Education Policy Analysis
Archives/Archivos Analíticos de Políticas
Educativas
ISSN: 1068-2341
EPAA@asu.edu
Arizona State University
Estados Unidos

Smith, William C.; Kubacka, Katarzyna
The Emphasis of Student Test Scores in Teacher Appraisal Systems
Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, vol. 25, 2017, pp. 1-26
Arizona State University
Arizona, Estados Unidos

SPECIAL ISSUE
Global Perspectives on High-Stakes Teacher Accountability Policies

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

epaa aape

Arizona State University

Volume 25 Number 86 August 21, 2017 ISSN 1068-2341

The Emphasis of Student Test Scores in Teacher Appraisal Systems

William C. Smith
&
Katarzyna Kubacka
UNESCO – Global Education Monitoring Report
France

Citation: Smith, W. C., & Kubacka, K. (2017). The emphasis of student test scores in teacher appraisal systems. *Education Policy Analysis Archives, 25*(86). http://dx.doi.org/10.14507/epaa.25.2889 This article is part of the special issue, *Global Perspectives on High-Stakes Teacher Accountability Policies*, guest edited by Jessica Holloway, Tore Bernt Sorensen, and Antoni Verger.

Abstract: Over the past 30 years teachers have been held increasingly accountable for the quality of education in their classroom. During this transition, the line between teacher appraisals, traditionally an instrument for continuous formative teacher feedback, and summative teacher evaluations has blurred. Student test scores, as an 'objective' measure, are increasingly used in teacher appraisals in response to historic questions that evaluations are based on 'subjective' components. Their central position in appraisals is part of a larger Global Testing Culture, where standardized tests are linked with high stakes outcomes. Although most teacher appraisal systems are based on multiple components, the prominence of testing as the taken for granted measure of quality suggests that not all components are given equal weight or seen as equally important. This article further explores the role of testing in high stakes teacher appraisal systems across 33 countries using data from the 2013 TALIS; addressing both the prominence of student test scores and their relative importance in teacher's perceived feedback utility. Results indicate that,

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 10/1/2017
Revisions received: 7/8/2017
Accepted: 7/8/2017

while rarely applied in isolation, student test scores are the most common component used in teacher appraisals. Relative to other components, student achievement is more often emphasized and, when emphasized in feedback, teachers are more likely to feel their appraisal had limited impact on their instruction and was completed solely as an administrative exercise.
Keywords: accountability; teacher appraisal; standardized test; Global Testing Culture; feedback, TALIS

El énfasis de los resultados de los exámenes estudiantiles en los sistemas de evaluación del profesorado

Resumen: A lo largo de los últimos 30 años, los profesores siempre han sido responsabilizados por la calidad de la educación en su aula. Durante esta transición, la línea entre evaluaciones de profesores, tradicionalmente un instrumento para retroalimentación continua de profesores formativos y evaluaciones de profesores sumativos, se tornó borrosa. Los resultados de los exámenes estudiantiles, como una medida "objetiva", son cada vez más utilizados en las evaluaciones de profesores en respuesta a cuestiones históricas de que las evaluaciones se basan en componentes "subjetivos". Su posición central en las evaluaciones forma parte de una cultura de prueba global mayor, donde las pruebas estandarizadas están vinculadas a resultados de altas participaciones. Aunque la mayoría de los sistemas de evaluación de profesores se basan en componentes múltiples, la prominencia de pruebas como la medida de calidad garantizada sugiere que no todos los componentes reciben igual peso o se consideran igualmente importantes. Este artículo analiza aún más el papel de las pruebas en sistemas de evaluación de profesores de alto riesgo en 33 países usando datos del 2013 TALIS; Abordando la prominencia de los resultados de los exámenes de los alumnos y su importancia relativa en la utilidad de retroalimentación percibida por el profesor. Los resultados indican que, aunque raramente se aplican aisladamente, los resultados de los exámenes de alumnos son el componente más común utilizado en las evaluaciones de los profesores. En relación a otros componentes, la realización de los alumnos se enfatiza con más frecuencia y, cuando se enfatiza en la retroalimentación, los profesores son más propensos a sentir que su evaluación tuvo un impacto limitado en sus instrucciones y se completó sólo como un ejercicio administrativo.
Palabras clave: Rendición de cuentas; Evaluación de profesores; Prueba estandarizada; Cultura global de las pruebas; Retroalimentación TALIS

A ênfase dos resultados dos exames estudantis em sistemas de avaliação de professores

Resumen: Ao longo dos últimos 30 anos, os professores sempre foram responsabilizados pela qualidade da educação em sua sala de aula. Durante esta transição, a linha entre avaliações de professores, tradicionalmente um instrumento para feedback contínuo de professores formativos e avaliações de professores sumativos, tornou-se desfocada. Os resultados dos exames estudantis, como uma medida "objetiva", são cada vez mais utilizados nas avaliações de professores em resposta a questões históricas de que as avaliações são baseadas em componentes "subjetivos". A sua posição central nas avaliações faz parte de uma cultura de teste global maior, onde os testes padronizados estão ligados a resultados de altas participações. Embora a maioria dos sistemas de avaliação de professores se baseie em componentes múltiplos, a proeminência de testes como a medida de qualidade garantida sugere que nem todos os componentes recebem igual peso ou são considerados igualmente importantes. Este artigo analisa ainda mais o papel dos testes em sistemas de avaliação de professores de alto risco em 33 países usando dados do 2013 TALIS; Abordando a proeminência dos resultados dos exames dos alunos e sua

importância relativa no utilitário de feedback percebido pelo professor. Os resultados indicam que, embora raramente sejam aplicados isoladamente, os resultados dos exames de alunos são o componente mais comum usado nas avaliações dos professores. Em relação a outros componentes, a realização dos alunos é enfatizada com mais frequência e, quando enfatizada no feedback, os professores são mais propensos a sentir que sua avaliação teve impacto limitado em suas instruções e foi completada apenas como um exercício administrativo.
Palavras-chave: Prestação de contas; Avaliação de professores; Teste padronizado; Cultura global de testes; Feedback TALIS

Introduction

Testing is a core practice in education. Regarded as a symbol of quality, testing permeates all aspects of education, shaping the experiences of the actors involved. Teachers, as the front line providers of education, are positioned to feel the brunt of the pressure when student test scores are the valued outcome. The role of testing in teachers lives is part of the larger Global Testing Culture (Smith, 2016a) where around the world education quality is being simplified into student measures on high-stakes standardized tests. Celebrated as seemingly objective measures, student test scores are increasingly used as a tool to evaluate teacher's performance and determine their future.

Teachers are commonly regarded as the main actors within schools, contributing to and shaping student development and learning (Jimerson & Haddock, 2015). At the same time, investments in teachers constitute the largest percentage of education budgets. Thus, it comes as no surprise that teachers are at the center of many education policy initiatives and reforms. Teacher appraisals, traditionally an instrument for continuous formative teacher feedback, are increasingly morphing into summative tools for high stakes accountability purposes. Student test scores, as the most 'objective' component used in appraisals are commonly used in high stakes decisions. In Portugal, for example, teacher salary scales were redesigned in 2007 to include student test scores as an indicator of teacher performance (Barnes et al., 2016). In 2010, Denmark instituted national standardized tests. The provision of school level results through the mandatory *Quality Report* increased accountability pressure on teachers and school leaders alike (Andreasen et al., 2015). In teacher appraisals, the application of high stakes based on student test scores and the narrow attention paid to student test scores in feedback can impact teacher's perceived utility of the appraisal and ultimately their motivation and satisfaction.

This article explores the central role of student test scores in teacher appraisal systems using a cross national data set from 33 countries. It examines the prominence of student test scores among other components of teacher appraisals and the relative importance given to test scores in teacher feedback. The study continues in four sections. First, the literature review introduces the Global Testing Culture and looks at the (in)distinction between teacher appraisal systems and summative teacher evaluations. This is followed by a data and methods section introducing key definitions and the hierarchical generalized linear model used to examine factors associated with feedback utility. The results section provides an analysis of the common components in test-based high stakes appraisal systems, explores the prominence of test scores within high stakes appraisal systems, and illustrates how overemphasis on test scores can have detrimental effects on teachers perception of feedback utility. Finally, the conclusion section situates the overall findings within the Global Testing Culture, identifies country specific outliers, and suggests areas for future research.

The Importance of Testing and the Global Testing Culture

The use of standardized tests in education has increased sharply over the last 50 years (Smith, 2014) with national or state testing systems seen as "an important, perhaps the key, strategy for improving education quality" (Chapman & Snyder, 2000, p. 457). Student test scores, as a measure of student performance, are embedded in many forms of accountability as a seemingly objective measure of quality (Henry & Gutherie, 2016). This is reflected in what some have called a Global Testing Culture with standardized test scores aggregated at the classroom or school level to apply high stakes to schools or teachers (Smith, 2016b). The taken for granted acceptance of testing as the correct measure of quality in a pressurized world of increased high stakes lays out norms for all actors, shaping their behaviour and public opinion.

Under a Global Testing Culture we see less diversity in the practical uses of testing. For example, student examinations that used to be designed to make decisions about student advancement and teacher competency tests, which historically have been used for pre-service teacher certification, are now more commonly used for multiple purposes, including holding educators accountable (Smith, 2014). Even amongst international assessments, the line between formative and summative purposes is increasingly blurred as the world moves toward more accountability. For instance, the Early Grade Reading Assessment (EGRA) developed by RTI and implemented in at least 60 countries (UNESCO, 2015) was originally intended to be a formative assessment, designed from a curriculum based measurement model. However, the purpose of EGRA quickly shifted from providing feedback to teachers to monitor in class progress to providing summative snapshots at the country level (Ticha & Abery, 2016). Based in part on the well documented power of Programme for International Student Assessment (PISA) (Meyer & Benavot, 2013; Pons, 2017), summative scores on international assessments often define education quality for a country (Murgatroyd & Sahlberg, 2016).

