The Reliability of Classroom Observations and Student Surveys in Non-Research Settings: Evidence from a Middle-Income Country[1]

Alejandro J. Ganimian[2]
Harvard University/
New York University

Andrew D. Ho[3]
Harvard University

Alejandra Campos Quintero[4]
Columbia University

Abstract: We present one of the first Generalizability studies of non-test measures of teaching effectiveness administered by practitioners in a middle-income country. The reliability of observations varies widely (from 0 to 0.75 on a 0-1 scale) and depends upon their context (whether they are conducted during training or on the job) and rater assignment configurations. The reliability of surveys varies substantially, coinciding with a change in which students were sampled across occasions. Our estimates are comparable to the reliability from research in high-income countries, but the variation within our estimates and between them and those from individual studies suggests that practitioners should conduct their own reliability analyses. We offer guidance on leveraging such analyses to improve the reliability of their measures.

[1] We gratefully acknowledge the funds provided by the Inter-American Development Bank for this study. We thank Emiliana Vegas, Mariana Alfonso, and the *Enseñá por Argentina* team—especially, Oscar Ghillione, Fernando Viola, Mariana Albarracín, and Laura de Jorge—for making this study possible. We also thank Samuel Hansen Freel for excellent research assistance.

[2] Visiting Associate Professor of Education, Harvard Graduate School of Education. Associate Professor of Applied Psychology and Economics, Steinhardt School of Culture, Education, and Human Development, New York University. alejandro_ganimian@gse.harvard.edu.
[3] Charles William Eliot Professor of Education, Harvard Graduate School of Education. andrew_ho@gse.harvard.edu.
[4] Doctoral Student, Teachers College at Columbia University. aac2271@tc.columbia.edu.

1. Introduction

There is a growing consensus on the need to measure teaching effectiveness using multiple instruments (Baker et al., 2010; Goe et al., 2008; Kane et al., 2014). Yet, with notable exceptions (Glazerman et al., 2011; Whitehurst et al., 2014), guidance on how to achieve reliable ratings from non-test measures derives largely from studies in which classroom observations and student surveys were administered for research purposes in high-income countries. The applicability of such guidance beyond such contexts remains underexamined. When these measures are collected by researchers, there are multiple mechanisms in place (e.g., rater training, certification, master coding, validation engines) to ensure the integrity of the information being gathered. Further, nearly all studies in this literature were conducted in the United States, where teachers may be more used to being evaluated than those in other countries (Commission/EACEA/Eurydice, 2021; Isoré, 2009; Martinez et al., 2016; OECD, 2015, 2025; Pouezevara et al., 2016). Both factors could render these metrics more reliable than those administered by practitioners in low- and middle-income countries (LMICs). If this were the case, practitioners making decisions based on research settings (Manzi et al., 2011; Quezada & Salcedo, 2019) could design their teacher feedback systems that produce less reliable results than they intended and realize.

In this paper, we leverage classroom observations and student surveys on 100 teachers collected by an alternative pathway into teaching (called *Enseñá por Argentina* or ExA) in Argentina to understand how their reliability compares to those from research settings in the U.S. Our study differs from prior analyses of the reliability of non-test measures in four main ways. First, our data were collected by practitioners in a middle-income country, complementing previous efforts led by researchers in high-income countries. Second, the instruments that we examine were adapted from widely used protocols in the U.S. for formative purposes (there were

no stakes attached to them), reducing concerns that any differences in reliability with prior studies may be driven by how the instruments were developed or how their results were used. Third, we observe the same measures across two years and contexts (the measures were collected during practice teaching and once they began teaching in hard-to-staff schools) and multiple rater conditions (coaches and peers, alone or combined, held constant or rotating), allowing us to determine whether the reliability of our measures are stable or vary across administrations. Lastly, because the non-profit that provided the data is part of a global network of 60 organizations using similar procedures and measures (Teach for All), we see our results as potentially relevant beyond the specific context of our study.

Our analysis goes beyond traditional metrics of reliability that quantify consistency in scores across one source of error at a time (e.g., items or raters). We simultaneously estimate the contribution of different facets of measurement error (e.g., item difficulty and rater stringency) and of interactions between these facets (e.g., some raters being more stringent on some items). The main advantage of this approach is that, by being more precise about the sources of error, we can also be more strategic about reducing it (e.g., if rater stringency is contributing more to measurement error than item difficulty, we can reduce error more efficiently by increasing the number of raters instead of items). This approach, "G(eneralizability) theory" (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991), is increasingly used in teacher feedback systems in the U.S. (Bell et al., 2012; Hill et al., 2012a; Ho & Kane, 2013; Kane & Staiger, 2012; Meyer et al., 2011). To our knowledge, it has not been widely applied in low- or middle-income countries.

We report five main findings. First, classroom observations conducted by practitioners vary widely in their reliability for making both *relative* distinctions (deciding which teachers are

more effective) and *absolute* judgments about teachers (yielding consistent scores for similar performance levels) with current numbers of items, raters, and occasions. Across two settings in which teachers were observed—clinical practice and the school year—the generalizability coefficient for relative error—a measure of reliability for relative distinctions from 0 (perfectly unreliable) to 1 (perfectly reliable)—ranged from 0 to 0.75, and the coefficient for absolute error—a metric for absolute judgments also from 0 to 1—ranged from 0 to 0.71. The vast degree of variability indicates that observations in these settings can range from pure noise (i.e., 0% of variation in observation scores reflecting actual differences in measured teaching effectiveness, as opposed to measurement error) to helpful mechanisms to assign teachers in need of support to interventions (i.e., 75% of variation reflecting actual differences). The single-observation generalizability coefficients for relative error range from 0 to 0.61 across conditions, with a mean of 0.39, which is remarkably close to the mean of 0.39 and median of 0.38 across the 44 estimates from Generalizability studies conducted by researchers in high-income countries.

Second, the level and variation in the reliability of these observations seems to be partly explained by the context in which they were administered and the ways in which raters were assigned to teachers. The generalizability coefficients were lower during clinical practice than during the school year: the mean coefficient for relative (absolute) error in clinical practice was 0.48 (0.42) and the one for relative (absolute) error in the school year was 0.64 (0.52). These coefficients also varied more during clinical practice: the coefficient for relative (absolute) error varied from 0 to 0.71 (0 to 0.75), compared to 0.63 to 0.66 (0.44 to 0.59) during the school year. Within clinical practice, reliability varied widely across different types of rater assignment: both coefficients were highest and least variable when the same peer scored both lessons, lower and slightly more variable when the same coach scored both lessons, and lowest and most variable

when a different peer scored each lesson. These figures suggest that the type of considerations that influence rater assignment among practitioners may have non-trivial impacts on reliability.

Third, with one exception, it seems possible to improve the reliability of observations by increasing the number of times teachers are scored. During clinical practice, observing each teacher thrice instead of twice would improve the generalizability coefficients for relative error by 6-9 percentage points (pp.) and the ones for absolute error by 4-7 pp., depending on the rater type (coaches or peers) and whether all lessons per teacher are observed by the same or a different rater. Adding an observation during the school year would improve the coefficients for relative and absolute error by 8 pp. across both years for which we have data. Further increases in the number of observations would improve reliability by a smaller margin.

Fourth, student surveys administered by practitioners also vary in their reliability. During clinical practice, the generalizability coefficient for relative error ranged from 0.33 to 0.58, and the one for absolute error from 0.29 to 0.57. In the school year, the corresponding figures were 0.59 to 0.85 and from 0.36 to 0.70, respectively. These results indicate that somewhere between 33 and 85% of variation in observed scores reflects actual differences in measured teaching effectiveness. The single-rater generalizability coefficients for relative error range from 0.03 to 0.25, broadly consistent with the only estimate of 0.16 we found from prior research in the U.S.

Fifth, improving the reliability of student surveys is possible by increasing the number of respondents with reasonable extensions to existing administration conditions. During clinical practice, surveying 15 instead of 10 students would improve the generalizability coefficients for relative error by 7 to 10 pp. and the ones for absolute error by 6 to 10 pp. Adding five students during the school year would improve the coefficients for relative and absolute error by 6-9 pp.

and 4 pp., respectively. Further increases in the number of students surveyed would improve reliability by smaller margins.

The rest of the paper is structured as follows. Section 2 reviews prior research on the reliability of classroom observations and student surveys, showing that measures collected for research purposes exhibit relatively high levels of reliability. Section 3 describes the data used for this study, which draws on classroom observations and student surveys administered in two different settings across two years of an alternative pathway into teaching in Argentina. Section 4 explains how we use generalizability theory to be more precise about the sources of measurement error in observations and surveys and more strategic about how to reduce them. Section 5 presents our estimates of reliability for both metrics and how they may be improved by increasing the number of the relevant facets of error (e.g., increasing raters and/or lessons).

2. Prior research

In the past two decades, policymakers and practitioners became increasingly interested in measuring teaching effectiveness. Initially, this interest was largely motivated by research suggesting that teachers vary widely in their capacity to improve their students' achievement. Several studies found that the students of some teachers consistently score higher in standardized tests than those of others, even when both groups have similar demographics and start at comparable levels of achievement (Aaronson et al., 2007; Chetty et al., 2014; Kane et al., 2008; Koedel et al., 2015; Nye et al., 2004; Rivkin et al., 2005; Rockoff, 2004). This evidence prompted efforts to try to identify effective teachers to inform hiring, retention, training, and pay (Dee & Wyckoff, 2013; Goldhaber et al., 2017; Rockoff et al., 2011).

The statistical methods used to estimate teachers' influence on student achievement ("value-added models") have been criticized for sometimes leading to impossible results (e.g., a teacher affecting their students' prior-year test scores; Rothstein, 2010), yielding conflicting results across tests (Papay, 2011), ignoring other ways in which teachers contribute to students' well-being (Blazar, 2018; Jackson, 2020; Kraft, 2019), and neglecting how school-level factors (e.g., principals, counselors, and peers) mediate teachers' capacity to help students (Jackson, 2013; Jackson & Bruegmann, 2009; Johnson et al., 2012; Mulhern, 2023; Papay et al., 2020).

The shortcomings of value-added models, coupled with a recognition that teachers' impact on students' lives goes beyond improving their test scores, led many to advocate for combining student-achievement gains with other measures of teacher quality. Several studies illustrated the advantages of this approach, showing that other measures of teaching quality (e.g., classroom observations and student surveys) add valuable information not captured by tests (Baker et al., 2010; Darling-Hammond et al., 2012; Goe et al., 2008; Kane et al., 2014; Kane & Staiger, 2011, 2012; Kane et al., 2011). Some have offered practical advice on how to administer such instruments (e.g., the number of times a teacher should be observed to obtain consistent ratings of their performance; Ho & Kane, 2013).

The reliability of non-test measures administered by practitioners, however, has been relatively underexamined—especially, in LMICs. In this study, we attempt to address this gap by focusing on the reliability of classroom observations and student surveys of a non-profit organization that is part of a global network with similar teacher feedback practices. In this section, we review prior studies that have analyzed the reliability of these two instruments using a common approach (described below) to understand how they compare to those in our study. We focus on the reliability of each measure because that is how the organization that provided us

with the data for our study use the information. For research on the reliability of composites that combine multiple measures, see Douglas and Mislevy (2010); Martínez and Fernández (2021).

The study of the reliability of measures of teaching effectiveness in general, and of classroom observations and student surveys in particular, has evolved considerably in recent decades. Conventionally, educational measurement scholars conceive of the score a teacher receives in a procedure as partly due to that teacher's effectiveness and partly to errors in measurement. They distinguish between these parts by taking multiple measures and interpreting similarities across measurements as indicative of the former and differences as indicative of the latter. This idea is crystallized in "classical test theory" (CTT) and its equation $X_i = \tau + \varepsilon_i$, which indicates that any observed score X_i is equal to a true score τ plus the error from that procedure ε_i (Allen & Yen, 2001; Lord & Novick, 2008; Nunnally, 1978). The true score is the long-run average of scores over replications, measurement errors are replication-specific deviations from that average, and reliability is the correlation between scores across replications (the ratio of true to total score variance).

This framework is often used to quantify measurement error from the questions (items) in a test. If all items are measuring the same construct, we can interpret the expectation across item scores as the true score and any deviations from it as error. For example, Cronbach's alpha measures internal-consistency reliability as the proportion of total score variance due to shared variation across items (Cronbach, 1951). In classroom observations and student surveys, "items" refer to the indicators on which raters (in observations, coaches or peers; in surveys, students) are asked to score teachers (for examples of the items in the instruments used in this study, see sections 3.2.1 and 3.2.2). This idea is also applied to error from raters or occasions. If we see the

score from each rater (or occasion) as a replication, we can interpret the correlation in scores across raters (or occasions) as inter-rater (or test-retest) reliability.

A key limitation of classical analyses of reliability is that they do not distinguish between different sources of measurement error. They decompose observed-score variance into true and undifferentiated error variance. An alternative is to use random-effects models to parse out the contribution of each facet (e.g., items or raters) and interactions between them (e.g., raters being more stringent on some items). This approach, "G(eneralizability) theory" (Brennan, 2001; Cronbach et al., 1972; Shavelson & Webb, 1991), allows us to describe error variance more accurately and be more strategic about reducing it by increasing replications over the facets that add the most noise. For example, if rater stringency contributes more to error than item difficulty, increasing the number of raters will reduce error by a larger margin than increasing the number of items. In classroom observations and student surveys, each item is rated on a scale from 1 (lowest level of performance) to 5 (highest level). Item difficulty reflects systematic differences across items (e.g., items consistently rated lower than others are considered as more "difficult").

In recent decades, G(eneralizability) studies became an increasingly prominent method for examining the reliability of non-test measures of teaching effectiveness in K-12 education. These studies offered practical guidance on how to design teacher-feedback systems to produce reliable results. Perhaps most famously, the Measures of Effective Teaching (MET) study, which compared the reliability of four widely used classroom-observation protocols across five school districts in the U.S., concluded that "to achieve reliability in the neighborhood of 0.65... we had to score four different lessons, each with a different rater" (Kane & Staiger, 2012) and subsequently identified multiple approaches to reliable observations (MET Project, 2013). This

evidence has been cited in the design of teacher feedback systems in the U.S. and abroad (Manzi et al., 2011; Quezada & Salcedo, 2019).

Much of what we know about the reliability of alternative metrics of teaching effectiveness, however, stems from a relatively small set of context and measures. Table A.1 in Appendix A summarizes findings from 15 generalizability studies of classroom observations and student surveys. The studies were all conducted in high-income countries—primarily, the U.S. (87%) and only two studies in Europe, including one in Germany and Switzerland (Praetorius et al., 2014) and one in the Netherlands (van der Lans et al., 2016). The most commonly studied instrument is the Classroom Assessment Scoring System (CLASS, Hamre et al., 2013; Mashburn et al., 2008; Pianta et al., 2008a), which appears in seven studies across pre-primary, primary, and secondary levels; followed by the Framework for Teaching (FfT, Danielson, 2011) in five studies; and the Mathematics Quality of Instruction (MQI, Hill et al., 2012a; Hill et al., 2011) in three. All other instruments appear in only one study each.

Sample sizes vary widely, ranging from 8 teachers (Hill et al., 2012a) to 1,333 (Kane & Staiger, 2012), with a median of 48. Most studies, however, are small: 73% include fewer than 100 teachers. The number of observations per teacher typically ranges from 3 to 8, with a median of approximately 5. Extremes include Hill et al. (2012a) with only 1 observation per teacher and Ho and Kane (2013) with an average of 46—though the latter is exceptional.

