., Witham, P. J., Milanowski, A. T., & Thorn, C. A. (2008). *The other 69 percent: Fairly rewarding the performance of teachers of nontested subjects and grades*. Washington, D.C. Retrieved from http://www1.gcsnc.com/whatmatters/pdf/other69Percent.pdf

Richter, F. G. C., & Brorsen, B. W. (2006). Aggregate Versus Disaggregate Data in Measuring School Quality. *Journal of Productivity Analysis, 25*(3), 279-289. doi:10.1007/s11123-006-7644-6

Roeser, R. W., Eccles, J. S., & Sameroff, A. J. (2000). School as a context of early adolescents' academic and social-emotional development: A summary of research findings. *The elementary school journal*, 443-471.

Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2017). Heterogeneity of student perceptions of the classroom climate: a latent profile approach. *Learning Environments Research, 20*(3), 289-306. doi:10.1007/s10984-017-9235-z

Schenke, K., Ruzek, E., Lam, A. C., Karabenick, S. A., & Eccles, J. S. (2018). To the means and beyond: Understanding variation in students' perceptions of teacher emotional support. *Learning and Instruction, 55*, 13-21. doi: 10.1016/j.learninstruc.2018.02.003

Scherer, R., & Gustafsson, J. E. (2015). Student assessment of teaching as a source of information about aspects of teaching quality in multiple subject domains: an application of multilevel bifactor structural equation modeling. *Front Psychol, 6*, 1550. doi:10.3389/fpsyg.2015.01550

Schmitt, N., & Kuljanin, G. (2008). Measurement invariance: Review of practice and implications. *Human Resource Management Review, 18*(4), 210-222. doi: 10.1016/j.hrmr.2008.03.003

Seidel, T., & Shavelson, R. J. (2007). Teaching Effectiveness Research in the Past Decade: The Role of Theory and Research Design in Disentangling Meta-Analysis Results. *Review of Educational Research, 77*(4), 454-499. doi:10.3102/0034654307310317

Spooren, P., & Mortelmans, D. (2006). Teacher professionalism and student evaluation of teaching: will better teachers receive higher ratings and will better students give higher ratings? *Educational Studies, 32*(2), 201-214. doi:10.1080/03055690600631101

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2007). *Using multivariate statistics* (Vol. 5): Pearson Boston, MA.

Telli, S. (2016). Students' perceptions of teachers' interpersonal behaviour across four different school subjects: control is good but affiliation is better. *Teachers and Teaching: Theory and Practice, 22*(6), 729-744. doi:10.1080/13540602.2016.1158961

Thoonen, E. E. J., Sleegers, P. J. C., Oort, F. J., Peetsma, T. T. D., & Geijsel, F. P. (2011). How to Improve Teaching Practices: The Role of Teacher Motivation, Organizational Factors, and Leadership Practices. *Educational Administration Quarterly, 47*(3), 496-536. doi:10.1177/0013161X11400185

Voight, A., & Hanson, T. (2012). Summary of Existing School Climate Instruments for Middle School. *Regional Educational Laboratory West.* Retrived from https://files.eric.ed.gov/fulltext/ED566402.pdf

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1-11. doi: 10.1016/j.learninstruc.2013.03.003

Wagner, W., Göllner, R., Werth, S., Voss, T., Schmitz, B., & Trautwein, U. (2016). Student and teacher ratings of instructional quality: Consistency of ratings over time, agreement, and predictive power. *Journal of Educational Psychology, 108*(5), 705-721. doi:10.1037/edu0000075

Wallace, T. L., Kelcey, B., & Ruzek, E. (2016). What Can Student Perception Surveys Tell Us About Teaching? Empirically Testing the Underlying Structure of the Tripod Student Perception Survey. *American Educational Research Journal, 53*(6), 1834-1868. doi:10.3102/0002831216671864

Worrell, F. C., & Kuterbach, L. D. (2001). The Use of Student Ratings of Teacher Behaviors With Academically Talented High School Students. *Journal of Secondary Gifted Education, 12*(4), 236-247. doi:10.4219/jsge-2001-670