Additionally, the Global Testing Culture shapes what is acceptable and what is possible. The public expects the government to administer tests to demonstrate their competency (Kijima & Leer, 2016) and maintain quality standards (Smith, 2017b). Furthermore, testing can consume parents, students, and the larger community. For example, in South Africa, the year 12 matric test is so engrained that families situate their life around the 'matric year' and no other purpose for education is imagined (Balwanz, 2016). Teachers under increasing pressure to raise student test scores are more likely to use shortcuts or limit instruction to test specific content and activities (Allen et al., 2016; Somerset, 2016). The Global Testing Culture also shapes how teachers see themselves and their peers. Test-based high stakes accountability is associated with increased anxiety and feelings of shame (Certo, 2006; Larsen, 2005) as well as the branding of teachers based on their effort on test improvement (Booher-Jennings, 2005).

Centrality of Teachers and the Call for Greater Accountability

Evidence identifies quality teaching as vital for student learning (Darling-Hammond, 2000; Rockoff, 2004). The importance of teacher quality for student education is often used as an argument for increasing teacher accountability (Duke, 1995). Over the past 30 years teachers have been held increasingly accountable for the quality of education in their classroom (Volante, 2007) through an emphasis on managerialism which has often led to the erosion of trust in the teaching profession (Fitzgerald, 2008; Whitford, 2013). In addition, teachers are often held up as the problem in struggling education systems (Bantwini & King-McKenzie, 2011; Goldstein, 2011; Kumashiro, 2012). For example, in Turkey, after scores on PISA showed no improvement between 2003 and 2006, the Ministry of National Education focused the blame on poorly qualified teachers who lacked the skills to implement their new curriculum (Gur et al., 2012).

The increased spotlight on teachers comes at a time when teacher roles have expanded to school counsellor, curriculum developer, and researcher (Madden & Lynch, 2014; O'Hare & Bo,

2010; Yan, 2012), making it challenging for teachers to provide their full energy and sufficient attention on quality instruction and student learning. However, this environment of potentially conflicting responsibilities and increasingly diverse classrooms had not slowed the march towards greater accountability placed on teachers. Teachers, more so than administrators, parents, or the government, are cast as the primary actor responsible and accountable for education today.

Public pressure on schools and education systems to show that they match the expectations of quality education has put demands on systems to document teachers' effectiveness and spurred policy-makers interest in using teacher accountability. In the United States and United Kingdom, education accountability gained momentum in the 1970s and 1980s, with teacher accountability playing an important role (Duke & Stiggins, 1986; McLaughlin & Pfeiffer, 1988). In addition, the increased availability of educational data, including large longitudinal datasets, and the use of the data to rank schools and systems, reinforces interest in teachers as the accountable party (Jackson et al., 2014). Research suggesting teachers' differ in their skills and their effects on student learning (Rivkin et al., 2005) further supports initiatives in many countries to reward schools based on student performance (Fullan & Mascall, 2000; Kim & Sunderman, 2005). Although teacher appraisals, and their subsequent feedback, have, at times been seen as something more informal focused on the formative development of teacher practices, Fullan and Mascall (2000) point out that appraisals are now "part of a political movement of accountability" where "teachers are seen as public servants who should be accountable for their work" (p.41).

The (in)Distinction Between Teacher Evaluations and Teacher Appraisals

Teacher appraisals have historically been considered the formative part of teacher evaluation systems, distinct from a final summative teacher evaluation linked to high stakes. Many researchers have noted a conflict between the more controlling role of evaluations as a tool of monitoring teacher performance and the supporting role of promoting teacher development, questioning whether these two roles can coexist (McLaughlin & Pfeiffer, 1988). In countries such as the United States, teachers were fearful and suspicious of teacher evaluations, and researchers have questioned the validity and reliability of its implementation in school districts. Some of challenges included: lack of evaluator competency, badly designed evaluation materials and too much focus on teachers relative to other stakeholders (Styles Johnston & Camp Yeakey, 1979). The frustrations and failures of summative teacher evaluation systems led to the re-imagining of teacher appraisals in the 1960s and 1970s as a continual process that could provide more timely feedback to teachers. Professional development was to be emphasized over strict monitoring (Shinkfield & Stufflebeam, 1995). Teachers generally lacked trust and failed to see the utility in summative evaluations, which generally failed to impact teacher practices or student learning (Danielson & McGreal, 2000). In contrast, the more inclusive nature of appraisals was linked to increased satisfaction and more reflective pedagogical decisions in the classroom (Danielson & McGreal, 2000).

Over time the number of components used in teacher appraisals has expanded. Among the most common elements used today are direct teacher observation and teacher's self-assessment. Classroom observations of teacher practices are seen as "key both for understanding the mechanisms linking classroom processes and desired improvements in student outcomes, and for informing formative and developmental feedback to guide teacher improvement efforts" (Martinez et al. 2016, p. 15). Self-assessments are seen by some as essential to increase teacher buy-in and engagement in the appraisal process and increase the likelihood that results are used for instructional purposes (Danielson, 2011). Examples of self-assessment can be found in Peru, where interviews about their assessment are used for evaluation, and in Switzerland, where teachers assess their teaching, interaction with peers, parents, and students, and participation in

professional development (Schmelkes, 2015). Additional elements used to appraise teachers include student surveys, teacher portfolios, measures of teacher's content knowledge, interviews with teachers, parent feedback, and indicators of student performance (OECD, 2014; Schmelkes, 2015).

For teacher appraisals, the importance of test scores to measure student performance emerged from longstanding critiques of traditional teacher evaluation systems (Marzano & Toth, 2013) which, according to Toch and Rothman (2008), were "superficial, capricious, and often don't even directly address the quality of instruction much less measure students' learning" (p.1). The Race to the Top (RTT) grant programme, created in the United States in 2009 to stimulate improvements in low-performing schools, illustrates one attempt to embed student test scores directly into teacher appraisals. As part of RTT, the U.S. Department of Education suggested that measures on student growth be included in evaluation systems that impact teachers' professional development and career progression (USDOE, 2009). The federal guidance for states applying for RTT funding consists of a mix of formative and summative purposes for teacher evaluation, melding teacher appraisals and teacher evaluations (Popham, 2013).

Increased managerialism, combined with importance of teachers in quality education, the dominance of teacher salary in national budgets, and the taken for granted equivalence of test scores and education quality, has contributed to the transformation of appraisals into summative instruments. Managerialism in education was part of the larger neo-liberal turn in education felt in the 1980s (Hursh, 2005). The belief that the private sector was more efficient than the public sector and sense that best practices can be transferred from one organization to another (Bottery, 1989), increased attention to cost-cutting and standards setting in education (Larsen, 2005). Student test scores have been increasingly embraced as a marker to direct funds appropriately based on standardized measures of quality. Managerialism also embraces the role of external evaluators, diminishing and distorting the internal reflections central to teacher appraisals, leaving teachers focused on whether they can "demonstrate publicly that they fulfil accountability requirements" (Larsen, 2005, p. 300).

High stakes are another reason why it is difficult to distinguish between teacher appraisals and teacher evaluations. High stakes are often applied as an incentive to motivate educator behavior (Smith, 2017a) and have been increasingly used for teacher appraisals as part of performance management policy (Evans, 2013). High-stakes appraisals, in general, have been an object of a fair amount of controversy and views on them tend to be polarizing (AERA, 2015). For proponents, linking teacher appraisal to teacher professional outcomes can be seen as a way to make the system more meaningful to teachers and stimulate teacher professional development, beyond holding them accountable (OECD, 2013b).

Critics, on the other hand, focus on the undesirable side effects of high stakes appraisals. Teachers express concerns about how their working conditions increase their stress in general and stress related to student testing (von der Embse et al., 2016), and negatively impact their motivation (Figazzolo, 2013). High-stakes, test-based approaches can also have important, unintended consequences in the classroom: with teachers narrowing the curriculum, teaching to the test or focusing on more talented students, at the cost of others, in order to boost test results (Darling-Hammond, 2015; Jennings & Sohn, 2016; UNESCO, 2014).

In addition to these issues, previous concerns on evaluation and feedback utility and teacher motivation have emerged. Teacher job satisfaction is associated with perceptions that the appraisal system is more than a mere administrative task (OECD, 2014). One of the problems with the transformation of appraisals into pseudo-evaluations is the potential lack of continuous feedback provided to teachers. One time, summative pieces of information are less likely to shape teachers practices (Ahsan & Smith, 2016). Greater perceived feedback utility effects motivation and is associated with increased openness to engage with and learn from the information received (Malik & Aslam, 2013; Mok & Zhu, 2014). For example, in a study using a sample of 1,983 teachers across 65 Flemish schools, Devaux and colleagues (2013) identified

perceived utility of feedback during post-appraisal interviews as the most important feature related to teacher pursuit of professional development.

Finally, some may argue that the negative effects associated with test-based high stakes appraisal systems may be partially mediated by taking a multi-metric approach to teacher appraisal. Amongst proponents, there is an emerging consensus that using multiple methods can be a more effective approach to appraisal than relying solely on one metric (Garrett & Steinberg, 2015; OECD, 2013b). This is due in part to the recognition that teaching is complex and multidimensional and a range of methods are needed to properly capture a more complete picture of teacher performance (Goe & Croft, 2009). Given the prominence of testing, however, questions arise on whether equal importance is given to all included components.

Current Research

This article further explores the role of testing in high stakes teacher appraisal systems, addressing both its prominence and its relative importance in perceived feedback utility. Included in this analysis is a mapping of teacher appraisal patterns across 33 national or regional education systems, which largely confirms the central position of student test scores as an appraisal component across countries. Specific questions addressed in this study include:

1. How common are the use of student test scores and high stakes in teacher appraisals?
2. How much importance is placed on different components when teachers receive feedback from their appraisal?
3. How does teacher's perception of feedback utility differ by the degree test scores are emphasized in their feedback?