Mean scores on classroom observations provide evidence of potential floor and ceiling effects. On the CLASS (scored from 1 to 7), Classroom Organization means are consistently high (5.18–5.75), approaching the upper bound and suggesting ceiling compression that may limit the instrument's ability to differentiate among teachers at the top of the distribution. Instructional Support means are substantially lower (1.93–3.58), with Praetorius et al. (2014)

Cognitive Activation domain at 1.93 suggesting a possible floor effect. Emotional Support means fall in the mid-range (3.64–5.37). On the FfT (scored from 1 to 4), means cluster near the center of the scale (1.87–2.58) with no strong evidence of floor or ceiling effects.

The standard deviation (SD) of the teacher effect and the standard error of measurement (SEM) of a single observation reveal that measurement error frequently rivals or exceeds true teacher variation. For CLASS domains, SDs range from 0.22 to 0.73 while SEMs range from 0.39 to 0.97. In many cases—particularly for Instructional Support—the SEM exceeds the SD, meaning that the error variance from a single observation is larger than the variance attributable to genuine differences between teachers. A similar pattern holds for the FfT, where SDs range from 0.19 to 0.37 and SEMs from 0.24 to 0.44. The RESET instrument shows especially large SEMs (0.82–2.11) relative to its 0–3 scale.

The reliability of a single observation (i.e., the proportion of observed-score variance attributable to persistent differences in teacher practice) is low across studies and instruments. Across the 45 reliability estimates in Table A.1, values range from 0.00 to 0.81 with a median of 0.38 and a mean of 0.39. Over three-quarters (76%) fall below 0.50, and only three estimates (7%) reach or exceed the conventional threshold of 0.70: two MQI subscales at the pre-primary level (Mantzicopoulos et al., 2018) and the FfT for reading instruction (Patrick et al., 2020). At the secondary and multiple-levels classifications, no single-observation reliability reaches 0.70. The sole student survey estimate (Schweig, 2014) yields a reliability of only 0.16. These findings indicate that a single observation by a single rater is generally insufficient for dependable measurement of teaching quality, regardless of the instrument used.

The classroom observations in these studies were conducted for research purposes and they incorporated several quality-assurance mechanisms that likely improve their reliability, such

as: rater training, assessment, certification, and additional practice (e.g., deep-dive training, one-on-one coaching, paired observations, and group calibration, Jerald, 2012); master coding (in which experts discuss and agree on correct scores and score rationales, McClellan, 2013), and a validation engine (including an online video library, scoring rubric, comparisons with other metrics, and automated reports, MET Project, 2010), among others (e.g., piloting the observation protocol, Joe et al., 2013). Whether observations conducted by practitioners, with fewer of these mechanisms, can achieve similar levels of reliability remains an open question.

3. Data

3.1 Context

We conducted our study in Argentina, an upper-middle income country with high levels of enrollment in primary and secondary school, but lower learning outcomes than its neighbors. Argentina's income per capita (USD 13,730) is comparable to that of China, Mexico, Russia, and Turkey (World Bank 2024a), but it has recently undergone several economic and political crises that distinguish it from both these countries and most of its South American neighbors. According to the latest data, 4 in 10 people live below the poverty line (World Bank 2024c). Over 99% of children and youth enroll in primary and lower-secondary school, but only 90% do so in upper-secondary school and just 70% graduate from high school (World Bank 2024b). Even among those who reach the last year of high school, 43% score at the lowest levels of the national assessment in language and 82% do so in math (Ganimian 2025). The share of 15-year-olds at the lowest levels of global tests is higher: 55% in language and 73% in math (OECD 2023). Additionally, the poorest students are 21 and 42 percentage points more likely to score at these levels in reading and math than their richest peers, respectively.

We focused on the Province of Buenos Aires (PBA), the largest sub-national school system in the country. In Argentina, the provinces (akin to U.S. states) are responsible for providing pre-primary to tertiary education and the federal government for providing higher education as well as technical and financial assistance to the provinces (Ley de Educación Nacional 2006). PBA serves 4.3 million students: 654,958 in pre-primary education, 1.7 million in primary education, 1.7 million in secondary education, and 260,082 at the tertiary level (MdCH 2024). PBA is representative of country as a whole, with a median household income of ARS 117,278 (USD 121) per month, which is almost identical to the national average. It is also comparable in income inequality, with a Gini coefficient slightly below the national mean (INDEC 2024). Its learning outcomes mirror this economic reality: its scores on the national assessment closely resemble those of the average province in the country (Ganimian 2025).

We obtained the data for our study from *Enseñá por Argentina* (ExA), a non-profit that recruits college graduates to teach in hard-to-staff schools for two years. By 2024, 15 years after its founding, ExA had placed 400 teachers serving 130,000 students across seven provinces (the Province and City of Buenos Aires, Chaco, Mendoza, Neuquén, Salta, and Santa Fé). Further, it follows similar processes to train and develop its teachers as 60 other organizations around the world that form the Teach for All network. We see our study as relevant for this broader group and for other organizations that use comparable instruments and procedures.

3.2 Procedure

In this study, we examine the reliability of two measures of teaching effectiveness (classroom observations and student surveys) developed and administered by ExA for feedback purposes. Like many members of the Teach for All network, ExA uses both classroom observations and student surveys to decide which teachers should receive additional support and the type of

support they need (e.g., classroom management v. content knowledge). These uses of data on teacher effectiveness are likely more common in no-research settings than decisions about hiring, pay, and dismissal, which are rarely based on performance (Bruns et al., 2011; Kraft & Gilmour, 2017; Weisberg et al., 2009). This is particularly true in Latin America, where principals have little discretion over these decisions (Adelman & Lemos, 2018; Anand et al., 2023). Just because measures in this context are used for formative purposes, it does not follow that they have no consequences. In fact, in resource-constrained non-profits, allocating support among teachers is perhaps the most consequential decision for both the organization and its teachers.

In 2014 and 2015, ExA administered these measures right after teachers were hired, during its summer training institute (a four-week pre-service training, which concludes with two weeks of practice teaching) and during the school year, once teachers were already in the classroom. We refer to the former process as "clinical practice" and to the latter one as the "school year." All new teachers participated in clinical practice only on the year in which they were hired (e.g., if a teacher was hired in 2014, they only participated in clinical practice in 2014) and both new and existing teachers taught during the school year for two years (e.g., the 2014 school-year dataset includes both teachers hired in 2013 [second-year] and 2014 [first-year]).

All participants in the summer training institute, including the coaches and teachers, participated in a one-hour training session on the classroom observations and student surveys organized by ExA's Training and Coaching team. The session introduced both instruments, their intended uses and frequency of administration during clinical practice and the school year, and the instruments on which it was based. Then, it described each of the domains and items in each domain, and it provided examples of the rating scale for selected items as an illustration.

Throughout the session, coaches and teachers had the opportunity to ask clarifying questions. After this session, they reviewed the instruments in groups, with members of ExA's staff circulating through the groups to answer any questions that may arise. Lastly, coaches and teachers had an opportunity to practice rating a demonstration lesson and discuss their results.

3.2.1 Clinical practice

During clinical practice, each teacher taught a group of volunteer students for two weeks, and they were observed on two lessons, with one rater scoring each lesson across six domains (a domain is the average score across multiple items; see Appendix B). In G-studies, this configuration of teachers, lessons, raters, and domains is denoted as a domain-by-lesson-within-teacher, or $d \times (l{:}\, t)$, design. In this design, domains are crossed with teachers and lessons (as indicated by the $\times$ sign) because all teachers were scored on the same classroom-observation protocol (see section 3.4.1) across all lessons. Lessons are nested within teachers (as indicated by the : sign) because each teacher taught different lessons (e.g., teacher A taught grade 5 math; teacher B taught grade 6 language).

The effect of rater stringency on reliability cannot be estimated because there was only one rater per lesson, so we cannot know how another rater would have scored the same lessons. Some teachers were scored by the same coach on both lessons, others by the same peer, and yet others by a different peer. Coaches (but not peers) observed multiple teachers, so we conduct separate studies for each coach (7 in 2014 and 8 in 2015) and report the average result across coaches for each year. This setup allows us to compare the reliability of these three approaches to assigning raters, which may be of interest to practitioners seeking to balance rater experience and availability. Teachers were not randomly assigned to these rater conditions, but practitioners

reported no systematic attention to pairing. Their assignment depended on logistical factors, such as the availability of coaches and peers during each lesson.

The students of each teacher were also surveyed on the last lesson of clinical practice on seven domains. In this case, students act as raters. In G-theory notation, this arrangement is represented as a domain-by-rater-within-teacher or $d \times (r:t)$ design. Domains are crossed with raters and teachers because all teachers were scored on the same survey (see section 3.4.2). Raters are nested because each teacher taught a different group of students (e.g., teacher A was rated by students 1-10, whereas teacher B by students 11-20). In each analysis, we included all survey respondents per lesson. The effect of lesson difficulty on reliability cannot be estimated because students were surveyed only once, so we cannot know how the same students would have rated their teacher on a different lesson. Importantly, this design treats students as interchangeable raters, despite the fact that they may have different experiences and they cannot be conceived as interchangeable (Seidel, 2006). As the G-studies in the prior research section indicate, however, other studies also make this simplifying assumption.

3.2.2 School year

During the school year, teachers taught in multiple schools, grades, and subjects for 11 months. Each teacher was scored on two occasions by one rater on the same domains as in clinical practice. We refer to occasions here because these observations occurred at different time points, unlike the lessons in clinical practice, which took place in close succession. This is a teacher-by-domain-by-occasion or $t \times d \times o$ design. Domains are crossed with everything else for the same reasons as above. Occasions are also crossed because all teachers were observed at the middle and end of the year. In 2014, both observations occurred in the same school, grade, section, and subject to keep them comparable; in 2015 they were conducted in different classes to be more

comprehensive. The effect of rater stringency on reliability cannot be estimated because there was only one rater per occasion. Each rater observed multiple teachers, so we conduct a separate study for each rater (3 in 2014 and 4 in 2015) and report the average result across raters. As in clinical practice, the assignment of teachers to raters was not random but determined by logistical constraints.

The students of each teacher were surveyed twice using the same tool from clinical practice. These are separate domain-by-rater-within-teacher or $d \times (r{:}\, t)$ designs per occasion. Domains are crossed with everything else for the same reasons as above. Raters are nested within teachers because each teacher has a different set of students. As in clinical practice, we included all survey respondents per occasion. We run a separate analysis per occasion—instead of crossing occasions with everything else—because surveys were anonymous, so we cannot ensure that the students on both occasions are the same students. Further, in 2014, ExA surveyed the same group of students on both occasions, but in 2015 it surveyed different classes. Therefore, raters are unlikely to be crossed with occasions in 2014 and they are definitely not crossed with occasions in 2015.

3.3 Sampling

Our sampling frame includes 100 unique teachers who participated in ExA in 2014 and 2015: 23 began the program prior to 2014 and remained, 32 started in 2014, and 45 started in 2015. We have data on the last two cohorts for clinical practice and the school year, but we only see the first cohort during the school year because it completed clinical practice before our study.

Our samples for each analysis do not include all teachers in a given cohort. Some teachers were observed fewer times than the rest, so we drop them to ensure all teachers have enough data to estimate relevant variance components. During clinical practice, some teachers

were observed by the same coach or peer on both lessons and others by a different peer per lesson (see section 3.2.1). We analyze each group separately. Some teachers are included in multiple analyses, but none contributes more than once to the same analysis. Table 1 shows the number of teachers, lessons or occasions, raters, and domains, and the design for each analysis.

3.4 Measures

The classroom observations and student surveys that ExA used were adapted from U.S. protocols. This is quite common in LMICs, where researchers simplify, combine, and/or translate U.S.-developed instruments, often conducting validation studies (Araujo et al., 2016; Bruns et al., 2018; Bruns et al., 2016; Bruns & Luque, 2014; Filmer et al., 2022; Molina et al., 2018). This feature of our study reduces concerns that any differences in reliability between our study and prior research are driven purely by instrument construction. Our evidence therefore speaks to the "portability" of reliability when similar instrument families are implemented in a middle-income context and under operational, practitioner-led administration. We do not claim to separate these influences; instead, we treat the study as a joint "stress test" of whether research-setting reliability evidence is a reasonable guide in a policy and practitioner reality.

Reliability is an important but not sufficient property when using observation and survey scores to support inferences about teaching. Validity is a degree of evidence supporting the interpretation and use of scores in a particular context and is supported by multiple sources of evidence (AERA/APA/NCME, 2018). Our contribution is not a comprehensive validation of these measures in Argentina, but an examination of whether operational, practitioner-led administration can yield sufficiently reliable scores under realistic constraints. We cite prior validation research on related instruments (largely U.S.-based) while emphasizing that such evidence may not fully transport to this setting. For context, we report descriptive correlations

between these measures in Appendix A (Tables A.3–A.5). These associations are not intended as confirmatory validity tests and may be attenuated by measurement error and affected by contextual differences in implementation and interpretation.

3.4.1 Classroom observations

ExA developed its classroom-observation protocol based on five measures created and administered in the United States: the Classroom Assessment Scoring System (CLASS, Allen et al., 2013; Bell et al., 2012; Hafen et al., 2014; Hamre et al., 2013; Mashburn et al., 2008; Pianta et al., 2020; Pianta et al., 2008b; Sandilos & DiPerna, 2014); the Framework for Teaching (FfT, Danielson, 2011; Kane et al., 2011; Kimball et al., 2009; Milanowski, 2009; Patrick et al., 2019); Teaching As Leadership (TAL, Farr 2010); the Protocol for Language Arts Teaching Observation (PLATO, Grossman et al., 2015; Grossman et al., 2014; Grossman et al., 2013; Lockwood et al., 2015); and Mathematical Quality of Instruction (MQI, Blazar, 2015; Blazar et al., 2017; Blazar & Kraft, 2017; Hill et al., 2008; Hill et al., 2012a; Hill et al., 2011; Hill et al., 2012b; Learning Mathematics for Teaching Project, 2011).

The process to translate and combine these instruments proceeded as follows. The director of ExA's Training and Coaching team translated the English version of the instruments into Spanish. Then, the team discussed which aspects of each instrument were valued by the organization and could be feasibly assessed by a coach or peer through clinical practice and the school year. Next, the director drafted an instrument that combined aspects from all the instruments, organizing them around six domains (e.g., "presenting content clearly"), writing items for each of them (e.g., "does the teacher announce clearly what students are going to learn at the start of class?") and descriptors for each performance level for each indicator (e.g., from 1 "no, the teacher never mentions to what students are going to learn" to 5 "yes, the teacher

mentions what students are going to learn at the start of class using an outline"). The director and members of the team engaged in multiple rounds of discussions to decide on changes. The initial implementation of the protocol in 2013 informed further changes for 2014 and beyond. There was no back-translation or piloting, as there may be when instruments are used for research.

The protocol covered six domains: presenting content clearly, checking for understanding, managing student behavior, implementing class procedures, creating an environment conducive to learning, and developing a sense of possibility. Each domain was scored based on five to seven items on a 1 (pre-novice) to 5 (exemplary) scale. Each item included a brief description for each possible score to assist raters with their selection. We include histograms of the lesson- and teacher-level scores (Figure A.1), bar graphs of the domain-level ratings (Figure A.3), and tables with correlations among them (Tables A.2-A.3) in Appendix A. We describe the domains and provide translated example items for the protocol, and link to the full protocol, in Appendix B. The director of ExA's Training and Coaching team translated the English version of the survey into Spanish, sought feedback from its team members, and deployed it for the first time in 2013. This initial implementation led to minor changes for the version used in 2014 and beyond.