Data and Methods

Data from the 2013 TALIS were used in this study. TALIS is a cross-national survey of teachers and school environments, focusing on lower secondary education. The initial release of data from the 2013 TALIS contained information from 33 countries or participating economies through teacher and principal questionnaires. The stratified samples are nationally representative, with teachers nested in schools. The pooled sample includes a total of 85,400 teachers. Missing data is dealt with through listwise deletion. Information from both the teacher and principal questionnaire is used to identify the stakes associated with teacher appraisals, the components included in the appraisal, the feedback provided to teachers, and teachers' perception of the feedback's utility.

Key definitions. According to Larsen (2005), teacher appraisals are high stakes if appraisal results are "tied to increases in salary, promotion and maintenance of employment" (p. 296). Using this definition as a basis, appraisals are identified as high stakes in this study if any of the following happen at least sometimes following a teacher appraisal: material sanctions such as reduced annual increases in pay are imposed, there is a change in a teacher's salary or a payment of a financial bonus, a change in the likelihood of a teacher's career advancement takes place, or the teacher is dismissed or contract is non-renewed.

Student test scores are one of six components included in teacher appraisals. An appraisal is considered *test-based* if it is used at the school regardless of which entity (external bodies, school management team) or individual (principal, mentor, other teachers) performed the task as part of the formal appraisal. *Test-based high stakes teacher appraisals* speaks to teacher appraisals that have both high stakes outcomes and are based, at least in part, on student test scores.

Test-based high stakes appraisal patterns. To identify the most common patterns of test-based high stakes teacher appraisals, both overall and across country patterns of components were identified. These patterns included the use of student test scores in high stakes decisions and at least one of the five other components – the inclusion of teacher observations, student surveys, assessments of teacher's content knowledge, teacher's self-assessment, and/or parent feedback – and resulted in 33 potential unique patterns or combinations.

Analyzing appraisal feedback. Principal responses were matched to responses in the teacher questionnaire to examine whether teachers receive feedback and what part of the feedback is emphasized. Mirroring the six components included in teacher appraisals is a question asking teachers whether or not they received feedback on that component. No feedback is received on the component if the teacher responded with "I have never received this feedback in this school".

Although multiple components may be used in teacher appraisal they may not receive equal consideration in the high stakes decision. An overall ranking and a relative measure of importance is used to identify how much emphasis is placed on student performance, and thus student test scores, when teachers receive feedback on their appraisal. To identify which parts of teacher appraisal are emphasized, teachers are asked to evaluate eleven potential areas of feedback. Each area is coded on a Likert scale from not considered at all when feedback is received (1) to considered with high importance (4). Some of the factors can be mapped directly onto a component; however, factors associated with self-assessment and teacher evaluation are harder to distinguish (see Table 1). The *overall ranking* ranges from 1 (most emphasized factor) to 11 (least emphasized factor) and is aggregated at the country level.

Table 1
Linking Appraisal Components to Associated Feedback

Appraisal Component	**Associated Feedback from Appraisal**	**Importance Placed on Feedback: Mean (SD)**
Test scores	Student achievement	3.47 (0.76)
Student surveys	Student feedback	3.14 (0.92)
Parent feedback	Parent feedback	2.93 (1.00)
Assessment of teacher's content knowledge	Knowledge and understanding of subject field	3.32 (0.84)
Teacher observations or self-assessments	Pedagogical competency	3.38 (0.79)
	Student assessment practices	3.28 (0.80)
	Student behavior and classroom management	3.36 (0.79)
	Teaching of students with special needs	2.89 (1.02)
	Teaching in a multicultural or multilingual setting	2.32 (1.08)
	Feedback teachers provide to other teachers	2.62 (1.05)
	Collaboration with other teachers	3.15 (0.89)

The *relative importance* is calculated by taking the difference between the score for student performance and the mean score of the other ten factors. For example, in the overall pooled sample student performance was the most emphasized factor with a score of 3.47 (out of 4). The difference between this score and the mean of all other factors (3.04) revealed a relative

importance score of 0.43. This measure of relative importance is used at the teacher level in the inferential analysis to predict teacher's perception of feedback utility.

Predicting teachers perception of feedback utility. The *relative importance* placed on student achievement in feedback is the primary independent variable used to examine the association between overemphasis on student test scores and teachers perception of feedback utility. Feedback utility is captured in teacher's sense of whether appraisal feedback *makes little impact* on their instruction and whether the appraisal feedback is used for *only administrative* purposes. Teacher responses to the statements "teacher appraisal and feedback have little impact upon the way teachers teach in the classroom" and "teacher appraisal and feedback are largely done to fulfil administrative requirements" were coded as binary variables (agree/strongly agree = 1; disagree/strongly disagree = 0). Overall, 43.1% of teachers agreed that the appraisal had little impact and 50.0% of teachers felt that the process was simply an administrative task. Teachers *sex* (68.1% female), *age* (mean = 42.5, sd = 10.5), *contract status* (81.4% on permanent contract), *years of experience* (mean = 16.1, sd = 10.3), and *education level* (2.2% less than ISCED 5) are included as control variables in the analysis.

Given the dichotomous measures of the feedback utility outcome variables (feedback *makes little impact* and feedback was *only administrative*), hierarchical generalized linear modeling (HGLM) was the method used in this analysis. A HGLM acknowledges the nested, or hierarchical, nature of data (Raudenbush & Bryk, 2002), adjusting the standard error as necessary. This is necessary given the likelihood that teachers in the same school and in the same country are more similar than their peers in different schools or countries. The xtmelogit command in Stata version 13 was used for the analysis. Odds ratios are provided to ease interpretation of results.

The complete random intercept HGLM model it illustrated in Equation 1, which predicts feedback utility for *i* teacher in *j* school in *k* country. The equation is replicated, using both *makes little impact* and *only administrative* as separate dependent variables capturing the general concept of feedback utility (see Table 2 for HGLM results). Teacher level variables include the primary independent or predictor variable (β_{1jk}) and control variables (β_{2jk} - β_{6jk}). Also included are the initial intercept (δ_{000}) and error terms for the country (ν_{00k}), school (u_{0jk}), and teacher level (e_{ijk}).

$$\text{Feedback Utility}_{ijk} = \delta_{000} + \beta_{1jk}(\text{Relative Importance}) + \beta_{2jk}(\text{Female}) + \beta_{3jk}(\text{Age}) + \\ + \beta_{4jk}(\text{Contract Status}) + \beta_{5jk}(\text{Years of Experience}) + \beta_{6jk}(\text{Education Level}) + \nu_{00k} + u_{0jk} + e_{ijk}.$$

Equation 1

Results

Components and Stakes of Teacher Appraisals

Student test scores are the most commonly used component in teacher appraisals. Nearly 97% of teachers in the TALIS sample work in schools that include student test scores in their teacher appraisal. The inclusion of student test scores ranked just above teacher observations (96%), with assessments of teacher content knowledge being the least commonly included component (78%). Furthermore, approximately 79% of teachers work in schools that have high stakes consequences associated with teacher appraisal. Figure 1 charts the inclusion of student test scores and stakes of the appraisal by country. In the bottom left quadrant include countries, such as Italy and Portugal, whose teachers are less likely to be in a high stakes appraisal system and are less likely to have student test scores used in their appraisal, relative to the overall mean. Additional outliers include Finland, the only country in the sample where less than 80% of

teachers have student test scores included as a component in their appraisal, and Mexico, Spain, and Japan, where less than half of their teachers are in schools which use appraisals for high stakes decisions.

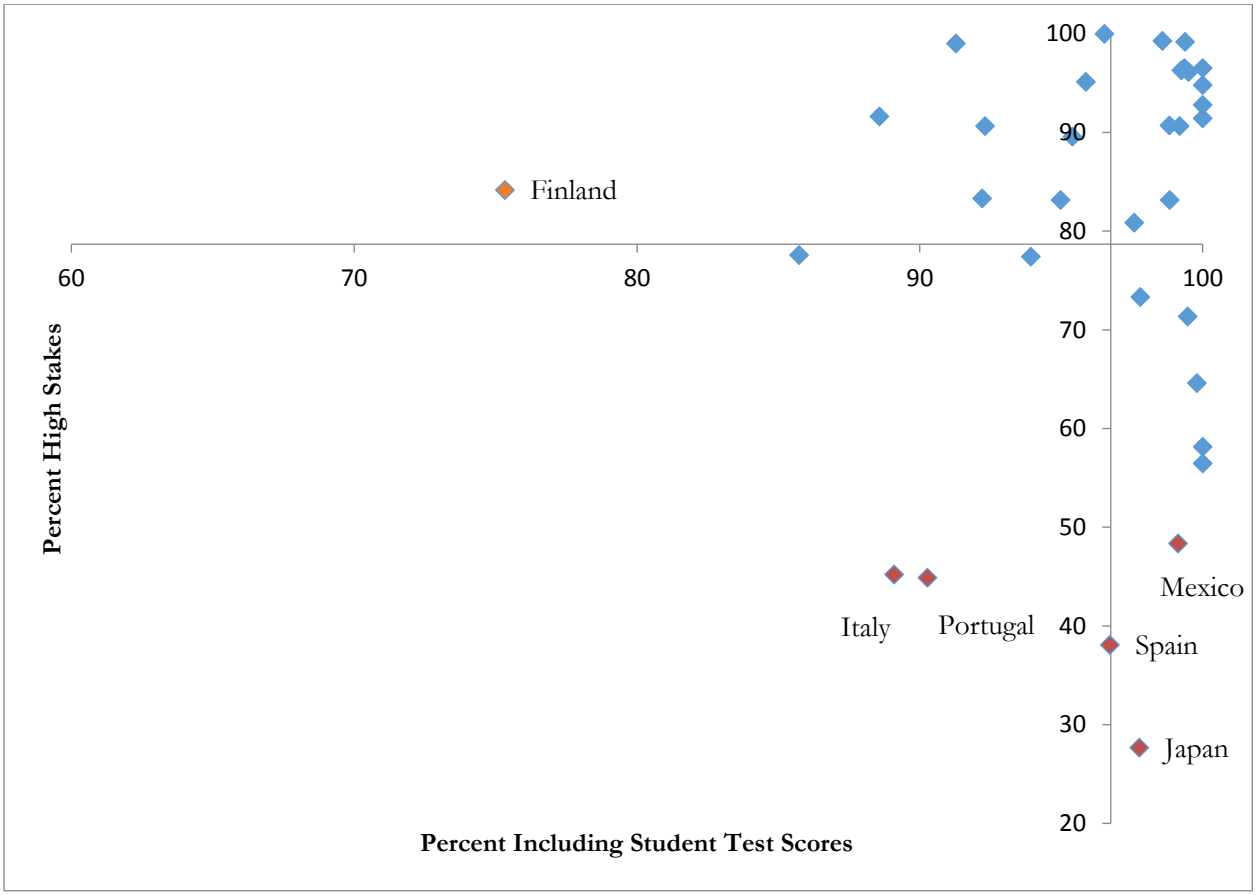

Figure 1. Cross National Differences in Inclusion of Student Test Scores and Stakes of Teacher Appraisals

Note: X and y-axis intercept set at the overall mean.