3.4.2 Student surveys

ExA translated the Tripod survey (Ferguson 2010, 2012). The survey covers seven domains: care (attending to students' needs), confer (engaging students in conversations), captivate (motivating students to learn), clarify (checking for students' understanding), consolidate (helping students integrate concepts), challenge (having high standards for students), and control (managing students' behavior). Each domain was scored based on two to seven items on a 1 ("never") to 5

("always") scale. The distributions of rater-, teacher-, and domain-level scores are in Appendix A (Figures A.2 and A.4), and the descriptions of domains and example items in Appendix B.

The distributions of scores are skewed toward the high end of the scale, which may affect our reliability estimates in two ways. First, ceiling effects at the item level could attenuate reliability, though this concern is mitigated when scores are averaged across multiple items, since composite scores tend to be less skewed than individual item scores (Ho & Kane, 2013). Second, if the population of teachers in our study exhibits less true-score variance than the populations for which these instruments were developed—a restriction of range—observed reliability coefficients will be lower regardless of ceiling effects.

4. Analysis

4.1 Generalizability studies

We estimate the reliability of the classroom observations and student surveys during clinical practice and the school year conducting G-studies. In all studies, we conceive of the observed score X_i that a teacher receives in replication i as composed of a universe score τ (i.e., long-run average over replications) and *multiple* facets of error (e.g., deviations from τ due to differences in domain difficulty or rater stringency). In each study, we decompose observed-score variance into universe-score variance (i.e., actual differences in effectiveness) and different types of error variance (i.e., differences due to facets of error and interactions between them).

As discussed in section 3.2, the study design or way in which teachers were assigned to domains, lessons or occasions, and raters differed across both contexts and years. Each of these designs allows us to distinguish between different sources of error variance. Below, we explain how we analyze data from each design using random-effects models. Standard errors are

computed via the delta method. Because variance components are bounded below zero (and generalizability coefficients are bounded between 0 and 1), Wald-style symmetric confidence intervals based on roughly two standard errors may extend beyond the parameter space, especially for small samples or components near zero. We report standard errors to convey estimation uncertainty and avoid formal "significance" claims for individual variance components.

As also stated in section 3.2, ExA uses individual domain scores to decide what type of support teachers need. Our analyses, however, treat domains as a random facet, analogous to items in conventional G-studies. We decompose variance at the domain level and report reliability for the average score across all domains, since averaging over domains increases the precision of the composite score, as reflected in the D-study equations below.

4.1.1 The $d \times (l:t)$ and $d \times (r:t)$ designs

As explained in sections 3.2.1- 3.2.2, classroom observations during clinical practice follow a $d \times (l:t)$ design and student surveys during clinical practice and the school year follow a $d \times (r:t)$ design. In both, all teachers are scored on the same domains, but each teacher faces different lessons or raters. These designs let us distinguish between five sources of variance:

$$X_{dl:t} = \mu + \nu_t + \nu_d + \nu_{l:t} + \nu_{dt} + \nu_{dl:t,e} \quad (1)$$

or

$$X_{dr:t} = \mu + \nu_t + \nu_d + \nu_{r:t} + \nu_{dt} + \nu_{dr:t,e}, \quad (2)$$

where $X_{dl:t}$ or $X_{dr:t}$ is the observed score for teacher t on domain d, assessed on lesson l or by rater r; μ is the grand mean (i.e., the average score across all teachers, domains, and lessons or raters); ν_t is the teacher effect (i.e., how much teacher t differs in their performance); ν_d is the domain effect (i.e., how much domain d differs in its difficulty); $\nu_{l:t}$ or $\nu_{r:t}$ are the lesson or

rater effect (i.e., how much lesson l differs in its difficulty or rater r in their stringency), nested within teachers; ν_{dt} is the domain-by-teacher effect (i.e., how much domain d differs in its difficulty for teacher t); and $\nu_{dl:t,e}$ or $\nu_{dr:t,e}$ is the domain-by-lesson or domain-by-rater effect (i.e., how much domain d differs in its difficulty for lesson l or rater r), nested within teachers and confounded with residual variation. The parameters of interest are not these random effects, but their variances, which are estimated directly via restricted maximum likelihood.

In these designs, we can estimate relative error variance $\hat{\sigma}_\delta^2$ (i.e., variation in scores from the facets of error that affect the relative standing or ranking of teachers) using the formulas:

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{l:t}^2}{n_l} + \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{dl:t,e}^2}{n_d n_l}, \tag{3}$$

or

$$\hat{\sigma}_\delta^2 = \frac{\hat{\sigma}_{r:t}^2}{n_r} + \frac{\hat{\sigma}_{dt}^2}{n_d} + \frac{\hat{\sigma}_{dr:t,e}^2}{n_d n_r}, \tag{4}$$

where $\hat{\sigma}_{l:t}^2$ and $\hat{\sigma}_{r:t}^2$ are the estimated variances from lessons and raters, nested within teachers; $\hat{\sigma}_{dt}^2$ is the variance from the domain-by-teacher interaction; $\hat{\sigma}_{dl:t,e}^2$ or $\hat{\sigma}_{dr:t,e}^2$ is the variance from the interaction between domains and lessons or raters, nested within teachers and confounded with residual error; and n_d, n_l, and n_r are the numbers of domains, lessons, and raters.

We can also estimate absolute error variance $\hat{\sigma}_\Delta^2$ (i.e., variation in scores from the facets of error that affect not only rankings, but also teachers' locations on the score scale) as:

$$\hat{\sigma}_\Delta^2 = \frac{\hat{\sigma}_d^2}{n_d} + \hat{\sigma}_\delta^2, \tag{5}$$

where $\hat{\sigma}_d^2$ is the estimated domain variance and everything else is as above.

We can use our estimates of relative and absolute error variance to obtain generalizability coefficients for relative and absolute error $\mathbb{E}\hat{\rho}^2$ and Φ. These are akin to a reliability coefficients

from CTT like Cronbach's alpha, but they are more general because they take into account error variance stemming from multiple facets of error and from interactions among them. They define reliability as the share of total variance explained by universe score variance:

$$\mathbb{E}\hat{\rho}^2 = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{\sigma}_\delta^2} \tag{6}$$

and

$$\hat{\Phi} = \frac{\hat{\sigma}_t^2}{\hat{\sigma}_t^2 + \hat{\sigma}_\Delta^2} \tag{7}$$

where $\hat{\sigma}_t^2$ is the estimated universe-score variance and all else is as above. These formulas are always the same regardless of the study design, so we do not repeat them below.

4.1.2 The $t \times d \times o$ design

As explained in section 3.2.2, during the school year classroom observations follow a $t \times d \times o$ design. In this design, all teachers are scored on the same domains and occasions. This design allow us to decompose observed scores into seven sources of variance:

$$X_{tdo} = \mu + \nu_t + \nu_d + \nu_o + \nu_{dt} + \nu_{to} + \nu_{do} + \nu_{tdo,e}, \tag{8}$$

where X_{tdo} is the observed score for teacher t on domain d and occasion o; μ is the grand mean; ν_t is the teacher effect; ν_d is the domain effect; ν_o is the occasion effect (i.e., how much occasion o differs in its difficulty); ν_{dt} is the domain-by-teacher effect; ν_{to} is the teacher- by-occasion effect (i.e., how much teacher t differs in their performance on occasion r); ν_{do} is the domain-by-occasion effect (i.e., how much domain d differs in its difficulty on occasion o); and $\nu_{tdo,e}$ is the teacher-by-domain-by-occasion effect (i.e., how much teacher t differs in their performance on domain d and occasion o), confounded with residual error.

We can estimate relative error variance as:

$$\hat{\sigma}_{\delta}^{2} = \frac{\hat{\sigma}_{dt}^{2}}{n_d} + \frac{\hat{\sigma}_{to}^{2}}{n_o} + \frac{\hat{\sigma}_{tdo,e}^{2}}{n_d n_o}, \quad (9)$$

where $\hat{\sigma}_{to}^{2}$ and $\hat{\sigma}_{tdo,e}^{2}$ are the estimated variances for the teacher-by-occasion and teacher- by-domain-by-occasion interactions; n_d and n_o are the numbers of domains and occasions; and everything else is as above. We can also estimate absolute error variance as:

$$\hat{\sigma}_{\Delta}^{2} = \frac{\hat{\sigma}_{d}^{2}}{n_d} + \frac{\hat{\sigma}_{o}^{2}}{n_o} + \frac{\hat{\sigma}_{do}^{2}}{n_d n_o} + \hat{\sigma}_{\delta}^{2}, \quad (10)$$

where $\hat{\sigma}_{d}^{2}$, $\hat{\sigma}_{o}^{2}$, and $\hat{\sigma}_{do}^{2}$ are the estimated variances for domains, occasions, and the domain- by-occasion interaction; and everything else is as above.

4.2 Decision studies

We then identify the optimal approach to increase the reliability of classroom observations and student surveys using D(ecision) studies. In each D-study, we take the generalizability coefficients for relative and absolute error of a study design, which capture the reliability of these instruments under the current conditions, and calculate how they would change if we averaged over more raters and lessons or occasions in each measurement procedure. As explained in section 4.1, these coefficients are derived from the estimates of relative and absolute error variance based on the variance components from each G-study. The calculation of these variances includes the number of replications for each facet of error in each design. By letting some of these numbers vary, we can anticipate their expected impact on reliability.

4.2.1 The $\boldsymbol{d} \times (\boldsymbol{l}:\boldsymbol{t})$ and $\boldsymbol{d} \times (\boldsymbol{r}:\boldsymbol{t})$ designs

As equations (5)-(7) show, in these designs, relative and absolute error variance and their generalizability coefficients depend partly on the number of domains and lessons or raters. Thus, if we increased any of them, error variance would decrease and reliability would increase. This makes intuitive sense: if teachers are scored on more domains or lessons or by more raters, their

scores should be more reliable (because we are increasing the number of replications). We will assume that the observation protocol and survey have strong theoretical justifications and estimate how increasing the number of lessons or raters would impact their reliability. We will let the number of lessons or raters vary in the calculation of relative error variance:

$$\hat{\sigma}_{\delta}^{2} = \frac{\hat{\sigma}_{l:t}^{2}}{n_l} + \frac{\hat{\sigma}_{dt}^{2}}{n_d} + \frac{\hat{\sigma}_{dl:t,e}^{2}}{n_d n_l'}, \tag{11}$$

or

$$\hat{\sigma}_{\delta}^{2} = \frac{\hat{\sigma}_{r:t}^{2}}{n_r} + \frac{\hat{\sigma}_{dt}^{2}}{n_d} + \frac{\hat{\sigma}_{dr:t,e}^{2}}{n_d n_r'}, \tag{12}$$

and also in the calculation of absolute error variance:

$$\hat{\sigma}_{\Delta}^{2} = \frac{\hat{\sigma}_{d}^{2}}{n_d} + \hat{\sigma}_{\delta}^{2}, \tag{13}$$

where n_l' and n_r' are the number of lessons and raters that are allowed to vary and everything else is as above. If we increased these numbers, error variance would decrease (because they are in the denominator of both sets of expressions) and the generalizability coefficients would increase (because error variance in their denominators; see equations (6) and (7).

4.2.2 The $t \times d \times o$ design

As equations (9)-(10) show, in this design, relative and absolute error variance and their generalizability coefficients depend partly on the number of domains and occasions. If we again hold the number of domains constant in observations and surveys, we can let the number of occasions vary to estimate how increasing them would impact relative error variance:

$$\hat{\sigma}_{\delta}^{2} = \frac{\hat{\sigma}_{dt}^{2}}{n_d} + \frac{\hat{\sigma}_{to}^{2}}{n_o'} + \frac{\hat{\sigma}_{tdo,e}^{2}}{n_d n_o'}, \tag{14}$$

and absolute error variance:

$$\hat{\sigma}_{\Delta}^{2} = \frac{\hat{\sigma}_{d}^{2}}{n_{d}} + \frac{\hat{\sigma}_{o}^{2}}{n_{o}'} + \frac{\hat{\sigma}_{do}^{2}}{n_{d}n_{o}'} + \hat{\sigma}_{\delta}^{2}, \quad (15)$$

where n_o' is the varying number of occasions and everything else is as above.

5. Results

5.1 Classroom observations

Classroom observations in this setting vary widely in their reliability for making both relative distinction and absolute judgments about teachers. As Table 2 shows, the coefficients for relative error (on the third row from the bottom) ranged from 0 to 0.75 and those for absolute error (on the second-to-last row) ranged from 0 to 0.71. To illustrate what these coefficients imply in practice, consider a teacher who receives an average observation score of 3 on a 1 to 5 scale (close to the mean in Table 1). The standard error of measurement (SEM) for a single observation ranges from 0.184 to 0.340 across conditions. In the best case, this yields a 95% confidence interval of [2.65, 3.35], spanning only 0.70 points on the scale. In the worst case, it yields an interval of [2.33, 3.67] or 1.34 points—about a third of the scale. The vast degree of variability indicates that observations in these settings can range from pure noise to helpful mechanisms to assign teachers in need of support to resources, training, and/or coaching. As the standard errors around the variance components indicate, both point estimates and differences among them should be interpreted with caution given known imprecision from small sample sizes.

To situate these results in the prior literature, we computed the reliability of a single observation (i.e., one lesson scored by one rater, averaged across all six domains) from the variance components in Table 2, setting n_l or n_o to 1 and n_d to 6. These single-observation generalizability coefficients for relative error range from 0 to 0.61 across conditions, with a mean of 0.39. This is remarkably close to the mean of 0.39 and median of 0.38 across the 44 estimates

in Table A.1—even though our observations were conducted by practitioners rather than trained research raters and that they were conducted in a middle-income country.

On average, observations conducted during clinical practice had lower levels of reliability than those during the school year. The mean generalizability coefficient for relative error across all rater assignments and years in clinical practice was 0.48 and the one for the school year was 0.64. The mean coefficient for absolute error was 0.42 and the one for the school year was 0.52. The reliability of observations also varied more during clinical practice. The coefficients for relative and absolute error in clinical practice ranged from 0 to 0.75 and from 0 to 0.71, respectively. Those for the school year ranged from 0.63 to 0.66 and from 0.44 to 0.59.

One factor that might explain the differences in the level and variability of reliability between clinical practice and the school year is the context in which observations were conducted. If we compare observations in which the same coach scored both lessons across clinical practice and the school year, the former had lower reliability than the latter. The generalizability coefficients for relative error ranged from 0.34 to 0.52 during clinical practice and from 0.63 to 0.66 during the school year, and the ones for absolute error ranged from 0.29 to 0.33 during clinical practice and from 0.44 to 0.59 during the school year.

Another factor that may explain these differences is the different ways in which peer raters were assigned to teachers during clinical practice. If we compare observations in which the same peer scored both lessons to those in which a different peer scored each lesson, the former has higher and less variable reliability than the latter. The coefficients for relative and absolute error were around 0.75 and 0.71 on both years when the same peer scored both lessons and they ranged between 0 and 0.51 and from 0 to 0.48 when a different peer scored each lesson. Observations in which the same coach scored both lessons had lower and more variable

reliability than those in which the same peer scored both lessons. The coefficients for relative and absolute error ranged from 0.34 to 0.52 and from 0.29 to 0.33 when the same coach scored both lessons and they were around 0.75 and 0.71 on both years when the same peer scored both lessons. We consider several mechanisms that could explain this difference in the discussion.

As we state in section 3.2.1 and 3.2.2, however, rater assignment was not randomized. To the extent that teachers were systematically assigned to particular rater types or to more/less stringent raters, our variance components and therefore our generalizability coefficients may partly reflect selection and pairing rather than measurement properties alone. For example, because generalizability coefficients depend on between-teacher variance, range restriction in the subset of teachers observed under a given configuration could mechanically lower reliability. We therefore interpret differences across rater types and configurations with caution.

Except for clinical-practice observations in which a different peer scored each lesson, the coefficients for relative and absolute error varied little from one year to the next. This is consistent with our conversations with practitioners, in which they did not flag any changes in the training of peers that could explain fluctuations in the reliability of observations conducted by a different peer per lesson. For clinical-practice observations with a different peer per lesson, the year-on-year variability could be due to composition of the cohort of teachers across years or other factors we do not observe. As stated elsewhere, however, variance components are estimated with error, and differences in reliability across years should be interpreted with caution.

Tables presenting the results from score variance decompositions typically also include columns indicating the percentage of total variance that each variance component represents. It is important to remember, however, that variance components are estimated variances of

distributions of the most elemental scores (e.g., in the $t \times d \times o$ design, X_{tdo} or the observed score for teacher t on domain d and occasion o), not the average scores that we typically use (e.g., in the same design, $\bar{X}_t$ or the average score for teacher t across domains and occasions). To describe the importance of a source of error in terms of its impact on reliability for the scores that we more commonly use, we report the results of our D-studies.

Increasing the reliability from observations is feasible during all but one of the scenarios we considered. As Figure 1 shows, the generalizability coefficient for relative error in clinical-practice observations is between 0.2 and 0.6 when each teacher is rated twice, regardless of the rater type (see the y-coordinate of the blue lines at 2 lessons in panels A–F), except for clinical-practice observations in which a different peer scored each lesson in 2014. Adding a lesson would improve this coefficient by 6-9 pp. (see the y-coordinate of the same lines at 3 lessons). Further increases in the number of lessons would only improve reliability by 3-6 pp., despite making such observations more logistically complex (see the increasingly flat slopes of these lines beyond 3 lessons). The coefficient for relative error in school-year observations is between 0.4 and 0.5 when each teacher is rated twice (see panels G-H). Adding an occasion would improve this coefficient by 8 pp., but further increases would improve it further by only 5 pp.

Adding lessons or occasions would have a smaller impact on the reliability of absolute judgments. As Figure 1 shows, the generalizability coefficient for absolute error is between 0.3 and 0.7 during clinical practice when a teacher is rated twice by the same coach or a different peer on each lesson (see the y-coordinates of the red lines at 2 lessons in panels A-B and E-F). Adding a lesson would raise this coefficient by 4-7 pp. (see y-coordinates of the same lines at 4 lessons in panels A, E-F). Further increases in the number of lessons would improve reliability by 2-5 pp. The pattern is similar for the school year. The coefficient for absolute error is between

0.4 and 0.6 when each teacher is rated twice (see panels G-H). Adding an occasion would improve this coefficient by 8 pp., but further increases would improve it by only 5 pp.

To decide whether these improvements are worth pursuing, however, it is important to consider the practical costs (e.g., time, money, logistics) of adding observations. During clinical practice, when all teachers and coaches are in the same venue for several consecutive weeks, adding a lesson seems less costly than during the school year, when it entails an additional trip from a coach to the teacher's school (and potentially, less attention to the other teachers they coach). Within clinical practice, adding a lesson scored by peers is easier (assuming that there are more teachers available than observing a lesson at any given time) than adding a lesson scored by coaches (assuming that they continue to be paid ExA employees and that they are likely observing a lesson at every slot during clinical practice).

5.2 Student surveys

Student surveys also vary widely in their reliability. As Table 3 shows, the coefficients for relative error ranged from 0.33 to 0.85 and those for absolute error from 0.29 to 0.70. To illustrate what these coefficients imply in practice, consider a teacher who receives an average score of 3.9 on a 1 to 5 scale (close to the mean in Table 1). The SEM for a single survey ranges from 0.119 to 0.173 across conditions. In the best case, this yields a 95% confidence interval of [3.67, 4.13], spanning 0.47 points on the scale. In the worst case, it yields an interval of [3.56, 4.24] or 0.68 points—less than a fifth of the scale. These intervals suggest that student surveys, while generally more precise than classroom observations, still carry meaningful uncertainty when administered only once.

To compare with the prior literature, we computed the reliability of a single student's response (setting n_r to 1 and n_d to 7) from the variance components in Table 3. These single-

rater generalizability coefficients for relative error range from 0.02 to 0.35, broadly consistent with Schweig (2014)'s estimate of 0.16 for a single Tripod administration (Table A.1). As with classroom observations, these comparisons suggest that the measurement challenges documented in U.S. research settings extend to practitioner-led implementations in different countries.

As in the case of observations, surveys administered during clinical practice had lower and more variable reliability than those during the school year. The mean generalizability coefficient for relative error for clinical practice was 0.46 and the one for the school year was 0.72. The mean coefficient for absolute error for clinical practice was 0.43 and the one for the school year was 0.53. The coefficients for relative and absolute error in clinical practice ranged from 0.33 to 0.58 and from 0.29 to 0.57. Those for the school year ranged from 0.59 to 0.85 and from 0.36 to 0.70. During the school year, reliability was lowest in 2015, when ExA switched from surveying the same students twice to surveying different groups of students (see section 3.2.2). More broadly, ExA made few changes in the study designs for student surveys, so the apparent stability in reliability estimates may be partly a function of only two designs being compared.

Increasing the number of raters would improve the reliability of relative judgments in all but one of the scenarios considered. As Figure 2 shows, the generalizability coefficient for relative error in clinical-practice surveys is between 0.15 and 0.55 with 10 students (see the y-coordinates of the blue lines at 10 students in panels A-B). Adding 5 students from the same class would improve reliability by 9 pp. in 2015 (see y-coordinates of these lines at 15 students); adding 5 more would only do so by 6 pp. (see y-coordinates at 20 students). The impact of adding raters in the school year is slightly lower. The coefficient for relative error is between

0.39 and 0.77 with 10 students. Adding 5 students would increase it by 5-8 pp., and adding 5 more would only do so by 3-6 pp.

Adding raters would have a similar impact on the reliability of absolute judgments. As Figure 2 shows, the generalizability coefficient for absolute error is between 0.14 and 0.54 for clinical-practice surveys with 10 students (see the y-coordinates of the red lines at 10 students in panels A-B). Adding 5 students would improve reliability by 9 pp. in 2015 (see y-coordinates at 15 students), but adding 5 more would only do so by 6 pp. (see y-coordinates at 20 students). Again, adding raters would have a smaller impact on reliability in the school year. The coefficient for absolute error is between 0.28 and 0.67 with 10 students. Adding 5 students would increase it by 4 pp., but 5 more would only do so by 2-3 pp.

Once again, these improvements in reliability ought to be considered against their costs. Given that ExA surveyed all attending students, some of the increases considered here (e.g., adding five more students) would improve reliability at no additional costs. Even surveying 20 students should be doable given average number of survey respondents during clinical practice and the school year (see Table 3).

6. Discussion

In this paper, we presented one of the first G-studies of two non-test measures of teaching effectiveness in a middle-income country: classroom observations and student surveys. Our motivation was twofold. First, prior G-studies relied on data collected for research, with several quality-assurance mechanisms in place, so we evaluated whether their results are indicative of the reliability of instruments administered for practice. Second, past G-studies have focused on high-income countries (primarily, the U.S.), so we evaluated whether they are representative of

the realities of less established instruments administered in LMICs. We obtained data from an education non-profit that is part of a global network and examined the reliability of its metrics.

Our results suggest that measures of teaching effectiveness administered by practitioners in a middle-income country vary widely. We believe that this finding is important because it demonstrates that practitioners cannot rely on the reliability estimates from studies conducted in research settings in high-income countries to design their own teacher feedback systems. Despite the growing adoption of these instruments in LMICs (Araujo et al., 2016; Bruns et al., 2018; Bruns et al., 2016; Bruns & Luque, 2014; Filmer et al., 2022; Molina et al., 2018), there had been no empirical examinations of this question, which has implications for teacher development. Importantly, however, our study differs from prior G-studies in multiple respects simultaneously—including the country, the instruments, the raters, and the quality-assurance mechanisms—so we cannot isolate which of these factors drives the similarities or differences in our findings relative to prior work.

We also found that the reliability of classroom observations varied based on the context in which they were administered: clinical practice or school year. Specifically, we find that observations are less reliable and more variable during clinical practice. This makes intuitive sense: it seems plausible that new teachers will vary more in their first opportunities to teach, with students who are not their own, than once they settle into their teaching position; it also seems possible that peers and new coaches may still be getting used to using the observation protocol during this period. Both factors could result in lower and more variable reliability.

The reliability of observations also varied based on the ways in which teachers were assigned to raters: whether the same coach or peer scored both lessons, or whether a different peer scored each lesson. Observations scored by the same coach are more reliable during the

school year than during clinical practice, which makes sense for the same reasons as above. Observations scored by the same peer are more reliable than those scored by a different peer, which also makes sense: agreement among new teachers is likely rarer than consistency in the ratings of a teacher across lessons. Several mechanisms could explain this difference, including range restriction in which teachers were observed under each rater type, systematic pairing of teachers to raters or rater types, and coaches narrowing and raising scores to encourage teachers. Yet we do not find clear evidence for any of these mechanisms. In the absence of random assignment of raters to teachers, we interpret differences between rater assignments as further evidence that reliability is sensitive to design features, rather than as a property of rater types.

We partnered with this alternative pathway into teaching because it is a member of an international network that uses similar practices to train and provide feedback to its teachers. We reported on the contexts, rater configurations, rater types, and parameters (generalizability coefficients for relative v. absolute error) that maximize reliability in our setting. We showed that G-studies can be helpful not only to understand the reliability of a measurement procedure, but also to improve it. We demonstrated that how increases in the number of lessons or occasions in classroom observations, or of raters in student surveys, can meaningfully improve reliability, while drawing attention to the diminishing marginal returns of further expansions. We hope that the processes and insights from our study will be relevant to these similar programs, potentially impacting thousands of teachers every year and the students they serve.

Yet, to answer the question of whether classroom observations and student surveys administered by practitioners can produce reliable results, there ought to be more analyses of data collected by governments and non-profits, especially in LMICs. This is particularly important given evidence that non-test measures of teaching effectiveness are associated with

student characteristics and environmental factors (Cohen & Goldhaber, 2016; Whitehurst et al., 2014), which may limit the external validity of findings across contexts. We hope future studies build on our work by conducting similar analyses of locally developed instruments.

The high degree of variability in our reliability estimates indicate that practitioners should use our study as a guide for conducting their own analyses, instead of treating our estimates as benchmarks for their specific contexts. Yet we encourage practitioners to estimate reliability under their actual operational design and assignment mechanism, and where feasible, build in limited randomization and double scoring to properly separate variance components and estimate and predict reliability. To support them in this process, we have explained how to understand the design of each measurement procedure and how to analyze the data that each produce. We have also made the datasets and code from our analyses available with this paper.

In using G-studies to design their teacher feedback systems, practitioners should note that the recommendations from D-studies are estimated with considerable imprecision. A recent study used Bayesian estimation to reanalyze data from a G-study in the medical field and found that the minimum number of raters needed to achieve adequate levels of reliability was higher and more variable than the study had implied (Himmelsbach 2025). Therefore, we do not recommend that practitioners conduct a single G- and D-study and assume that their reliabilities are precise nor that they apply to all subsequent administrations of their instruments. As our analysis illustrates, there can be significant year-on-year variation in the reliability of non-test measures of teaching effectiveness administered by practitioners. Instead, we encourage practitioners to regularly examine the reliability of their measures and make adjustments as needed. Our results suggest that this approach would be preferable to assuming that the recommendations from formal research settings generalize to their own.

Lastly, it should also be noted that like all G-studies, ours only estimate variance from the facets explicitly included in each design. Facets that are not modeled (e.g., the timing of observations within the school year, the subject matter being taught, or student composition) are absorbed into residual variance. To the extent that these omitted facets introduce systematic rather than random error, our generalizability coefficients may overestimate the reliability of these instruments in practice (Bell et al., 2012).

Even with this limitation, G-theory is useful precisely because it clarifies which sources of error we consider. This clarity allows practitioners to treat reliability as an empirical property of their local design, while benefiting from theory that allows them to anticipate how reliability will change when features such as rater assignment, timing, or sample sizes change.

7. References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. *Journal of Labor Economics*, *25*(1), 95-135.

Adelman, M., & Lemos, R. (2018). *Managing for learning: Measuring and strengthening education management in Latin America and the Caribbean*. The World Bank.

AERA/APA/NCME. (2018). *Standards for educational and psychological testing*. American Educational Research Association (AERA), American Psychological Association (APA), and National Council on Measurement in Education (NCME).

Allen, J. P., Gregory, A., Mikami, A., Lun, J., Hamre, B. K., & Pianta, R. C. (2013). Observations of effective teacher-student interactions in secondary school classrooms: Predicting student achievement with the Classroom Assessment Scoring System--Secondary. *School Psychology Review*, *42*(1), 76-98.

Allen, M. J., & Yen, W. M. (2001). *Introduction to measurement theory*. Waveland Press.

Anand, G., Atluri, A., Crawfurd, L., Pugatch, T., & Sheth, K. (2023). Improving school management in low and middle income countries: A systematic review. *Economics of Education Review*, *97*, 102464.

Araujo, M. C., Carneiro, P., Cruz-Aguayo, Y., & Schady, N. (2016). Teacher quality and learning outcomes in kindergarten. *Quarterly Journal of Economics*, *131*(3), 1415-1453.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E. H., Ladd, H. F., Linn, R. L., Ravitch, D., Rothstein, R., Shavelson, R. J., & Shepard. (2010). *Problems with the use of student test scores to evaluate teachers*. Economic Policy Institute (EPI).

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C., & Qi, Y. (2012). An argument approach to observation protocol validity. *Educational Assessment*, *17*(2-3), 62-87.

Blazar, D. (2015). Effective teaching in elementary mathematics: Identifying classroom practices that support student achievement. *Economics of Education Review*, *48*, 16-29.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy*, *13*(3), 281-309.

Blazar, D., Braslow, D., Charalambous, C. Y., & Hill, H. C. (2017). Attending to general and mathematics-specific dimensions of teaching: Exploring factors across two observation instruments. *Educational Assessment*, *22*(2), 71-94.

Blazar, D., & Kraft, M. A. (2017). Teacher and teaching effects on students' attitudes and behaviors. *Educational Evaluation and Policy Analysis*, *39*(1), 146-170.

Brennan, R. L. (2001). *Generalizability theory*. Springer-Verlag.

Bruns, B., Costa, L., & Cunha, N. (2018). Through the looking glass: Can classroom observation and coaching improve teacher performance in Brazil? *Economics of Education Review*, *64*, 214-250.

Bruns, B., De Gregorio, S., & Taut, S. (2016). *Measures of effective teaching in developing countries*. Research on Improving Systems of Education (RISE).

Bruns, B., Filmer, D., & Patrinos, H. A. (2011). *Making schools work: New evidence on accountability reforms*. World Bank Publications.

Bruns, B., & Luque, J. (2014). *Great teachers: How to raise student learning in Latin America and the Caribbean*. The World Bank.

Chetty, R., Friedman, J., & Rockoff, J. E. (2014). Measuring the impacts of teachers I: Evaluating bias in teacher value-added estimates. *American Economic Review*, *104*(9), 2593-2632.