Although it is possible that appraisal systems that incorporate student test scores do not use them in high stakes decisions, this is rarely the case. Three out of four teachers in the sample work in a school that attaches high stakes to student test scores. Of those that work in high stakes systems, 97.3% of appraisals include student test scores as a component. The relationship between the inclusion of student test scores and the stakes of the teacher appraisal is statistically significant ($\chi^2 = 223.64$, df = 1, $p<.01$) in the pooled sample. Test-based high stakes appraisal systems are the focus for this article.

Figure 2 illustrates six common components used in teacher appraisals. The inclusion of student test scores is the most common component across all appraisals, as well as those with high stakes outcomes. There is little movement in the inclusion ranking of components in all versus just high stakes systems with student surveys moving from the fifth most common component across all appraisal systems (blue) to the fourth when just high stakes appraisals (orange) are considered, moving slightly above teacher self-assessments.

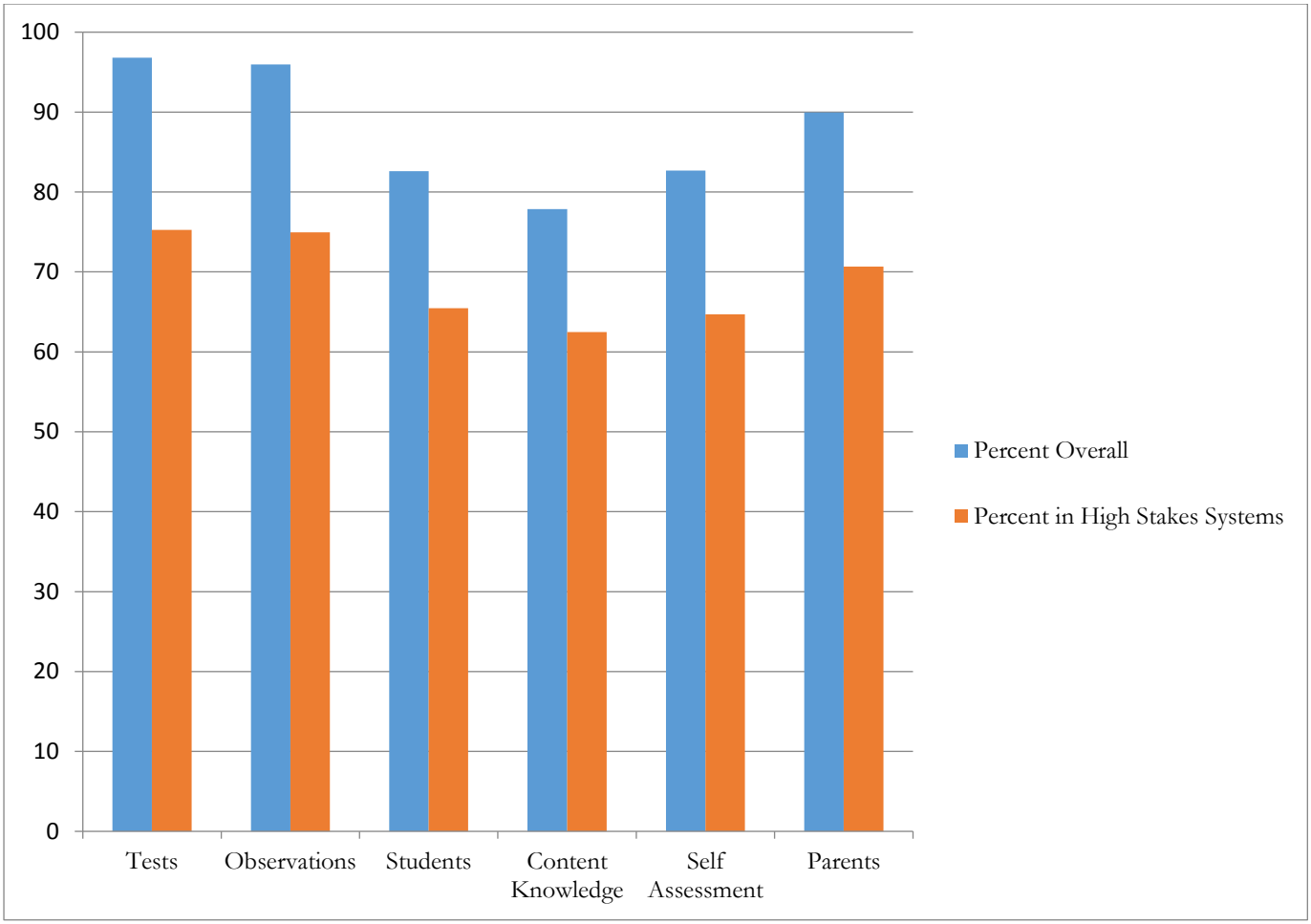

Figure 2. Components of Teacher Appraisals

Patterns of Test-based High Stakes Teacher Appraisals

Components included in high stakes appraisals are rarely done so in isolation. For instance, student test scores are used independent of other components in only 0.1% of teacher appraisals in the overall sample[1]. Commonly used patterns outlining the included components of test-based high stakes teacher appraisals can be derived from the data. Across the entire sample the majority of teachers work in schools that incorporate all six components in their appraisal. Out of the 33 potential patterns of test-based high stakes teacher appraisal, 63.3% include all components. The other 32 patterns combined represent less than 37% of teacher appraisals. Appendices A and B detail the top ten patterns of test-based high stakes teacher appraisals in the overall sample and top three patterns by country. Only one pattern was not used by any teacher appraisal; no appraisal was based on the combination of student test scores, student surveys, an assessment of teacher's content knowledge, and teacher's self-assessment.

Including multiple measures in test-based high stakes appraisals is the dominant practice. Among the top ten overall patterns student test scores and teacher observations both appear ten times, followed by parent feedback (8), teacher's content knowledge and student surveys (both 6), and teacher's self-assessment (5). Outside of basing high stakes decisions on the combination of student test scores and teacher observations (14th overall, 0.65%), all other two component combinations were ranked in the bottom 11 patterns. In this regards, Italy appears to be an

[1] High stakes teacher appraisals based only on student test scores are found in Brazil (0.47% of teachers) and Iceland (3.64% of teachers).

outlier, as the second most common pattern in the country consists of appraisals based on only student test scores and parent feedback (15.3%).

Diversity in appraisal patterns may suggest greater within-country autonomy, giving local administration the ability to craft appraisal systems. For example, France appears to be an interesting case as it is the only country where the most common pattern is found in less than one in five appraisals and is one of three countries where the three most common patterns represent less than 55% of all test-based high stakes teacher appraisals. France is one of the eight countries in the sample where at least 15 appraisal patterns are present (others include Australia, Brazil, Iceland, Israel, Portugal, Spain, and Sweden). However, even amongst countries with diversity in teacher appraisals, differences in the distribution of patterns remain. For instance, the two countries with the greatest number of patterns present (Brazil with 29 patterns and Iceland with 20 patterns present) appear different, as over 65% of teachers in Brazil work under the most common appraisal pattern and 18 patterns have less than 1% of teachers each while in Iceland the top pattern only includes 22% of teachers and all but two of the present patterns have greater than 1% of teachers. The variance present in these countries lies in contrast with Abu Dhabi (UAE) and Romania, where over nine in 10 teachers work in schools that use the most common pattern, and Latvia, where all teacher appraisals are captured in just three patterns.

While the vast majority of teachers work under test-based high stakes appraisals, it is important to recognize that in some countries this represents less than half of teachers. Specifically, in Mexico (47.5%), Portugal (40.2%), Italy (36.4%), Spain (35.8%), and Japan (26.5%) teachers tend to work in schools that do not include student test scores in their appraisals, do not make high stakes decisions based on their appraisals, or both. Therefore, the pattern breakdown for these countries represents the minority of teachers in the country.

Importance Placed on Appraisal Components

The presence of multiple components in a teacher's appraisal does not mean that each component has equal weight in the high stakes decision. Unfortunately, TALIS data cannot provide direct insight into how much relative weight is given to each appraisal component. As a proxy, teacher perception of emphasized appraisal feedback is used. For example, when teachers indicate that great importance is placed on student performance it suggests that student test scores are valued highly in the appraisal.

In the overall sample, student performance is the most emphasized piece of feedback from appraisals. Figure 3 suggests that although in all countries multiple measures are used in teacher appraisals, in practice the greatest importance is placed on student test scores. In 20 out of the 33 systems, student performance was the most emphasized factor in feedback. The y-axis in Figure 3 plots the relative importance of student achievement in appraisal feedback. Based on literature describing the test central education systems in England (UK) and the United States (Hursh, 2007; Lingard & Lewis, 2016; Smith, 2014), it is not surprising that student performance is not only the most emphasized factor in these systems but the relative importance of test scores in feedback is substantially larger than in other countries.