Cohen, J., & Goldhaber, D. (2016). Building a more complete understanding of teacher evaluation using classroom observations. *Educational Researcher*, *45*(6), 378-387.

Commission/EACEA/Eurydice, E. (2021). In *Teachers in Europe: Careers, development, and well-being*. Publications Office of the European Union.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles*. John Wiley & Sons. https://doi.org/dx.doi.org/10.2307/1162145

Danielson, C. (2011). *Enhancing Professional Practice: A Framework for Teaching*. Association for Supervision; Curriculum Development (ASCD).

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan*, *93*(6), 8-15. https://doi.org/10.1177/003172171209300603

Dee, T. S., & Wyckoff, J. (2013). Incentives, selection, and teacher performance: Evidence from IMPACT. *Journal of Policy Analysis and Management*, *34*, 267-297.

Douglas, K. M., & Mislevy, R. J. (2010). Estimating classification accuracy for complex decision rules based on multiple scores. *Journal of Educational and Behavioral Statistics*, *35*(3), 280-306.

Filmer, D., Molina, E., & Wane, W. (2022). *Identifying effective teachers: Lessons from four classroom observation tools*. Research on Improving Systems of Education (RISE).

Glazerman, S., Goldhaber, D., Loeb, S., Raudenbush, S. W., Staiger, D. O., & Whitehurst, G. J. (2011). *Passing muster: Evaluating teacher evaluation systems*. Brown Center on Education Policy, Brookings Institute.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*.

Goldhaber, D., Grout, C., & Huntington-Klein, N. (2017). Screen twice, cut once: Assessing the predictive validity of applicant selection tools. *Education Finance and Policy*, *12*(2), 197-223.

Grossman, P., Cohen, J., & Brown, L. (2015). Understanding instructional quality in English language arts: Variations in PLATO scores by content and context. In T. J. Kane, K. A. Kerr, & R. C. Pianta (Eds.), *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. Wiley Online Library.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher*, *43*(6), 293-303. https://doi.org/https://doi.org/10.3102/0013189X14544542

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school english language arts and teachers' value-added scores. *American Journal of Education*, *119*(3), 445-470.

Hafen, C. A., Hamre, B. K., Allen, J. P., Bell, C. A., Gitomer, D. H., & Pianta, R. C. (2014). Teaching through interactions in secondary classrooms: Revisiting the factor structure and practical application of the Classroom Assessment Scoring System—Secondary. *The Journal of Early Adolescence*, *35*(5-6), 651-680.

Hamre, B. K., Pianta, R. C., Downer, J. T., DeCoster, J., & Mashburn, A. J. (2013). Teaching through interactions: Testing a developmental framework of teacher effectiveness in over 4,000 classrooms. *The Elementary School Journal*, *113*(4), 461-487.

Hill, H. C., Blunk, M. L., Charalambous, C. Y., Lewis, J. M., Phelps, G. C., Sleep, L., & Ball, D. L. (2008). Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and instruction*, *26*(4), 430-511.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012a). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal*, *48*(3), 794-831.

Hill, H. C., Umland, K., Litke, E., & Kapitula, L. R. (2012b). Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. *American Journal of Education*, *118*(4), 489-519.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation.

Isoré, M. (2009). *Teacher evaluation: Current practices in OECD countries and a literature review*. Organization for Economic Cooperation and Development (OECD).

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics*, *95*(4), 1096-1116. https://doi.org/10.1162/rest_a_00339

Jackson, C. K. (2020). What do test scores miss? The importance of teacher effects on non-test-score outcomes. *Journal of Political Economy*, *126*(5), 2072-2107.

Jackson, C. K., & Bruegmann, E. (2009). Teaching students and teaching each other: The importance of peer learning for teachers. *American Economic Journal: Applied Economics*, *1*(4), 85-108.

Jerald, C. (2012). *Ensuring accurate feedback from observations*. Bill and Melinda Gates Foundation.

Joe, J. N., Tocci, C. M., Holtzman, S. L., & Williams, J. C. (2013). *Foundations of observation: Considerations for developing a classroom observation system that helps districts achieve consistent and accurate scores*. Bill and Melinda Gates Foundation.

Johnson, S. M., Kraft, M. A., & Papay, J. P. (2012). How context matters in high-need schools: The effects of teachers' working conditions on their professional satisfaction and their students' achievement. *Teachers College Record*, *114*(10), 1-39.

Kane, T. J., Pianta, R. C., & Kerr, K. A. (2014). *Designing teacher evaluation systems: New guidance from the Measures of Effective Teaching project*. John Wiley & Sons.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review*, *27*(6), 615-631.

Kane, T. J., & Staiger, D. O. (2011). *Learning about teaching: Initial findings from the Measures of Effective Teaching project*. Bill and Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains*.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources*, *46*(3), 587-613.

Kimball, S. M., White, B., Milanowski, A. T., & Borman, G. D. (2009). Examining the relationship between teacher evaluation and student assessment results in Washoe County. *Peabody Journal of Education*, *79*(4), 54-78.

Koedel, C., Mihaly, K., & Rockoff, J. E. (2015). Value-added modeling: A review. *Economics of Education Review*, *47*, 180-195.

Kraft, M. A. (2019). Teacher effects on complex cognitive skills and social-emotional competencies. *Journal of Human Resources*, *54*(1), 1-36.

Kraft, M. A., & Gilmour, A. F. (2017). Reivisiting the widget effect: Teacher evaluation reforms and the distribution of teacher effectiveness. *Educational Researcher*, *46*(5), 234-249. https://doi.org/https://doi.org/10.3102/0013189X17718797

Learning Mathematics for Teaching Project. (2011). Measuring the mathematical quality of instruction. *Journal of Mathematics Teacher Education*, *14*(1), 25-47.

Lockwood, J. R., Savitsky, T. D., & McCaffrey, D. F. (2015). Inferring constructs of effective teaching from classroom observations: An application of Bayesian exploratory factor analysis without restrictions. *The Annals of Applied Statistics*, *9*(3), 1484-1509.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. Information Age Publishing.

Mantzicopoulos, P., French, B. F., & Patrick, H. (2018). The Mathematical Quality of Instruction (MQI) in kindergarten: An evaluation of the stability of the MQI using generalizability theory. *Early Education and Development*, *29*(6), 893-908.

Manzi, J., González, R., & Sun, Y. (2011). *La evaluación docente en Chile*. MIDE UC, Centro de Medición, Pontificia Universidad Católica de Chile.

Martinez, F., Taut, S., & Schaaf, K. (2016). Classroom observation for evaluating and improving teaching: An international perspective. *Studies in Educational Evaluation*, *49*, 15-29.

Martínez, J. F., & Fernández, M. P. (2021). Teacher evaluation with multiple indicators: Conceptual and methodological considerations regarding validity. In J. Manzi, S. Taut, & M. R. García (Eds.), *Validity of educational assessments in Chile and Latin America* (pp. 373-394). Springer International Publishing.

Mashburn, A. J., Pianta, R. C., Hamre, B. K., Downer, J. T., Barbarin, O. A., Bryant, D., Burchinal, M., Early, D. M., & Howes, C. (2008). Measures of classroom quality in prekindergarten and children's development of academic, language, and social skills. *Child development*, *79*(3), 732-749.

McClellan, C. (2013). *What it looks like: Master coding videos for observer training and assessment*. Bill and Melinda Gates Foundation.

MET Project. (2010). *Validation engine for observational protocols*. Bill and Melinda Gates Foundation.

MET Project. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET Project's three-year study*. Bill and Melinda Gates Foundation.

Meyer, J. P., Cash, A. H., & Mashburn, A. J. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment*, *16*(4), 227-243.

Milanowski, A. (2009). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education*, *79*(4), 33-53.

Molina, E., Fatima, S. F., Ho, A. D., Melo Hurtado, C., Wilichowksi, T., & Pushparatnam, A. (2018). *Measuring teacher practice at scale: Results from the development and validation of the Teach classroom observation tool*. The World Bank.

Mulhern, C. (2023). Beyond teachers: Estimating individual school counselors' effects on educational attainment. *American Economic Review*, *113*(11), 2846-2893.

Nunnally, J. C. (1978). *Psychometric theory (2nd edition)*. McGraw.

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis*, *26*(3), 237-257.

OECD. (2015). *Education at a glance 2015: OECD indicators*.

OECD. (2025). *Results from TALIS 2024: The state of teaching*. Organization for Economic Cooperation and Development (OECD).

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal*, *48*(1), 163-193.

Papay, J. P., Taylor, E. S., Tyler, J. H., & Laski, M. E. (2020). Learning job skills from colleagues at work: Evidence from a field experiment using teacher performance data. *American Economic Journal: Economic Policy*, *12*(1), 359–388.

Patrick, H., French, B. F., & Mantzicopoulos, P. (2020). The Reliability of Framework for Teaching Scores in Kindergarten. *Journal of Psychoeducational Assessment*, *38*(7), 831-845.

Patrick, H., Mantzicopoulos, P., & French, B. F. (2019). The predictive validity of classroom observations: Do teachers' Framework for Teaching scores predict kindergarteners' achievement and motivation? . *American Educational Research Journal*, *57*(5), 2021-2058. https://doi.org/https://doi.org/10.3102/0002831219891409

Pianta, R. C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F. J. (2008a). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal*, *45*, 365-397. https://doi.org/10.3102/0002831207308230

Pianta, R. C., Hamre, B. K., & Nguyen, T. (2020). Measuring and improving quality in early care and education. *Early Childhood Research Quarterly*, *51*, 285-287.

Pianta, R. C., La Paro, K. M., & Hamre, B. K. (2008b). *The classroom assessment manual, pre-k*. Brookes.

Pouezevara, S., Pflepsen, A., Nordstrum, L., King, S., & Gove, A. (2016). *Measures of quality through classroom observation for the Sustainable Development Goals: Lessons from*

low- and middle-income countries. RTI International; United Nations Educational, Scientific, and Cultural Organization (UNESCO).

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12.

Quezada, M. C., & Salcedo, M. P. M. (2019). *Desarrollo de instrumentos de evaluación: Pautas de obesrvación*. Centro de Medición (MIDE UC) and Instituto Nacional para la Evalaución de la Educación (INEE).

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. *Econometrica*, *73*(2), 417-458.

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, 247-252.

Rockoff, J. E., Jacob, B. A., Kane, T. J., & Staiger, D. O. (2011). Can you recognize an effective teacher when you recruit one? *Education Finance and Policy*, *6*(1), 43-74.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics*, *125*(1), 175-214.

Sandilos, L. E., & DiPerna, J. C. (2014). Measuring quality in kindergarten classrooms: Structural analysis of the Classroom Assessment Scoring System (CLASS K-3). *Early Education and Development*, *25*(6), 894-914.

Schweig, J. D. (2014). Quantifying error in survey measures of school and classroom environments. *Applied Measurement in Education*, *27*(2), 133-157.

Seidel, T. (2006). The role of student characteristics in studying micro teaching-learning environments. *Learning Environments Research*, *9*(3), 253-271.

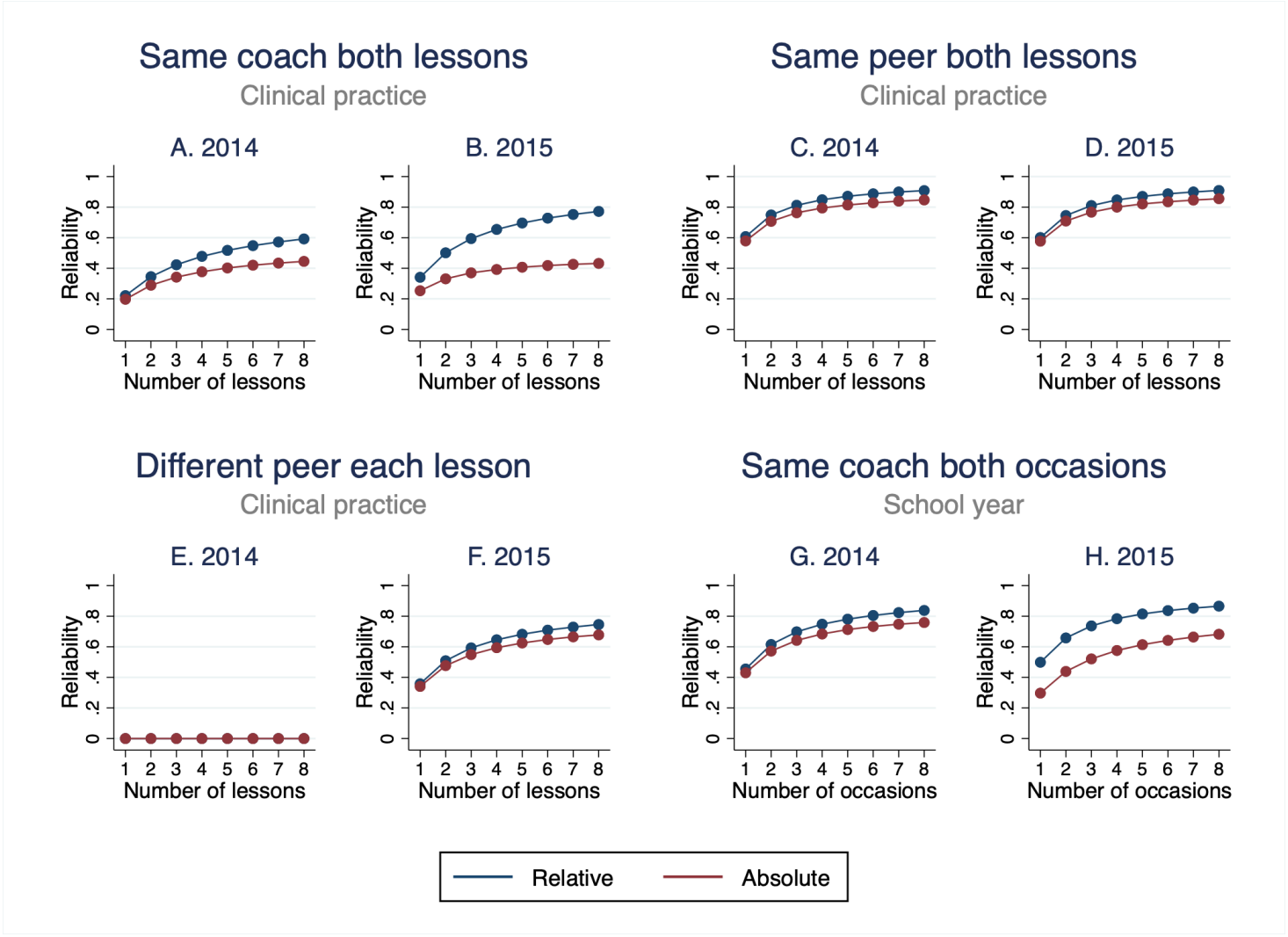
Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Sage.

van der Lans, R. M., van de Grift, W. J. C. M., van Veen, K., & Faokkens-Bruinsma, M. (2016). Once Is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, *50*, 88–95.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on diferences in teacher effectiveness*. The New Teacher Project (TNTP).