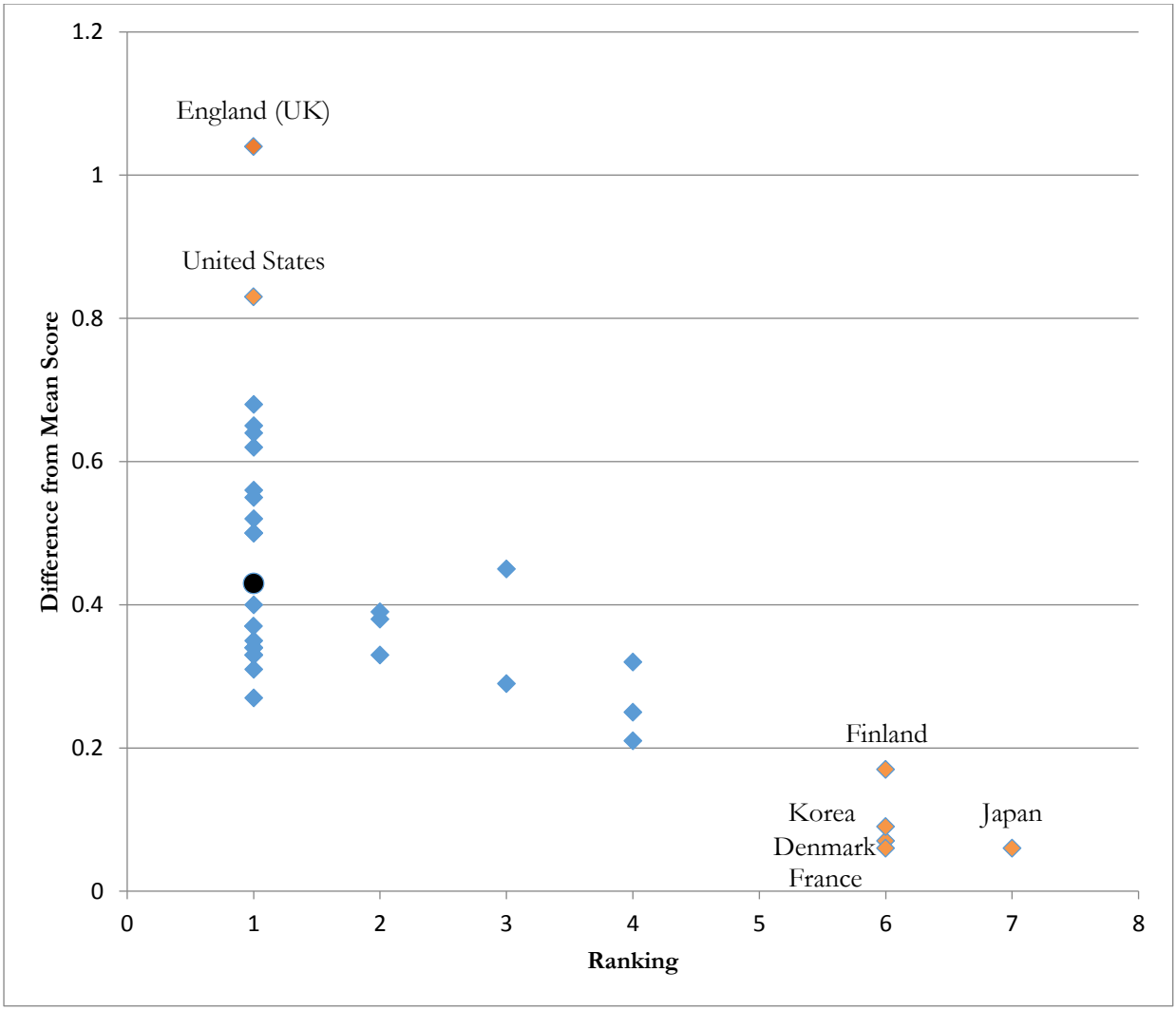

Figure 3. The Importance Placed on Student Performance in Teacher Feedback
Note: Black dot represent the rank and difference for the overall sample.

Although student test scores are included in the high stakes decisions of some teachers in Finland, South Korea, Denmark, France, and Japan, when feedback from teacher appraisals are received relatively less emphasis is placed on student performance. In all five countries student performance ranks as the six or seventh most emphasized factor. Additionally, the difference between student performance and mean of the other factors is close to zero[2]. In place of student performance relatively greater importance is placed on pedagogical competency in France (score = 3.65, rank = 1), Japan (score = 3.21, rank = 1), and South Korea (score = 3.30, rank = 1), while Denmark (score = 3.33, rank = 1) and Finland (score = 2.98, rank = 2) emphasize collaboration or work with other teachers.

The Importance and Use of Appraisal Feedback

From figure 3, it is clear that student achievement, generally reported in student test scores, is largely emphasized in appraisal feedback. Table 2 illustrates how the overemphasis on student test scores can influence teachers' perception of appraisal utility. Pooling data across all

[2] Differences with the mean score of other factors remain above zero due in large part to the very low scores in importance placed on teaching in a multicultural or multilingual setting.

countries, a three level random intercept HGLM is used to identify effects at the teacher level. Relative importance of student test scores in feedback is used to predict whether teachers believe the appraisal makes an impact on their teaching and whether it is only an administrative task. The analysis controls for teacher's sex, age, contract status, years of experience, and education level. Results are clustered at the country and school level to adjust for within country and within school similarities. Odds ratios are provided. Log odds ratio are available from the authors upon request.

Table 2
Predicting Feedback Utility from the Relative Importance Placed on Student Achievement Feedback

Appraisal Makes Little Impact		Appraisal is Only Administrative
Odds Ratio	**Variable**	**Odds Ratio**
.928**	Female	.939**
1.000	Age	.994**
1.211**	Permanent Contract	1.272**
1.005**	Years of Experience	1.005**
1.037	Education Level	.902
*1.144***	*Relative Importance*	*1.207***
.582	Intercept	.958

Notes: Pooled sample clustered at the country and school level. Sample size: makes little impact (n = 76,758); only administrative (n = 76,255). * = p<.05 ** = p<.01.

Teachers that feel student achievement is the most emphasized piece of feedback and that this feedback is disproportionately valued above other feedback options are more likely to perceive appraisals as an administrative tool that makes little impact on their classroom teaching. The odds that teachers believe the appraisal makes little impact is 1.14 times higher per point difference in emphasis on student achievement while the odds that the appraisals is purely administrative is 1.21 times higher per point. A one-point difference suggests that teachers that feel student performance is greatly emphasized in their feedback, relative to other components being moderately emphasized, finds their feedback to be less useful. Additionally, female teachers and older teachers perceive feedback to be of little use. Further characteristics associated with lower levels of feedback utility include permanent contracts and greater years of experience.

In addition to negative teacher perceptions about the utility of feedback, some teachers do not receive feedback. Feedback can be especially crucial when it comes from the component included in the high stakes decision. As Figure 4 makes clear, a large number of teachers do not receive any feedback on such components. On the high end 42% of teachers that work in an appraisal system that use parent feedback as an input into the high stakes decision do not receive any information about what the parents said. Similar high percentages are found for nearly all components.

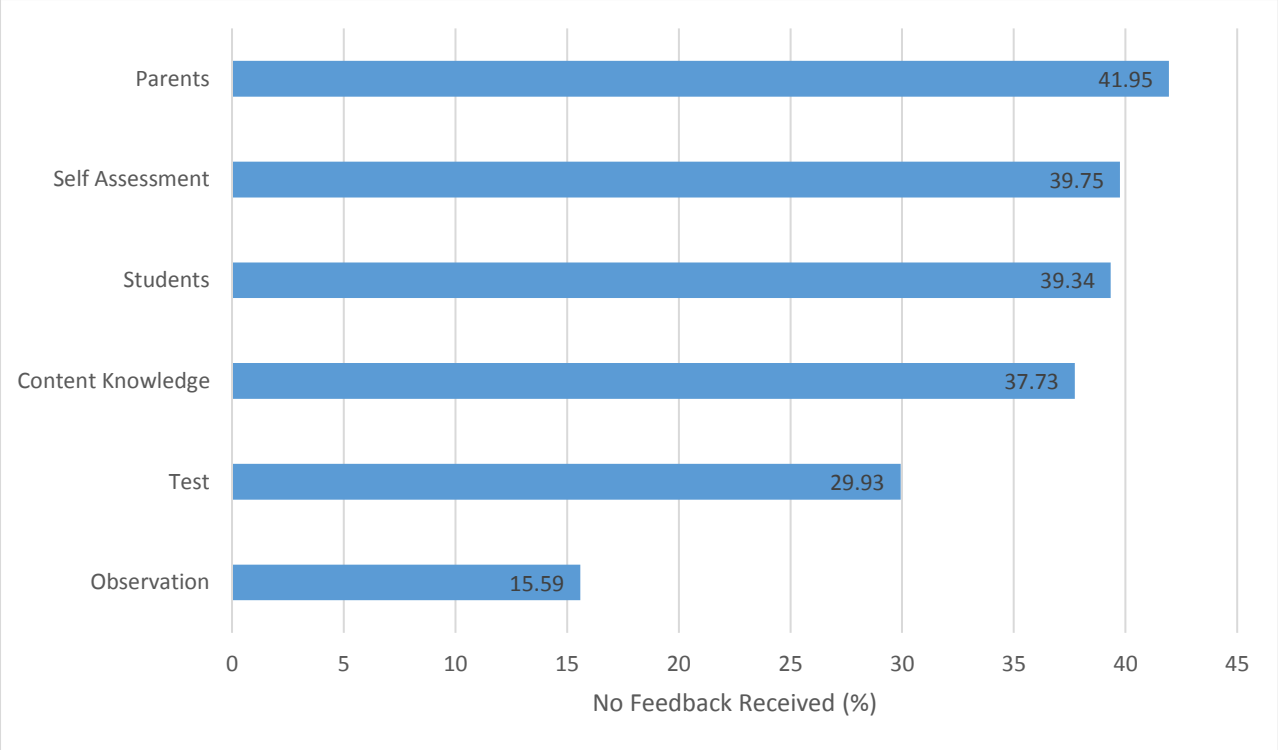

Figure 4. Percentage of Teachers that Do Not Receive Feedback by Component of their High Stakes Appraisal

Discussion

Amongst TALIS participating systems, the use of student test scores in teacher appraisals is nearly universal. The ubiquitous application of student test scores as the most common component in teacher appraisals, regardless of the stakes attached, is another example of the importance placed on these seemingly objective measures of education quality and part of the larger Global Testing Culture. The use of student test scores is significantly associated with the stakes of the appraisal with 75% of teachers working in schools that employ test-based high stakes teacher appraisal systems.

Test scores, however, are rarely included in isolation. Instead high stakes teacher appraisal patterns include multiple components, with over 60% of teacher appraisals in the total sample including teacher observation, student surveys, assessment of teacher's content knowledge, teacher's self-assessment, and parent feedback, in addition to student test scores. Notwithstanding the use of multi-metric patterns in teacher appraisals, student performance is still the most emphasized piece of feedback when appraisal results are communicated with teachers. This suggests that, although a variety of inputs may be used to make high stakes decisions, a greater weight or importance is put on the role of student test scores.

The disproportionate emphasis on student test scores is associated with lower levels of perceived feedback utility. In appraisal systems focused on tests, teachers believe the appraisal has limited impact on their teaching and is strictly an administrative exercise. This undermines the potentially formative aspects of appraisals. Furthermore, the sense that appraisals are simply an administrative checklist matches some of the concerns historically associated with teacher evaluations; demonstrating the ongoing melding of teacher appraisals and teacher evaluations.

In addition to perceptions that appraisal feedback is of little value, feedback was absent for a large number of teachers in test-based high stakes appraisal systems. Teachers whose pay, career trajectory, or continuation of employment depends on the outcome of their appraisal

should receive information on the components from which they are judged. Sadly, this is not always the case. At the high end, 42% of teachers whose appraisal is based in part on parent feedback receive no information on what the parents have said. Even information on teacher observation and student test scores, both included in over 95% of high stakes appraisals, is not always communicated.

The lack of feedback and perception that it is of little use beyond meeting administrative requirements can impact teacher's motivation. Teachers perceptions of appraisal utility is important, as individuals that do not see the value in feedback are less motivated and less likely to take action (Delvaux et al, 2013). Based on a three point satisfaction scale measuring whether teachers enjoyed working at their school, would recommend their school to others, and would not change their school if they could, teachers that felt the appraisal had little impact ($t = 32.45$, $p<.01$) or was solely administrative ($t = 55.70$, $p<.01$) were less satisfied with their work.

These cross-national findings tend to support the isomorphic march of test-based high stakes accountability laid out in the Global Testing Culture. Supporting the massive testing emphasis in England (UK) and the US that has been reported elsewhere (Hursh, 2007; Lingard & Lewis, 2016; Smith, 2014), results indicate that England (UK) may be the most test-obsessed system in the TALIS sample, with the US following narrowly behind. In England (UK) 98.5% of teachers work in a test-based high stakes system, the most of any participating system. Additionally, the emphasis placed on student performance in teacher feedback (3.81) is the second highest across all systems and the relative difference between student performance and other potential areas of feedback is over one, by far the largest relative importance across systems and a massive difference given the four-point scale. The US has the only other relative importance score over 0.70, as the emphasis placed on student performance is 0.83 points greater than the mean score of other potential areas of feedback.

Although the large majority of countries have test-based high stakes teacher appraisal systems in which student performance is the number one point of emphasis, a few outlier countries can be identified. France, at first glance, appears to be a typical country with nearly 73% of teachers working under test-based high stakes appraisals. However, upon exploring patterns of appraisals it is clear that schools in France have substantial autonomy – as 15 patterns are present and the top three represent less than 55% of all teachers – and that test scores are less emphasized in teacher feedback – as less importance is placed on student performance (ranking 6 out of 11 possible factors in feedback with the lowest relative importance score for student performance among the sampled countries). Finland also appears unique with the lowest inclusion rate of student test scores in teacher appraisals (75.3%) and greater emphasis placed on teacher collaboration over student performance. Finally, in Japan although over 97% of appraisals include student test scores, teacher appraisals are rarely associated with high stakes (27.7%). Furthermore, of the approximately one quarter of teachers in Japan that work under test-based high stakes appraisals, student performance is emphasized below six other pieces of feedback, with pedagogical competency highly valued.

This study was designed to examine the role of student test scores in high stakes teacher appraisals using the largest cross national dataset focused on teachers. Unfortunately, the nearly universal acceptance of incorporating student test scores into teacher appraisals and the dominance of one appraisal pattern over others limited the statistical power to identify significant differences between test-based high stakes appraisal systems and non-test based or lower stakes systems. This suggests that individual country studies, where great variance in appraisal patterns are present and appropriately large samples of teachers are included, may be the best way to evaluate the impact of such appraisals. Future research should extend this research by exploring the impact of test-based high stakes teacher appraisals on important outcomes such as teacher satisfaction and retention. Analyses of different policy initiatives linking teacher appraisals and pay around the world offer some recommendations for best practices. Literature suggests that teacher appraisal should be thought of as a tool for

professional development and not just an accountability measure. In addition, teacher appraisal needs to be based on good governance, which uses coherent frameworks negotiated together with teacher unions, policy makers and school management. Such systems should have clear and transparent procedures, in order to ensure trust in the system. Systems based on these principles should also use the results of the appraisal to feed into professional development and adequately address exceptional performance as well as underperformance (OECD, 2013a, 2013b; UNESCO, 2014).

Acknowledgements

The analysis, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of UNESCO's Global Education Monitoring Report.

References

AERA. (2015). AERA statement on use of value-added models (VAM) for the Evaluation of Educators and Educator Preparation Programs. *Educational Researcher*, 1-5.

Ahsan, S. & Smith, W. C. (2016). Facilitating student learning: A comparison of classroom and accountability assessment. In W. C. Smith (Ed.), *The Global Testing Culture: Shaping Education Policy, Perceptions, and Practice* (pp. 131-152). Oxford: Symposium Books.

Allen, R., Elks, P, Outhred, R., & Varly, P. (2016). Uganda's assessment system: A road-map for enhancing assessment in education. Oxford: Health & Education Advice & Resource Team (HEART).

Andreasen, K., Kelly, P., Kousholt, K., McNess, E., & Ydesen, C. (2015). Standardised testing in compulsory schooling in England and Denmark: A comparative study and analysis. *Bildung und Erziehung, 68*(3), 329-348. https://doi.org/10.7788/bue-2015-0306

Balwanz, D. (2016). The discursive hold of the matric: Is there space for a new vision for secondary education in South Africa? In W. C. Smith (Ed.), *Th Global Testing Culture: Shaping Education Policy, Perceptions, and Practice* (pp. 261-278). Oxford: Symposium Books.

Bantwini, B. D., & King-McKenzie, E. (2011). Some issues that teachers are confronted with: A case of the United States of America and South Africa. *Journal of Emerging Knowledge on Emerging Markets, 3*, 361-374. https://doi.org/10.7885/1946-651X.1053

Barnes, S.-A., Lyonette, C., Atfield, G,. & Owne, D. (2016). *Teacher's Pay and Equality: A Literature Reivew. Longitudinal Research into the Impact of Changes to Teachers' Pay on Equality in Schools in England.* Warwickshire, UK: Warwick Institute for Employment Research.

Booher-Jennings, J. (2005). Below the bubble: 'Educational triage' and the Texas accountability system. *American Education Research Journal, 42*(2), 231-268. https://doi.org/10.3102/00028312042002231

Bottery, M. (1989). The education of business management. *Oxford Review of Education, 15*(2), 129-146. https://doi.org/10.1080/0305498890150203

Certo, J. (2006). Beginning teacher concerns in an accountability-based testing environment. *Journal of Research in Childhood Education, 20*(4), 331-349. https://doi.org/10.1080/02568540609594571

Chapman, D., & Snyder, C. (2000). Can high stakes national testing improve instruction: Re -examining conventional wisdom. *International Journal of Educational Development, 20*, 457-474. https://doi.org/10.1016/S0738-0593(00)00020-1

Danielson, C. (2011). Evaluations that help teachers learn. *Educational Leadership, 68*(4), 35-39.

Danielson, C., & McGreal, T.L. (2000). *Teacher Evaluation: To Enhance Professional Practice.* Alexandria, VA: ETS.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Achives, 8*(1). http://dx.doi.org/10.14507/epaa.v8n1.2000

Darling-Hammond, L. (2015). Can value added add value to teacher evaluation? *Educational Researcher, 44*(2), 132-137. https://doi.org/10.3102/0013189X15575346

Delvaux, E., Vanhoof, J. Tuytens, M., Vekeman, E., Devos, G., & Van Petegem, P. (2013). How may teacher evaluation have an impact on professional development? A multilevel analysis. *Teaching and Teacher Education, 36*, 1-11.

Duke, D. L. (1995). *Teacher Evaluation Policy: From Accountability to Professional Development*. Albany, NY: SUNY Press.

Duke, D. L., & Stiggins, R. L. (1986). *Teacher Evaluation: Five Keys to Growth*. Washington, DC: National Education Association.

Evans, H. (2013). The impact of performance management policy on standards in schools. PhD Dissertation. Department of Sociology: University of Sussex.

Figazzolo, L. (2013). *The Use and Misuse of Teacher Appraisal. An Overview of Cases in the Developed World*. Brussels: Education International.