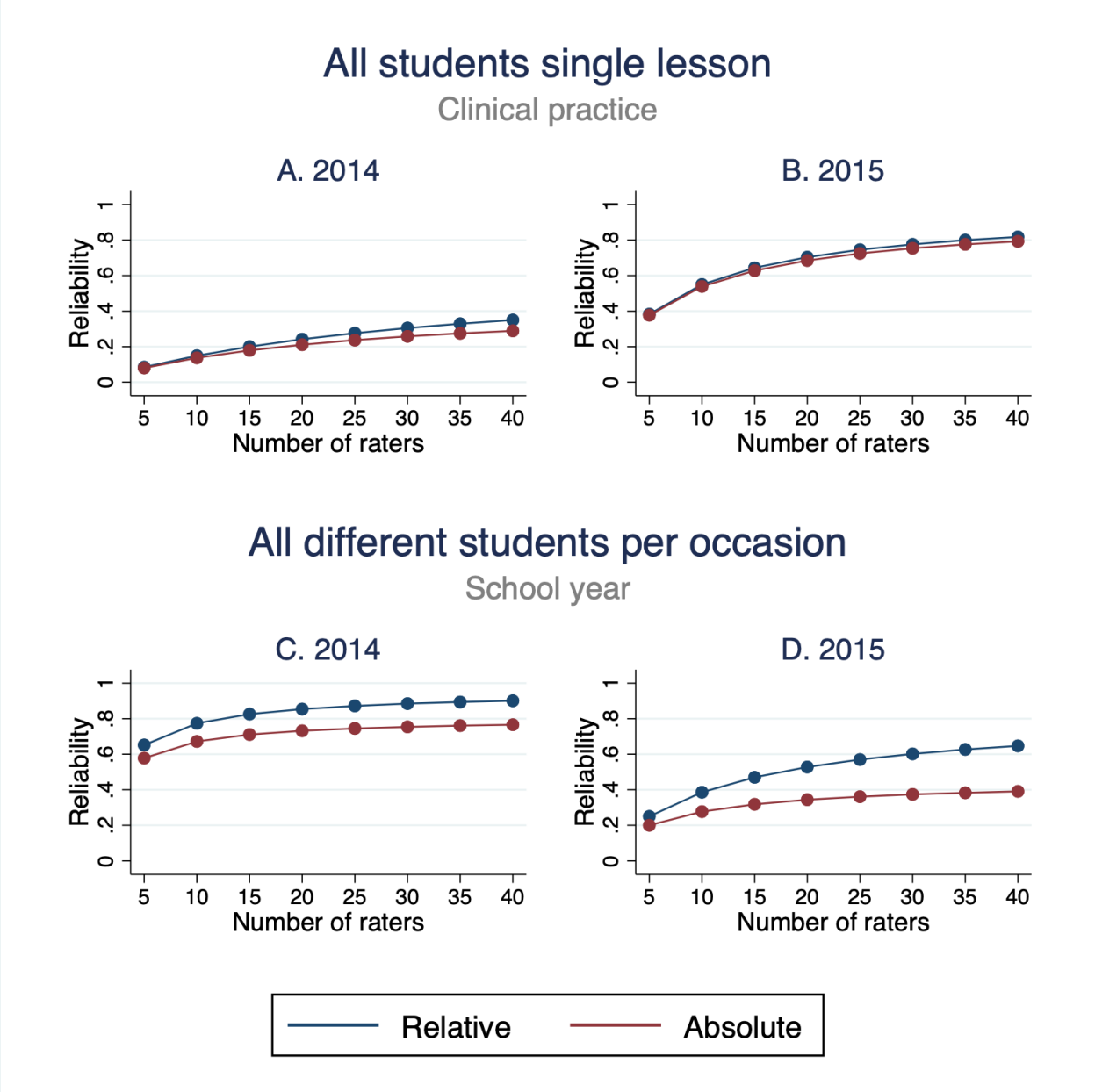
Whitehurst, G., Chingos, M. M., & Lindquist, K. M. (2014). *Evaluating teachers with classroom observations*. Brown Center on Education Policy, Brookings Institute.

Figure 1: Reliability of classroom observations at different numbers of lessons, clinical practice and school year, 2014 and 2015



Notes: This figure shows how the reliability of classroom observations would change by increasing the number of lessons. It features all designs in Table 2. The blue line refers to the reliability of the relative standing of teachers and the red one to that of the absolute scores of teachers.

Figure 2: Reliability of student surveys at different numbers of raters, clinical practice and school year, 2014 and 2015



Notes: This figure shows how the reliability of student surveys would change by increasing the number of raters. It features all designs in Table 3. The blue line refers to the reliability of the relative standing of teachers and the red one to that of the absolute scores of teachers.

Table 1: Data-analytic samples, 2014 and 2015

(1) Context	(2) Year	(3) Teachers	(4) Lessons/ occasions	(5) Raters per lesson/occasion	(6) Domains	(7) Study design	(8) Mean	(9) SD
A. Classroom observations								
Clinical practice	2014	30	2	Same coach both lessons	6	$7[d \times (l:t)]$	2.48	0.73
	2015	37	2	Same coach both lessons	6	$8[d \times (l:t)$	2.60	0.77
	2014	25	2	Same peer both lessons	6	$d \times (l:t)$	2.95	0.78
	2015	43	2	Same peer both lessons	6	$d \times (l:t)$	3.59	0.73
	2014	25	2	Different peer per lesson	6	$d \times (l:t)$	2.92	0.70
	2015	20	2	Different peer per lesson	6	$d \times (l:t)$	3.49	0.73
School year	2014	48	2	Same coach both occasions	6	$3(t \times d \times o)$	3.01	0.77
	2015	35	2	Same coach both occasions	6	$4(t \times d \times o)$	2.96	0.72
B. Student surveys								
Clinical practice	2014	23	1	All students single lesson	7	$d \times (r:t)$	4.47	0.75
	2015	31	1	All students single lesson	7	$d \times (r:t)$	4.44	0.88
School year	2014	33	2	All different students per occasion	7	$2[d \times (r:t)]$	3.75	0.94
	2015	28	2	All different students per occasion	7	$2[d \times (r:t)]$	3.86	0.88

Notes: This table lists the number of teachers, lessons or occasions, raters per lesson, domains, and study design of the classroom observations and student surveys during clinical practice and the school year in 2014 and 2015. It also displays the mean and standard deviation of the "elemental" scores (i.e., at the teacher-by-lesson-by- rater-by-item level). In observations during clinical practice, there were 7 coaches in 2014 and 8 in 2015. In observations during the school year, there were 3 coaches in 2014 and 4 in 2015. In surveys during the school year, there were 2 occasions in both years. In all cases, we conduct separate G-studies (one per coach) and report the average results (see sections 3.2.1-3.2.2 for further details).

Table 2: Variance in domain scores across classroom observations, clinical practice and school year, 2014 and 2015

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Clinical practice						School year	
	Same coach both lessons		Same peer both lessons		Different peer per lesson		Same coach both occasions	
	2014	2015	2014	2015	2014	2015	2014	2015
Variance component	Var.	Var.	Var.	Var.	Var.	Var.	Var.	Var.
Teacher	.034	.037	.210	.180	0	.090	.113	.073
	(.028)	(.030)	(.083)	(.054)	(0)	(.062)	(.041)	(.035)
Domain	.117	.244	.100	.075	.093	.072	.080	.068
	(.041)	(.073)	(.066)	(.049)	(.061)	(.051)	(.037)	(.043)
Lesson : Teacher	.086	.036	.097	.084	.194	.117		
	(.032)	(.017)	(.037)	(.025)	(.048)	(.047)		
Occasion							0	.084
							(0)	(.095)
Domain × Teacher	.059	.015	.030	.022	0	.072	.034	.017
	(.019)	(.010)	(.021)	(.015)	(0)	(.028)	(.018)	(.016)
Occasion × Teacher							.082	.032
							(.026)	(.019)
Domain × Occasion							.001	.045
							(.005)	(.024)
Residual	.142	.160	.203	.190	.227	.196	.229	.221
	(.016)	(.018)	(.026)	(.018)	(.021)	(.028)	(.022)	(.022)
SD of teacher effect	.184	.192	.458	.424	0	.300	.336	.270
SEM of a single observation	.254	.184	.265	.248	.340	.295	.256	.193
Reliability of single observation								
Relative standing of teachers	.34	.52	.75	.75	0	.51	.63	.66
Absolute scores of teachers	.29	.33	.71	.71	0	.48	.59	.44
Number of teachers	30	37	25	43	25	20	48	35

Notes: This table shows the variance in classroom-observations scores by context and year. All columns show the variance components for the object of measurement (teacher variance) and each facet of error. Standard errors, computed via the delta method, are shown in parentheses. The standard deviation of the teacher effect is the square root of the universe-score variance. The standard error of measurement of a single observation is the square root of relative error variance. Components estimated as negative were set to zero. Components left blank for a design were not estimated for that design.

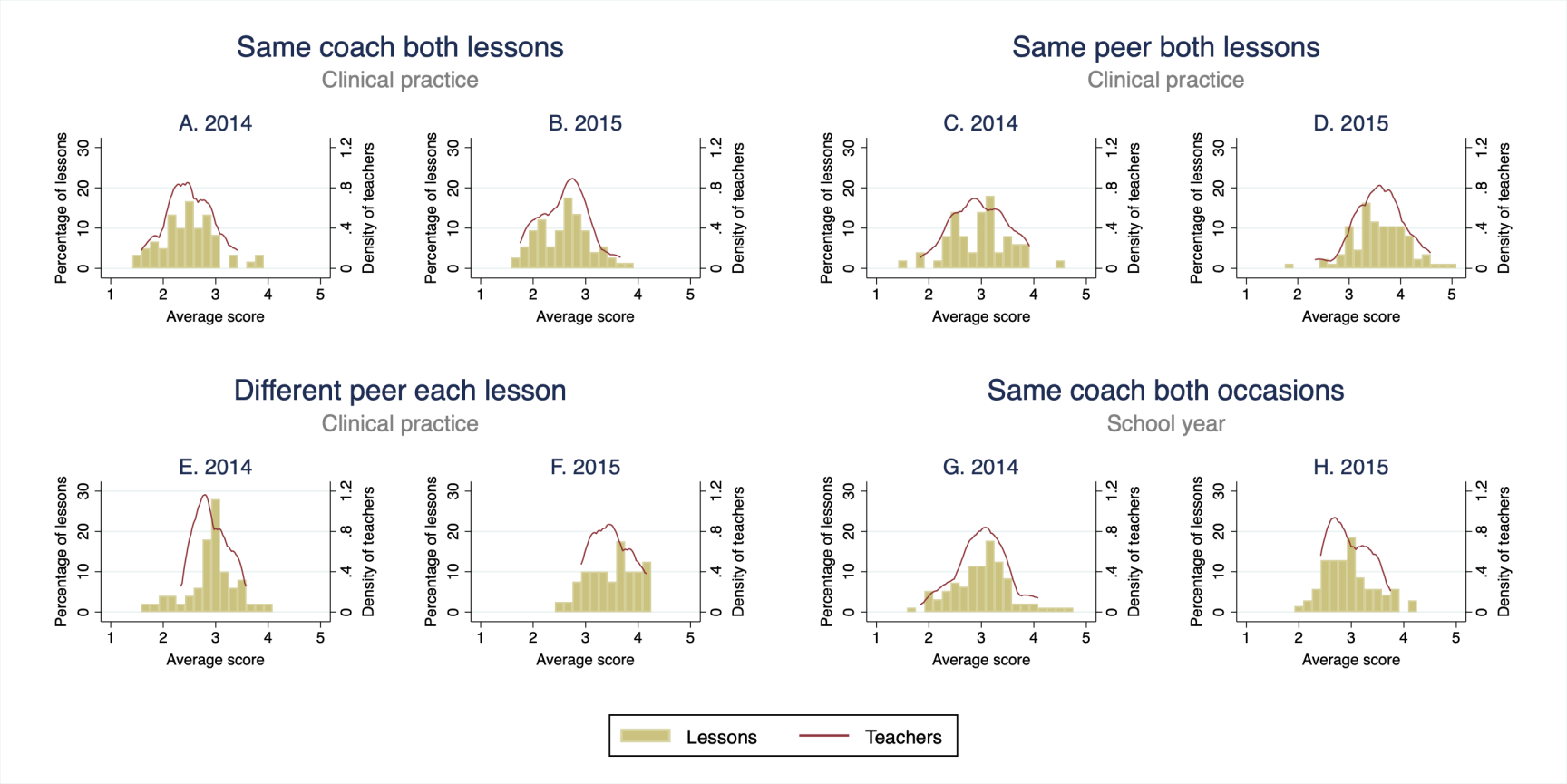
Table 3: Variance in domain scores across student surveys, clinical practice and school year, 2014 and 2015

	(1)	(2)	(3)	(4)
	Clinical practice		School year	
	All students single lesson		All different students per lesson	
	2014	2015	2014	2015
Variance component	Var.	Var.	Var.	Var.
Teacher	.007	.041	.094	.020
	(.006)	(.019)	(.020)	(.007)
Domain	.020	.011	.171	.152
	(.013)	(.007)	(.070)	(.063)
Rater : Teacher	.198	.506	.222	.213
	(.016)	(.033)	(.010)	(.011)
Domain x Teacher	.014	.004	.036	.026
	(.004)	(.002)	(.004)	(.004)
Residual	.296	.270	.399	.369
	(.008)	(.007)	(.006)	(.006)
SD of teacher effect	.084	.202	.307	.141
SEM of a single observation	.119	.173	.129	.119
Reliability of single observation				
Relative standing of teachers	.33	.58	.85	.59
Absolute scores of teachers	.29	.57	.70	.36
Number of teachers	23	31	33	25
Average number of raters	19.7	18.5	24	25.5

Notes: The table shows the variance in domain scores across four administrations of student surveys: two under clinical practice and two during the school year, in 2014 and 2015. All columns show the variance components for the object of measurement (teacher variance) and each facet of error. Standard errors, computed via the delta method, are shown in parentheses. The standard deviation of the teacher effect is given by the square root of the true score variance, and it corresponds to the distribution of average scores by teacher. The standard error of measurement of a single survey is given by the square root of relative error variance. Variance components estimated as negative have been set to zero. Variance components left blank for a design were not estimated for that design.