Fitzgerald, T. (2008). The continuing politics of mistrust: Performance management and the erosion of professional work. *Journal of Educational Administration and History, 40*(2), 113-128. https://doi.org/10.1080/00220620802210871

Fullan, M., & Mascall, B. (2000). *Human Resource Issues in Education: A Literature Review*. Wellington, New Zealand: Ministry of Education.

Garrett, R., & Steinberg, M. P. (2015). Examining teacher effectiveness using classroom observation scores: Evidence from the randomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*(2), 224 –242. https://doi.org/10.3102/0162373714537551

Goe, L., & Croft, A. (2009). *Methods of Evaluating Teacher Effectiveness*. Washington, DC: National Comprehensive Center for Teacher Quality.

Goldstein, R. A. (2011). Imaging the frame: Media representations of teachers, their unions, NCLB, and education reform. *Educational Policy, 25*(4), 543-576. https://doi.org/10.1177/0895904810361720

Gur, B. S., Celik, Z., & Ozoglu, M. (2012). Policy options for Turkey: A critique of the interpretation and utilization of PISA results in Turkey. *Journal of Education Policy, 27*(1), 1-21. https://doi.org/10.1080/02680939.2011.595509

Henry, G. T., & Guthrie, J. E. (2016). Using multiple measures for developmental teacher evaluation. In J. A. Grissom & P. Youngs (Eds.), *Improving Teacher Evaluation Systems: Making the Most of Multiple Measures* (pp. 1-7). New York: Teachers College Press.

Hursh, D. (2005). Neo-liberalism, markets and accountability: Transforming education and undermining democracy in the United States and England. *Policy Futures in Education*, *3*(1), 3-15. https://doi.org/10.2304/pfie.2005.3.1.6

Hursh, D. (2007). Assessing No Child Left Behind and the rise of neoliberal education policies. *American Education Research Journal, 44*(3), 493-518. https://doi.org/10.3102/0002831207306764

Jackson, C. K., Rockoff, J. E., & Staiger, D. O. (2014). Teacher effects and teacher-related policies. *Annual Review of Economics, 6*(1), 801-825. https://doi.org/10.1146/annurev-economics-080213-040845

Jennings, J., & Sohn, H. (2016). Measure for measure: How proficiency-based accountability systems affect inequality in academic achievement. *Sociology of Education, 87*(2), 125-141. https://doi.org/10.1177/0038040714525787

Jimerson, S. R., & Haddock, A.D. (2015). Understanding the importance of teachers in facilitating student success: Contemporary science, practice, and policy. *School Psychology Quarterly, 30*(4), 488-493. https://doi.org/10.1037/spq0000134

Kijima, R., & Leer, J. (2016). Legitimacy, state-building, and contestation in education policy development: Chile's involvement in cross-national assessments. In W.C. Smith (Ed.), *The Global Testing Culture: Shaping Education Policy,Perceptions, and Practice* (pp. 43-63). Oxford: Symposium Books.

Kim, J. S., & Sunderman, G. L. (2005). Measuring academic proficiency under the No Child Left Behind Act: Implications for educational equity. *Educational Researcher, 34*, 3-13. https://doi.org/10.3102/0013189X034008003

Kumashiro, K. K. (2012). *Bad Teacher!: How Blaming Teachers Distorts the Bigger Picture.* New York: Teachers College Press.

Larsen, M. A. (2005). A critical analysis of teacher evaluation policy trends. *Australian Journal of Education, 49*(3), 292-305. https://doi.org/10.1177/000494410504900306

Lingard, B., & Lewis, S. (2016). Globalization of the Anglo-American approach to top-down, test based acountability. In G. T. L. Brown & L. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 387-403). New York: Routledge.

Madden, J., & Lynch, D. (2014). Enabling teachers to better teach through engaging with research. *International Journal for Cross-Disciplinary Studies in Education, 5*(4), 1790-1797.

Malik, M. S., & Aslam, S. (2013). Performance appraisal and employee's motivation: A comparative analysis of telecom industry in Pakistan. *Pakistan Journal of Social Sciences, 33*(1), 179-189.

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving: An international perspective. *Studies in Educational Evaluation, 49*, 15-29. https://doi.org/10.1016/j.stueduc.2016.03.002

Marzano, R. J., & Toth, M.D. (2013). *Teacher Evaluation That Makes a Difference: A New Model for Teacher Growth and Student Achievement.* Alexandria, VA: ASCD.

McLaughlin, M., & Pfeiffer, S. (1988). *Teacher Evaluation. Improvement, Accountability and Effective Learning.* New York: Teachers College Press.

Meyer, H. D., & Benavot, A. (2013). *PISA, Power, and Policy: The Emergence of Global Educational Governance.* Oxford: Symposium Books. https://doi.org/10.15730/books.85

Mok, M. M. C., & Zhu, J. (2014). Cluster analysis of attitudes towards feedback and their mathematics achievement: A study of Hong Kong primary students. *Asia Paciic Journal of Educational Development, 3*(2), 1-13.

Murgatroyd, S., & Sahlberg, P. (2016). The two solitudes of educational policy and the challenge of development. *Journal of Learning for Development, 3*(3), 9-21.

OECD. (2013a). *Synergies for Better Learning. An International Perspective on Evaluation and Assessment.* Paris: Organisation for Economic Cooperation and Development.

OECD. (2013b). *Teachers for the 21st Century: Using Evaluation to Improve Teaching.* Paris: Organisation for Economic Cooperation and Development.

OECD. (2014). *TALIS 2013 Results: An International Perspective on Teaching and Learning.* Paris: Organisation for Economic Cooperation and Development.

O'Hare, K., & Bo, X. (2013). A case study of blended learing: The "Communicative Assessment - Development of Testing Skills" project. In B. Tomlinson & C. Whittaker (Eds.), *Blended Learning in English Teaching: Course Design and Implementation,* (pp. 83-89). London: British Council.

Pons, X. (2017). Fifteen years of research on PISA effects on education governance. *European Journal of Education.* https://doi.org/10.1111/ejed.12213

Popham, W. J. (2013). On serving two masters: Formative and summative teacher evaluation. *Principal Leadership*, 18-22.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods.* Thousand Oaks, CA: Sage Publications.

Rhoton, J., & Stiles, K. E. (2002). Exploring the professional development design process: Bringing an abstract framework into practice. *Science Educator*, *11*(1), 1-8. https://doi.org/10.1023/A:1013048828150

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*(2), 417-458. https://doi.org/10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252. https://doi.org/10.1257/0002828041302244

Schmelkes, S. (2015). Assessment of teacher performance - state of affairs. In UNESCO (Ed.), *Critical Issues for Formulating New Teacher Policies in Latin America and the Caribbean: The Current Debate*. Paris: UNESCO.

Shinkfield, A. J., & Stufflebeam, D. L. (1995). *Teacher Evaluation: Guide to Effective Practice*. Norwell, MA: Kluwer Academic Publishers.

Smith, W. C. (2017a). National testing policies and educator based testing for accountability: The role of selection in student achievment. *OECD Journal: Economic Studies, 2017*.

Smith, W. C. (2017b, forthcoming). Quality and inclusion in the SDGs: Tension in principle and practice. In C. Ydesen, A. Morin and B. Hamre (Eds.), *Testing and Inclusive Schooling*. Oxford: Routledge.

Smith, W. C. (2016a). *The Global Testing Culture: Shaping Education Policy, Perceptions, and Practice*. Oxford: Symposium Books. https://doi.org/10.15730/books.94

Smith, W. C. (2016b). An introduction to the Global Testing Culture. In W.C. Smith (Ed.), *The Global Testing Culture: Shaping Education Policy,Perceptions, and Practice* (pp. 7-24). Oxford: Symposium Books.

Smith, W. C. (2014). The global transformation toward testing for accountability. *Education Policy Analysis Archives, 22*(116). http://dx.doi.org/10.14507/epaa.v22.1571

Somerset, A. (2016). Questioning across the spectrum: Pedagogy, selection examinations, and assessment systems in low-income countries. In W.C. Smith (Ed.), *The Global Testing Culture: Shaping Education Policy,Perceptions, and Practice* (pp. 171-192). Oxford: Symposium Books.

Styles Johnston, G., & Camp Yeakey, C. (1979). The supervision of teacher evaluation: A brief overview. *Journal of Teacher Education, 30*(2), 17-22. https://doi.org/10.1177/002248717903000206

Ticha, R., & Abery, B. (2016). Beyond the large-scale testing of basic skills: Using formative assessment to facilitate learning. In W.C. Smith (Ed.), *The Global Testing Culture: Shaping Education Policy,Perceptions, and Practice* (pp. 153-170). Oxford: Symposium Books.

Toch, T., & Rothman, R. (2008). Rush to judgement: Teacher evaluation in public education. Education sector reports. Washington, D.C.: Education Sector Think Tank.

UIS. (2016). *The World Needs Almost 69 Million New Teachers to Reach the 2030 Education Goals*. UIS Fact Sheet, October 2016, No. 39. Montreal, Canada: UNESCO Institute for Statistics.

UNESCO. (2014). *EFA Global Monitoring Report 2013/4 - Teaching and Learning: Achieving Quality for All*. Paris: UNESCO.

UNESCO. (2015). *EFA Global Monitoring Report: Education for All 2000-2015: Achievements and Challenges*. Paris: UNESCO.

USDOE. (2009). *Race to the top program*. Washington, DC: U.S. Department of Education.

Volante, L. (2007). Evaluating test-based accountability perspectives: An international perspective. Paper presented at the Association for Educational Assessment – Europe, Stockholm, Sweden.

Von der Embse, N. P., Pendergast, L. L., Segool, N., Saeki, E., & Ryan, S. (2016). The

influence of test-based accountability policies on school climate and teacher stress across four states. *Teaching and Teacher Education, 59*, 492-502. https://doi.org/10.1016/j.tate.2016.07.013

Whitford, M. (2013). Performance appraisal in primary schools: Managing the integration of accountability and development. Masters Thesis. Educational Management and Leadership: Unitec Institute of Technology.

Yan, S. (2012). Teachers' role in autonomous learning. *Journal of Sociological Research, 3*(2), 557-562.

Appendix A

Top 10 Teacher Appraisal Patterns in Pooled Sample

Ranking	Abbreviation for Pattern	Components Included	%
1	T.O.S.CK.SA.P	Test, Observation, Students, Content Knowledge, Self Assessment, Parents	63.25
2	T.O.S.SA.P	Test, Observation, Students, Self Assessment, Parents	7.89
3	T.O.CK.SA.P	Test, Observation, Content Knowledge, Self Assessment, Parents	5.81
4	T.O.S.CK.P	Test, Observation, Students, Content Knowledge, Parents	5.17
5	T.O.S.P	Test, Observation, Students, Parents	2.92
6	T.O.SA.P	Test, Observation, Self Assessment, Parents	2.02
7	T.O.CK.P	Test, Observation, Content Knowledge, Parents	1.88
8	T.O.S.CK.SA	Test, Observation, Students, Content Knowledge, Self Assessment	1.76
9	T.O.S.CK	Test, Observation, Students, Content Knowledge	1.26
10	T.O.P	Test, Observation, Parents	1.11

Appendix B

Top Three Teacher Appraisal Patterns by Country

Country	Sample of Teachers	% Test-Based HS Appraisals	First	First (%)	Second	Second (%)	Third	Third (%)
Australia	1,690	85.56	T.O.S.CK.SA.P	56.78	T.O.CK.SA.P	11.96	T.O.S.SA.P	6.92
Brazil	11,034	72.23	T.O.S.CK.SA.P	65.00	T.O.S.SA.P	6.51	T.O.S.P	4.38
Bulgaria	2,616	96.29	T.O.S.CK.SA.P	60.74	T.O.S.CK.P	10.40	T.O.S.SA.P	6.51
Chile	1,128	78.99	T.O.S.CK.SA.P	54.55	T.O.CK.SA.P	19.98	T.O.SA.P	4.43
Cyprus	1,699	67.10	T.O.S.CK.SA.P	50.61	T.O.CK.SA.P	14.47	T.O.S.CK	7.02
Czech Republic	3,110	95.76	T.O.S.CK.SA.P	65.11	T.O.S.SA.P	19.54	T.O.S.CK.P/ T.O.S.P	2.62
Denmark	1,239	78.21	T.O.S.CK.SA.P	51.39	T.O.S.P	8.05	T.O.S.SA.P	7.74
Estonia	2,905	89.02	T.O.S.CK.SA.P	85.00	T.O.S.SA.P	7.23	T.O.CK.SA.P	1.78
Finland	1,922	61.50	T.O.S.CK.SA.P	40.27	T.O.S.P	17.85	T.O.S.SA.P	8.54
France	2,373	72.69	T.O.CK.P	19.48	T.O.CK.SA.P	19.02	T.O.S.CK.SA.P	14.20
Iceland	871	78.87	T.O.S.CK.SA.P	22.13	T.O.S.SA.P	19.36	***T.S.SA.P***	***10.19***
Israel	2,823	89.05	T.O.S.CK.SA.P	64.40	T.O.S.CK.SA	9.31	T.O.S.SA.P	5.29
Italy	***862***	***36.43***	T.O.S.CK.SA.P	33.80	***T.P.***	***15.30***	T.O.CK.SA.P	13.70
Japan	***3,357***	***26.48***	T.O.S.CK.SA.P	62.00	T.O.S.SA.P	21.70	T.O.CK.SA.P	8.89
South Korea	2,442	80.51	T.O.S.CK.SA.P	70.14	T.O.S.SA.P	8.04	T.O.S.CK.P/T.O.S.P	2.39
Latvia	1,892	91.07	T.O.S.CK.SA.P	77.02	T.O.S.SA.P	21.14	T.O.S.P	1.74
Malaysia	2,873	56.46	T.O.S.CK.SA.P	76.70	T.O.CK.SA.P	14.55	T.O.S.SA.P	2.47
Mexico	***2,914***	***47.53***	T.O.S.CK.SA.P	84.69	T.O.S.CK.P	4.55	T.O.S.P	3.47
Netherlands	1,671	90.13	T.O.S.CK.SA.P	64.48	T.O.S.CK.SA	11.42	T.O.S.SA.P	5.91
Norway	1,968	63.82	T.O.S.CK.SA.P	57.48	T.O.CK.SA.P	10.27	T.O.S.SA.P	7.09
Poland	3,448	92.31	T.O.S.CK.SA.P	78.26	T.O.S.SA.P	8.20	T.O.S.CK.P	4.93
Portugal	***3,290***	***40.24***	T.O.S.CK.SA.P	27.87	T.O.S.SA.P	23.04	T.O.CK.SA.P	15.18
Serbia	3,544	57.93	T.O.S.CK.SA.P	82.42	T.O.S.CK.P	8.09	T.O.S.SA.P	5.07
Singapore	2,709	97.82	T.O.S.CK.SA.P	72.30	T.O.CK.SA.P	18.00	***T.O.CK.SA***	***3.02***

Appendix B (Cont'd.)

Top Three Teacher Appraisal Patterns by Country

Country	Sample of Teachers	% Test-Based HS Appraisals	First	First (%)	Second	Second (%)	Third	Third (%)
Slovak Republic	3,051	96.20	T.O.S.CK.SA.P	68.69	T.O.S.SA.P	14.41	T.O.S.CK.P	5.01
Spain	***2,075***	***35.81***	T.O.S.CK.SA.P	29.60	***T.S.SA.P***	***14.90***	T.O.S.SA.P	13.30
Sweden	2,834	95.66	T.O.S.CK.SA.P	41.76	T.O.S.SA.P	17.56	T.O.S.CK.P	14.94
United States	1,598	90.93	T.O.S.CK.SA.P	44.25	T.O.CK.SA.P	16.04	T.O.S.CK.P	10.46
England (UK)	2,339	98.50	T.O.S.CK.SA.P	62.46	T.O.S.CK.P	7.12	T.O.CK.SA.P	6.03
Flanders (Belgium)	2,685	84.25	T.O.S.CK.SA.P	42.84	T.O.S.CK.P	14.01	T.O.CK.P	9.06
Abu Dhabi (UAE)	1,842	70.30	T.O.S.CK.SA.P	90.04	T.O.S.CK.P	4.25	T.O.CK.P	3.01
Alberta (Canada)	1,417	84.47	T.O.S.CK.SA.P	52.05	T.O.CK.SA.P	16.79	T.O.S.CK.P/T.O.S. SA.P	7.44
Romania	3,179	94.68	T.O.S.CK.SA.P	90.86	T.O.CK.SA.P	5.58	T.O.S.CK.P	1.73

About the Authors

William C. Smith
UNESCO – Global Education Monitoring Report
w.smith@unesco.org
William C. Smith, Ph.D., is a Senior Policy Analyst with UNESCO's Global Education Monitoring (GEM) Report and a Research Affiliate at Penn State University's Population Research Institute. His larger research is situated around the role social policy at the national and international level plays in education equity and outcomes. This article continues his current work on the power of student test scores to shape policy, influence student outcomes, and warp the education process. His recent publications in this line of work include the edited book, *The Global Testing Culture: Shaping Education Policy, Perceptions, and Practice*, and *National testing policies and educator based testing for accountability: The role of selection in student achievement* in the OECD Journal: Economic Studies.

Katarzyna Kubacka
UNESCO – Global Education Monitoring Report
k.kubacka@unesco.org
Katarzyna Kubacka, Ph.D., is a Research Officer at the Global Education Monitoring (GEM) Report. Before joining the GEM Report, Katarzyna spent five years at the Organisation for Economic Co-operation and Development (OECD), working at the Centre for Educational Research and Innovation (CERI) and Early Childhood and Schools division. She was a Policy Analyst on the Education and Social Progress project and the Teaching and Learning International Survey (TALIS). Her work concentrates on the relationships between teaching and learning, skills development and well-being.

About the Guest Editors

Jessica Holloway
Deakin University
Jessica.holloway@deakin.edu.au
Jessica Holloway, Ph.D., is a post-doctoral research fellow at the Centre in Research for Educational Impact (REDI) at Deakin University. She draws on post-structural theory to understand contemporary modes of accountability and its production of new teacher and leader subjectivities. Her current project, entitled *Teacher Leaders and Democracy: An International Study,* looks at modes of distributive leadership in U.S. and Australian schools.

Tore Bernt Sørensen
University of Bristol
t.b.sorensen@bristol.ac.uk
Tore Bernt Sørensen completed his doctorate at the Graduate School of Education, University of Bristol, UK, in 2017 with the dissertation "*Work In Progress: The Political Construction Of The OECD Programme Teaching And Learning International Survey*". Tore's research centres on comparative studies of education governance in a global context. Tore has a background as teacher and teacher trainer in Denmark. Before starting his doctorate, he worked in the Analysis and Studies Unit of the European Commission's Directorate-General for Education and Culture.

Antoni Verger
Universitat Autònoma de Barcelona
Antoni.verger@uab.cat
Antoni Verger is associate professor at the Department of Sociology of the UAB. A former post-doctoral fellow at the University of Amsterdam, Antoni's research analyses the relationship between global governance institutions and education policy, with a focus on the study of public-private partnerships and accountability policies in education. Currently, he is coordinating the research project REFORMED - *Reforming Schools Globally: A Multiscalar Analysis of Autonomy and Accountability Policies in the Education Sector* (ERC StG, 2016–2021).

SPECIAL ISSUE
Global Perspectives on High-Stakes Teacher Accountability Policies

education policy analysis archives

Volume 25 Number 86 August 21, 2017 ISSN 1068-2341