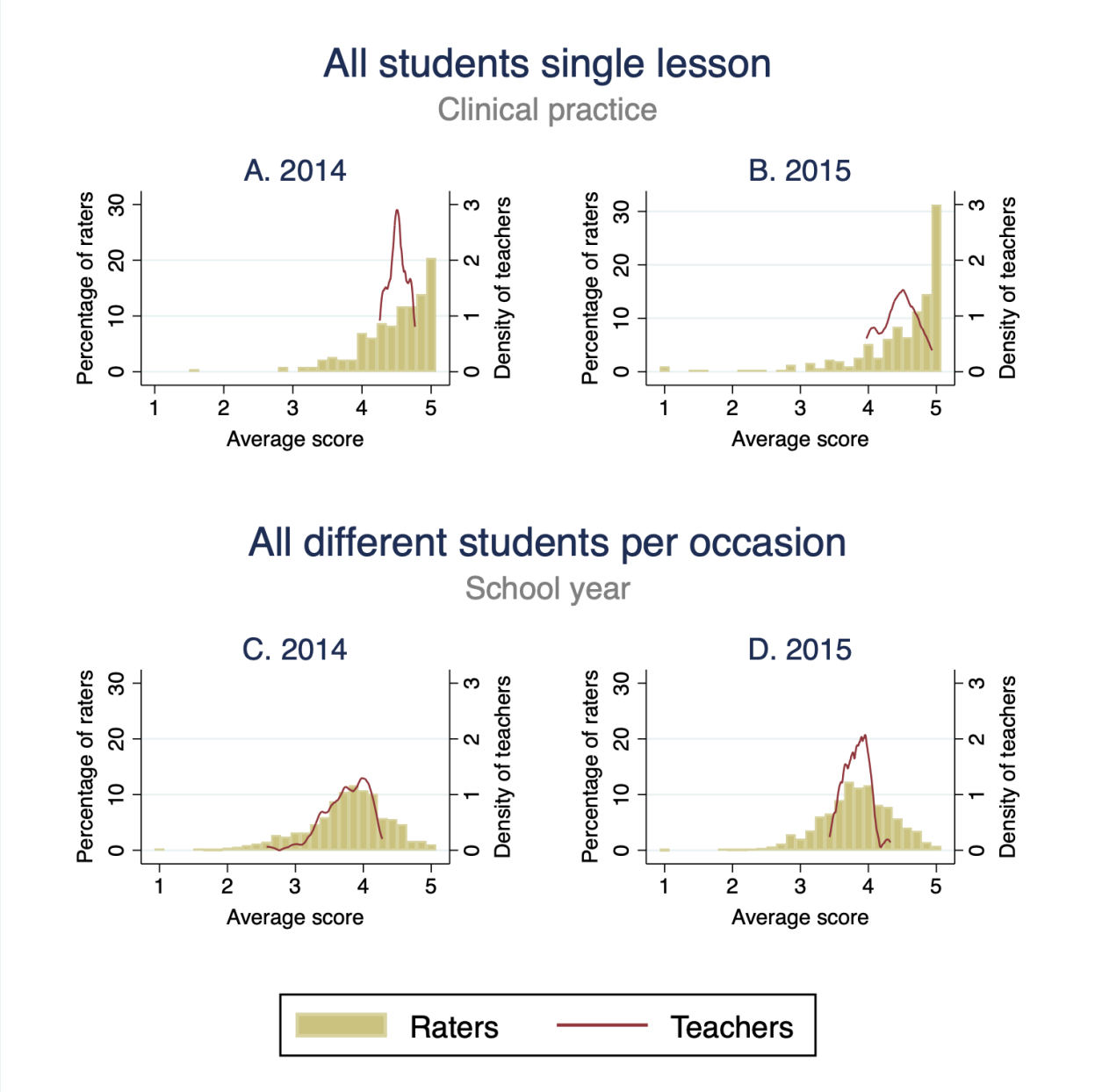
Appendix A: Additional Figures and Tables

Figure A.1: Distribution of lesson-/occasion-level and teacher-level average scores on classroom observations (2014 and 2015)



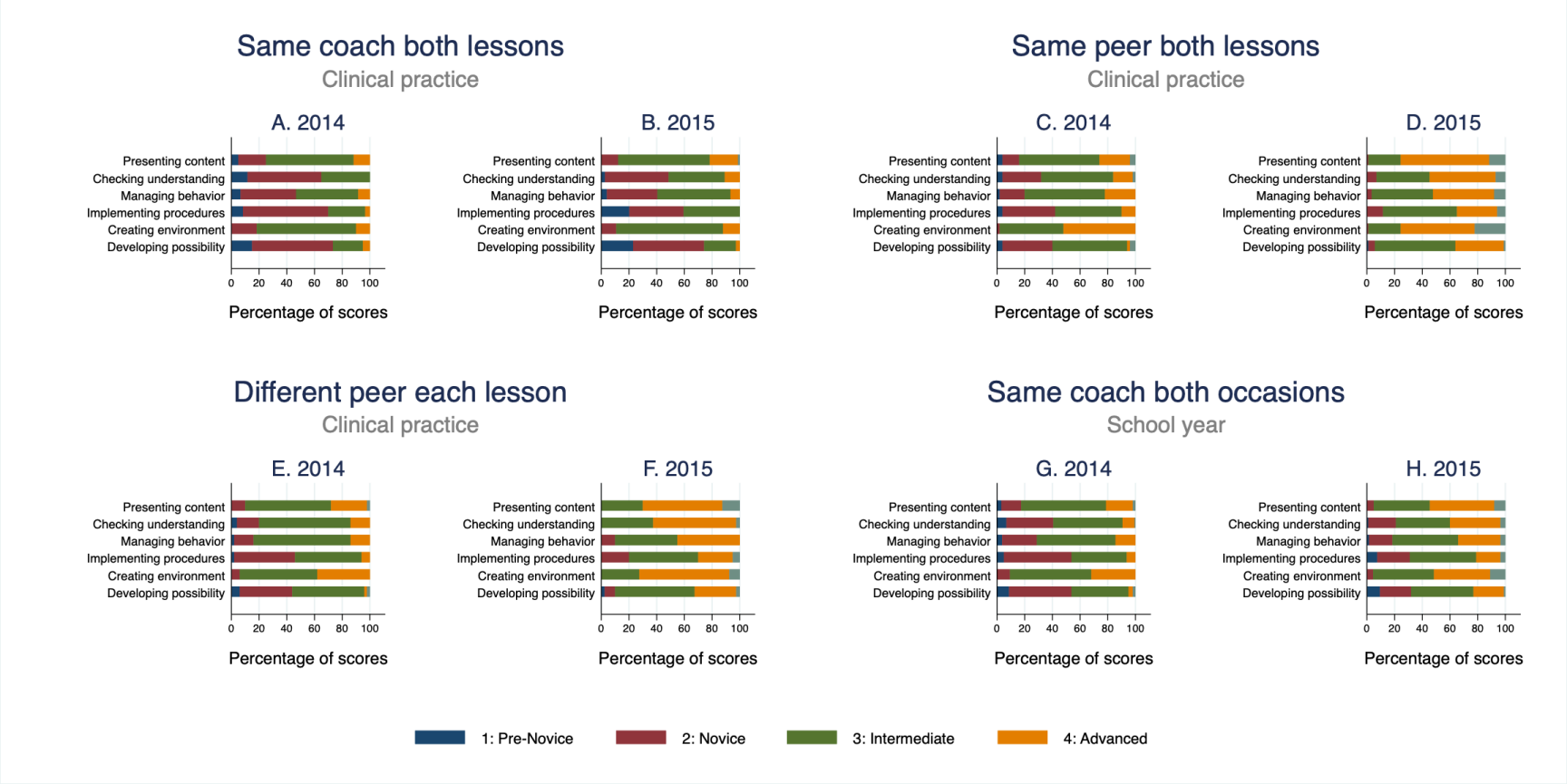
Notes: This figure shows the distribution of lesson- or occasion-level (histogram) and teacher-level (kernel plot) average scores on classroom observations of ExA teachers during clinical practice and the school year in 2014 and 2015.

Figure A.2: Distribution of rater-level and teacher-level average scores on student surveys (2014 and 2015)



Notes: This figure shows the distribution of rater-level (histogram) and teacher-level (kernel plot) average scores on student surveys of ExA teachers during clinical practice and the school year in 2014 and 2015.

Figure A.3: Distribution of domain-level scores on classroom observations (2014 and 2015)



Notes: This figure shows the distribution of domain-level scores on classroom observations of ExA teachers during clinical practice and the school year in 2014 and 2015. The six domains are: presenting content clearly, checking understanding, managing student behavior, implementing class procedures, creating a learning environment, and developing a sense of possibility (see section 3.4.1).

Figure A.4: Distribution of domain-level scores on student surveys (2015)

Notes: This figure shows the distribution of domain-level scores on student surveys on ExA teachers during clinical practice and the school year in 2014 and 2015. The seven domains are: care, confer, captivate, clarify, consolidate, challenge and control (see section 3.4.2).

Table A.1: Generalizability studies of classroom observations and student surveys

(1) Study	(2) Context	(3) Instrument	(4) Teachers	(5) Obs. per teacher	(6) Mean score	(7) SD of teacher effect	(8) SEM of single obs.	(9) Reliability of single obs.
Classroom observations								
A. Pre-primary								
Mantzicopoulos, French, Patrick, et al. (2018)	Midwestern U.S.	CLASS K-3-EMSUP	10	4	4.75/7	0.43	0.48	0.44
		CLASS K-3-CLORG			5.19/7	0.23	0.52	0.16
		CLASS K-3-INSUP			3.04/7	0.41	0.68	0.26
		FfT Class. environment			2.36/4	0.26	0.34	0.36
		FfT Class. instruction			1.87/4	0.19	0.44	0.16
Mantzicopoulos, French and Patrick (2018)	Midwestern U.S.	MQI-R	20	5		0.17	0.10	0.74
		MQI-WWSM				0.13	0.10	0.63
		MQI-EI				0.00		0.00
		MQI-CCASP				0.09	0.10	0.45
		Whole lesson				0.29	0.14	0.81
Patrick et al. (2020)	Indiana, IN	FfT Reading	20	10	2.47/4	0.37	0.24	0.70
		FfT Math			2.37/4	0.35	0.24	0.67
B. Primary								
Meyer et al. (2011)	Southeastern U.S.	CLASS-EMSUP	118	4	5.37/7	0.52	0.39	0.64
		CLASS-INSUP			2.88/7	0.22	0.43	0.20
		CLASS-CLORG			5.19/7	0.36	0.41	0.43
C. Secondary								
Hill et al. (2012)	Southwestern U.S.	MQI-R	8	1				0.45
		MQI-EI						0.37
		MQI-SPMMR						0.46
Casabianca et al. (2013)	Southeastern U.S.	CLASS-EMSUP (live)	82	4-5	3.69/7	0.73	0.70	0.52
		CLASS-EMSUP (video)			3.64/7	0.57	0.83	0.33
		CLASS-CLORG (live)			5.69/7	0.68	0.54	0.61
		CLASS-CLORG (video)			5.75/7	0.58	0.65	0.44
		CLASS-INSUP (live)			3.58/7	0.64	0.82	0.38

		CLASS-INSUP (video)			3.26/7	0.47	0.82	0.25
Mashburn, Meyer, et al. (2014)	Southeastern U.S.	CLASS-EMSUP	47	3	4.11/7	0.46	0.61	0.36
		CLASS-INSUP			3.21/7	0.43	0.74	0.25
		CLASS-CLORG			5.18/7	0.65	0.65	0.50
Mashburn, Downer, et al. (2014)	Brooklyn and Queens, NY	CLASS-EMSUP	48	6		0.48	0.53	0.45
		CLASS-INSUP				0.49	0.56	0.40
		CLASS-CLORG				0.56	0.68	0.44
Praetorius et al. (2014)	Germany and Switzerland	CLASS Class. mgmt.	38	5	3.64/7	0.54	0.40	0.64
		Personal learning support			2.60/7	0.37	0.56	0.45
		Cognitive activation			1.93/7	0.36	0.97	0.14
D. Multiple levels								
Kane and Staiger (2012)	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough Co., FL; New York, NY; Memphis, TN	FfT	1333	4-8		0.29	0.38	0.37
		CLASS						0.31
		PLATO						0.34
		MQI						0.14
		UTOP	1000					0.30
Ho and Kane (2013)	Hillsborough Co., FL	FfT	67	46	2.58/4	0.27	0.34	0.39
Semmelroth and Johnson (2014)	Idaho	RESET-Lesson obj.	9	3		0.43	0.82	0.22
		RESET-EBP imp.				0.91	1.98	0.17
		RESET-Whole lesson				1.16	2.11	0.23
van der Lans et al. (2016)	Netherlands	ICALT3	69	3		1.14		0.51
Briggs and Alzen (2019)	Charlotte-Mecklenburg, NC; Dallas, TX; Denver, CO; Hillsborough Co., FL; New York, NY; Memphis, TN	FfT	458	7-8	2.50/4	0.28	0.37	0.36

Student surveys								
E. Multiple levels								
Schweig (2014)	Urban California district	Tripod-Challenge	285	17		0.30	0.17	0.16

Notes: This table provides an overview of prior studies on the reliability of classroom observations and student surveys. The standard deviation (SD) of the teacher effect is the square root of true-score variance. The standard error of measurement (SEM) of a single observation is the square root of relative-error variance. The reliability of a single observation is the proportion of observed-score variance attributable to persistent differences in teacher practice, estimated for a single observation by a single rater. Cells left blank refer to values not reported. CLASS stands for Classroom Assessment Scoring System. EMSUP, INSUP, and CLORG are its domains: emotional support, instructional support, and classroom organization. MQI stands for Mathematics Quality of Instruction. R, WWSM, EI, and CCASP and SPMMR are its domains: richness, working with student and mathematics, errors and imprecision, common core-aligned student practices, and student participation in meaning making and reasoning. RESET for Recognizing Effective Special Education Teachers and its domains are: lesson objectives, evidence-based practice implementation, and whole-lesson review. FfT stands for Framework for teaching, PLATO for Protocol for Language Arts Teaching Observation, UTOP for UTeach Observation Protocol, ICALT3 for The International Comparative Analysis of Learning and Teaching. Luna-Bazaldua et al. (2021) reports results for four unidentified countries in the four listed regions. The SD of the teacher effect and SEM of a single observation for Kane and Staiger (2012) was obtained from Ho and Kane (2013). Ho and Kane (2013), each teacher was observed *on average* 46 observations per teacher by different observers and lessons. van der Lans et al. (2016) reported variance components on a logit scale.

Table A.2: Correlation between domain-level scores in classroom observations (2014)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Clinical practice						School year					
	Presenting content clearly	Checking understanding	Managing student behavior	Implementing class procedures	Creating learning environment	Developing sense of possibility	Presenting content clearly	Checking understanding	Managing student behavior	Implementing class procedures	Creating learning environment	Developing sense of possibility
A. Clinical practice												
Presenting content clearly	1.00											
Checking understanding	0.79***	1.00										
Managing student behavior	0.52***	0.39**	1.00									
Implementing class procedures	0.52***	0.40**	0.59***	1.00								
Creating learning environment	0.54***	0.38*	0.57***	0.67***	1.00							
Developing sense of possibility	0.69***	0.60***	0.42**	0.49**	0.65***	1.00						
B. School year												
Presenting content clearly	0.18	0.30	0.02	0.23	0.21	0.28	1.00					
Checking understanding	0.22	0.32	0.10	0.00	0.19	0.22	0.21	1.00				
Managing student behavior	0.27	0.39*	-0.03	0.12	0.24	0.35*	0.53***	0.50***	1.00			
Implementing class procedures	0.36*	0.32	-0.15	-0.01	0.06	0.25	0.25	0.58***	0.67***	1.00		
Creating learning environment	0.25	0.35*	0.02	0.12	0.21	0.40**	0.28	0.68***	0.53***	0.64***	1.00	
Developing sense of possibility	0.08	0.15	0.11	0.13	0.21	0.25	0.35*	0.47**	0.44**	0.50***	0.73***	1.00

Notes: This table shows the correlation coefficients between domain-level scores on classroom observations of ExA teachers during clinical practice and the school year in 2014. This table includes only the teachers with both sets of scores. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.3: Correlation between domain-level scores in classroom observations (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	Clinical practice						School year					
	Presenting content clearly	Checking understanding	Managing student behavior	Implementing class procedures	Creating learning environment	Developing sense of possibility	Presenting content clearly	Checking understanding	Managing student behavior	Implementing class procedures	Creating learning environment	Developing sense of possibility
A. Clinical practice												
Presenting content clearly	1.00											
Checking understanding	0.57**	1.00										
Managing student behavior	0.65***	0.65***	1.00									
Implementing class procedures	0.27	0.13	0.64***	1.00								
Creating learning environment	0.60***	0.66***	0.63***	0.39*	1.00							
Developing sense of possibility	0.61***	0.57**	0.74***	0.65***	0.72***	1.00						
B. School year												
Presenting content clearly	0.01	-0.14	0.09	0.20	-0.07	-0.02	1.00					
Checking understanding	-0.28	-0.01	-0.20	-0.33	-0.25	-0.54**	0.21	1.00				
Managing student behavior	0.21	0.39	0.25	-0.06	0.25	-0.10	0.50**	0.57**	1.00			
Implementing class procedures	-0.28	-0.18	-0.33	-0.49**	-0.38	-0.50**	0.25	0.75***	0.34	1.00		
Creating learning environment	0.15	0.13	0.27	0.33	0.14	-0.08	0.48**	0.59***	0.70***	0.18	1.00	
Developing sense of possibility	-0.11	0.07	-0.29	-0.47**	0.01	-0.38	0.16	0.68***	0.55**	0.58***	0.36	1.00

Notes: This table shows the correlation coefficients between domain-level scores on classroom observations of ExA teachers during clinical practice and the school year in 2015. This table includes only the teachers with both sets of scores. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.4: Correlation between domain-level scores in student surveys (2014)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Clinical practice							School year						
	Care	Confer	Captivate	Clarify	Consolidate	Challenge	Control	Care	Confer	Captivate	Clarify	Consolidate	Challenge	Control
A. Clinical practice														
Care	1.00													
Confer	0.21	1.00												
Captivate	0.58**	0.80***	1.00											
Clarify	0.05	0.62**	0.51*	1.00										
Consolidate	0.47	-0.21	-0.04	0.02	1.00									
Challenge	0.52*	0.19	0.45	-0.23	0.01	1.00								
Control	0.48*	0.23	0.44	-0.27	0.16	0.77***	1.00							
B. School year														
Care	0.09	0.71***	0.66**	0.29	-0.25	0.30	0.50*	1.00						
Confer	0.18	0.54*	0.71***	0.17	-0.30	0.41	0.42	0.75***	1.00					
Captivate	0.19	0.71***	0.73***	0.24	-0.34	0.31	0.38	0.90***	0.85***	1.00				
Clarify	0.15	0.73***	0.68**	0.19	-0.31	0.32	0.45	0.92***	0.75***	0.94***	1.00			
Consolidate	0.40	0.81***	0.88***	0.49*	-0.15	0.44	0.45	0.76***	0.83***	0.84***	0.77***	1.00		
Challenge	0.02	0.64**	0.65**	0.19	-0.41	0.19	0.33	0.91***	0.87***	0.96***	0.92***	0.75***	1.00	
Control	0.04	0.67**	0.62**	0.11	-0.42	0.49*	0.57**	0.92***	0.72***	0.88***	0.91***	0.73***	0.86***	1.00

Notes: This table shows the correlation coefficients between domain-level scores on student surveys on ExA teachers during clinical practice and the school year in 2014. This table includes only the teachers with both sets of scores. * significant at 10%; ** significant at 5%; *** significant at 1%.

Table A.5: Correlation between domain-level scores in student surveys (2015)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	Clinical practice							School year						
	Care	Confer	Captivate	Clarify	Consolidate	Challenge	Control	Care	Confer	Captivate	Clarify	Consolidate	Challenge	Control
A. Clinical practice														
Care	1.00													
Confer	0.45	1.00												
Captivate	0.98***	0.60	1.00											
Clarify	0.69	0.82*	0.76	1.00										
Consolidate	0.79	0.65	0.81*	0.94**	1.00									
Challenge	0.86*	0.45	0.83*	0.82*	0.96***	1.00								
Control	0.69	0.43	0.66	0.85*	0.95**	0.94**	1.00							
B. School year														
Care	-0.63	-0.68	-0.70	-0.71	-0.80	-0.76	-0.64	1.00						
Confer	-0.39	-0.66	-0.43	-0.91**	-0.84*	-0.69	-0.85*	0.52	1.00					
Captivate	-0.20	-0.87*	-0.33	-0.75	-0.65	-0.44	-0.50	0.74	0.76	1.00				
Clarify	0.23	0.71	0.36	0.44	0.15	-0.04	-0.05	-0.03	-0.26	-0.35	1.00			
Consolidate	-0.52	-0.75	-0.62	-0.65	-0.70	-0.63	-0.50	0.98***	0.47	0.80	-0.12	1.00		
Challenge	-0.71	-0.38	-0.71	-0.51	-0.71	-0.79	-0.60	0.92**	0.29	0.43	0.24	0.85*	1.00	
Control	-0.87*	-0.11	-0.82*	-0.26	-0.43	-0.60	-0.32	0.46	-0.11	-0.18	-0.01	0.36	0.68	1.00

Notes: This table shows the correlation coefficients between domain-level scores on student surveys on ExA teachers during clinical practice and the school year in 2015. This table includes only the teachers with both sets of scores. * significant at 10%; ** significant at 5%; *** significant at 1%.

Appendix B: Instruments

B.1 Classroom Observation

ExA developed its classroom-observation protocol based on prior measures and used it to provide feedback to its teachers during clinical practice and the school year. It covered six domains: presenting content clearly, checking for understanding, managing student behavior, implementing class procedures, creating an environment conducive to learning, and developing a sense of possibility. Each domain included five to seven items. Each item was scored from 1 (pre-novice) to 5 (exemplary). Each possible item score featured a brief description to help raters choose between them. The protocol can be accessed at: https://bit.ly/3NhS7nv.

Presenting content clearly included seven items: (a) does the teacher master the material?; (b) do they announce what students will learn at the beginning of class?; (c) do they use appropriate body language?; (d) does their explanation follow a clear structure?; (e) do they make effective use of visual aids?; (f) do they maintain an adequate pace?; (g) do they end the class reviewing key concepts or lessons learned?

Checking for understanding included seven items: (a) does the teacher ask students questions to check their understanding?; (b) do the questions span a wide range of skills?; (c) do all students participate in the questions?; (d) does the teacher offer feedback on students' answers?; (e) do they encourage students to talk to each other?; (f) do they respond to incorrect answers by helping students improve their answers?; and (g) do they manage to re-explain concepts that are not clear?

Managing student behavior included seven items: (a) does the teacher establish rules for behavior?; (b) do they enforce such rules consistently?; (c) do they minimize time spent on discipline issues?; (d) are there rewards and consequences when students follow the rules?; (e)

are such rewards and consequences commensurate to the rules being enforced?; (f) are teachers respectful to students when enforcing rules?; and (g) do they determine where students should sit to ensure the class runs as intended?

Implementing class procedures included five items: (a) has the teacher established routines for class procedures?; (b) do they implement these routines consistently?; (c) do they minimize time spent on class procedures?; (d) are there clear consequences for noncompliance with established routines?; and (e) does the teacher have a system to address exceptional circumstances?

Creating an environment conducive to learning included six items: (a) is the teacher respectful to students?; (b) do they ensure that students respect each other?; (c) do they make sure that students feel comfortable to ask questions?; (d) do they make sure that students feel comfortable to share mistakes in homework or classroom activities?; (e) do the classroom signs and rules facilitate a learning environment?; (f) does the teacher convey the learning goals for every lesson?

Developing a sense of possibility included seven items: (a) does the teacher recognize the students' strengths and improvements?; (b) do they demonstrate the appropriate procedures to solve problems in activities, homework, or assessments?; (c) do they advise students on how to study?; (d) do they provide model activities, homework, or assessments?; (e) do they convey the relevance of the content being taught?; (f) do they convey the importance of doing well in school?; (g) do they set high expectations for students?

B.2 Student Surveys

ExA adjusted and translated the Tripod survey (Ferguson, 2012) to provide feedback to teachers during clinical practice and the school year. It covered seven domains: care, confer, captivate,

clarify, consolidate, challenge, and control. Each item was scored from 1 (never) to 5 (always). The surveys are at: https://bit.ly/4dvIwV6 (primary) and https://bit.ly/3BGRNMw (secondary).

Care included six items: (a) I like the way my teacher treats me when I need help; (b) my teacher makes me feel that he/she really cares about me; (c) if I am sad or angry, my teacher helps me feel better; (d) my teacher encourages me to do my best; (e) my teacher knows if something is bothering me; and (f) my teacher gives us time to explain our ideas.

Confer included seven items: (a) when they are teaching us, my teacher asks us whether we understand; (b) my teacher asks questions to be sure we are following what they are saying; (c) my teacher checks to make sure we understand what he/she is teaching us; (d) my teacher tells us what we are learning and why; (e) my teacher wants us to share our thoughts; (f) students speak up and share their ideas about class work; (g) my teacher wants me to explain my answers—why I think what I think.

Captivate included two items: (a) schoolwork is interesting; and (b) homework helps me learn.

Clarify included seven items: (a) my teacher explains things in very orderly ways; (b) in this class, we learn to correct our mistakes; (c) my teacher explains difficult things clearly; (d) my teacher has several good ways to explain each topic that we cover in this class; (e) this class is neat—everything has a place and things are easy to find; and (f) if I don't understand something, my teacher explains it another way.

Consolidate included two items: (a) my teacher takes the time to summarize what we learn each day; and (b) when my teacher marks my work, they write on my papers to help me understand.

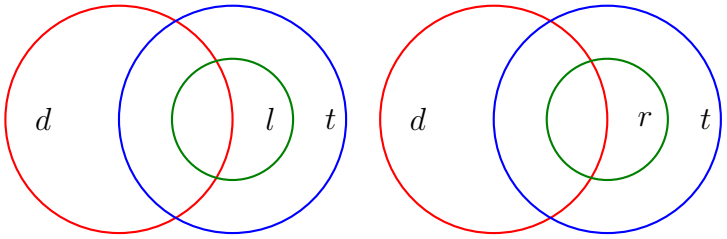
Challenge included two items: (a) my teacher pushes us to think hard about things we do; (b) in this class, my teacher accepts nothing less than our full effort.

Control included three items: (a) my classmates behave the way my teacher wants them to; (b) our class stays busy and does not waste time; and (c) everybody knows what they should be doing and learning in this class.

Appendix C: Study Designs

C.1 The $d \times (l{:}\,t)$ and $d \times (r{:}\,t)$ designs

In G-theory, the $d \times (l{:}\,t)$ and $d \times (r{:}\,t)$ designs are represented by Venn diagrams as follows:



$d \times (l : t)$ $\quad$ $d \times (r : t)$

This is a graphical representation of the study designs described in sections 3.2.1 and 3.2.2. In both cases, the circle for d intersects with everything else to indicate that domains are crossed with teachers and lessons (in the first case) or raters (in the second case). The circles for l and r are inside those for t to indicate that lessons and raters and nested within teachers.

Practically, what this means is that our datasets for each study look as follows:

Table C.1: Data segment for $d \times (l{:}\,t)$ design

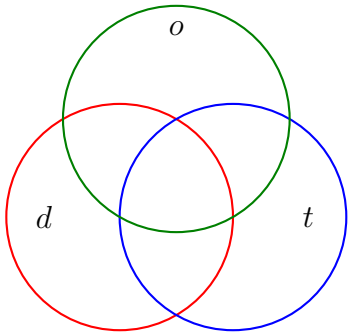
	Teacher 1		Teacher 2		Teacher 3	
	Lesson 1	Lesson 2	Lesson 1	Lesson 2	Lesson 1	Lesson 2
Domain 1	X	X	X	X	X	X
Domain 2	X	X	X	X	X	X
Domain 3	X	X	X	X	X	X
Domain 4	X	X	X	X	X	X
Domain 5	X	X	X	X	X	X
Domain 6	X	X	X	X	X	X

Table C.2: Data segment for $d \times (r{:}\,t)$ design

	Teacher 1		Teacher 2		Teacher 3	
	Rater 1	Rater 2	Rater 1	Rater 2	Rater 1	Rater 2
Domain 1	X	X	X	X	X	X
Domain 2	X	X	X	X	X	X
Domain 3	X	X	X	X	X	X
Domain 4	X	X	X	X	X	X
Domain 5	X	X	X	X	X	X
Domain 6	X	X	X	X	X	X

C.2 The $t \times d \times o$ design

In G-theory, the $t \times d \times o$ design is represented by the following Venn diagram:



In this case, the circles for t, d, and o intersect with each other to indicate that this is a fully crossed design: all teachers are scored on all domains and occasions.

This means that our dataset for this study look as follows:

Table C.2: Data segment for $t \times d \times o$ design

	Teacher 1		Teacher 2		Teacher 3	
	Occasion 1	Occasion 2	Occasion 1	Occasion 2	Occasion 1	Occasion 2
Domain 1	X	X	X	X	X	X
Domain 2	X	X	X	X	X	X
Domain 3	X	X	X	X	X	X
Domain 4	X	X	X	X	X	X
Domain 5	X	X	X	X	X	X
Domain 6	X	X	X	X	X	X

References:

Briggs, D. C., & Alzen, J. L. (2019). Making inferences about teacher observation scores over time. *Educational and Psychological Measurement*, *79*(4), 636-664.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K., & Pianta, R. C. (2013). Effect of observation mode on measures of secondary mathematics teaching. *Educational and Psychological Measurement*, *75*(3), 757-783.

Ferguson, R. F. (2012). Can student surveys measure teaching quality? *Phi Delta Kappan*, *94*(3), 24-28.

Hill, H. C., Charalambous, C. Y., & Kraft, M. A. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher*, *41*(2), 56-64.

Ho, A. D., & Kane, T. J. (2013). *The reliability of classroom observations by school personnel*. Bill & Melinda Gates Foundation.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teachers: Combining high-quality observations with student surveys and achievement gains*.

Luna-Bazaldua, D., Molina, E., & Pushparatnam, A. (2021). A generalizability study of Teach, a classroom observation tool. In M. Wilberg, D. Molenaar, J. González, U. Böckenholt, & J.-S. Kim (Eds.), *Quantitative psychology*. Springer.

Mantzicopoulos, P., French, B. F., & Patrick, H. (2018). The Mathematical Quality of Instruction (MQI) in kindergarten: An evaluation of the stability of the MQI using generalizability theory. *Early Education and Development*, *29*(6), 893-908.

Mantzicopoulos, P., French, B. F., Patrick, H., Watson, J. S., & Ahn, I. (2018). The stability of kindergarten teachers' effectiveness: A generalizability study comparing the Framework

for Teaching and the Classroom Assessment Scoring System. *Educational Assessment*, *23*(1), 24-46.

Mashburn, A. J., Downer, J. T., Rivers, S. E., Brackett, M. A., & Martinez, A. (2014). Improving the power of an efficacy study of a social and emotional learning program: Application of generalizability theory to the measurement of classroom-level outcomes. *Prevention Science 15*, 146-155.

Mashburn, A. J., Meyer, J. P., Allen, J. P., & Pianta, R. C. (2014). The effect of observation length and presentation order on the reliability and validity of an observational measure of teaching quality. *Educational and Psychological Measurement*, *74*(3), 400–422.

Meyer, J. P., Cash, A. H., & Mashburn, A. J. (2011). Occasions and the reliability of classroom observations: Alternative conceptualizations and methods of analysis. *Educational Assessment*, *16*(4), 227-243.

Patrick, H., French, B. F., & Mantzicopoulos, P. (2020). The Reliability of Framework for Teaching Scores in Kindergarten. *Journal of Psychoeducational Assessment*, *38*(7), 831-845.

Praetorius, A.-K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction*, *31*, 2–12.

Schweig, J. D. (2014). Quantifying error in survey measures of school and classroom environments. *Applied Measurement in Education*, *27*(2), 133-157.

Semmelroth, C. L., & Johnson, E. (2014). Measuring rater reliability on a special education observation tool. *Assessment for Effective Intervention*, *39*(3), 131-145.

van der Lans, R. M., van de Grift, W. J. C. M., van Veen, K., & Faokkens-Bruinsma, M. (2016). Once Is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation*, *50*, 88–95.