The Reliability of Classroom Observations and Student Surveys in Non-Research Settings: Evidence from Argentina[*]

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

Andrew D. Ho[‡]
Harvard University

Alejandra Campos Quintero[§]
Columbia University

October 14, 2025

Abstract

There is a growing consensus on the need to measure teaching effectiveness using multiple instruments. Yet, guidance on how to achieve reliable ratings derives largely from formal research in high-income countries. We study the reliability of classroom observations and student surveys conducted by practitioners in a middle-income country. Both instruments can achieve relatively high reliability (0.6–0.8 on a 0–1 scale) when averaged across raters and occasions, but the reliability of observations varies widely (from 0.4 to 0.8) based mostly on how raters are assigned to teachers. We use Generalizability Theory to estimate how reliability improves by increasing the number of times teachers are observed or the number of respondents to surveys. We recommend that practitioners design their teacher feedback systems based on analyses of their own data, instead of assuming that instruments and rubrics will generate scores with the same reliability as research settings.

Keywords: reliability, generalizability, teaching effectiveness, classroom observations, student surveys, Argentina.

[*] 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.

[†] 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.

[‡] Charles William Eliot Professor of Education, Harvard Graduate School of Education. andrew_ho@gse.harvard.edu.

[§] Doctoral Student, Teachers College at Columbia University. aac2271@tc.columbia.edu.

1 Introduction

In the past two decades, policy-makers 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 have 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 (Nye et al., 2004; Rockoff, 2004; Rivkin et al., 2005; Aaronson et al., 2007; Kane et al., 2008; Chetty et al., 2014; Koedel et al., 2015). This evidence then prompted efforts to try to identify effective teachers to inform hiring, retention, training, and pay (e.g., Rockoff et al., 2011; Goldhaber et al., 2017; Dee and Wyckoff, 2015).

This has resulted in growing consensus on the need to use multiple measures of effective teaching. The statistical methods used to estimate teachers' influence on student achievement 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; Kraft, 2019; Jackson, 2020), and neglecting how school-level factors (e.g., principals, counselors, and peers) mediate teachers' capacity to help students (Jackson and Bruegmann, 2009; Johnson et al., 2012; Jackson, 2013; Papay et al., 2020; Mulhern, 2023).

In recent years, several studies demonstrated that other measures of teaching quality (e.g., classroom observations and student surveys) add valuable information not captured by tests (Kane and Staiger, 2011, 2012; Kane et al., 2011, 2013, 2014). Many of these studies 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 and Kane, 2013). This evidence has been cited in the design of teacher feedback systems in the U.S. and abroad.

It is not clear, however, why one should expect that data collected for research purposes should produce results that are indicative of those that would be obtained in non-research settings. In most studies, teachers volunteer to participate, the individuals who rate them are trained, and there are many mechanisms in place to ensure the integrity of the information being gathered. By contrast, governments and non-profits have to devise solutions that work for all teachers and they may face constraints on training or their capacity to adopt quality-assurance checks. These differences could render the non-test metrics administered in research much more reliable than those in practice settings. If this were the case, practitioners could be making decisions about how to design their teacher feedback systems based on non-test measures that are more reliable than the ones they collect, resulting in teaching effectiveness metrics that are less reliable than they intended and realize.

This question is particularly pressing for practitioners in low- and middle-income countries, given that the vast majority of prior studies in this literature have been conducted in the U.S.

Perhaps, in the U.S., teachers are more used to receiving feedback and that their colleagues and students are more used to acting as raters than those in the rest of the world. The U.S. has also invested more funds on developing and researching teacher feedback systems. These differences could make metrics in the U.S. more reliable than those in other countries.

In the present study, we aim to shed light on both questions by estimating the reliability of classroom observations and student surveys administered as part of an education program (i.e., not for research purposes) and comparing our results to those of the relevant literature. We examine the reliability of classroom observations and student surveys of 100 participants in an alternative pathway into teaching in Argentina. These teachers were scored at two time points. The first was during two weeks of practice teaching, shortly after they completed a brief pre-service training course. The second was during the school year, once they began teaching in hard-to-staff schools for two years. The program (*Enseñá por Argentina* or ExA) developed these measures by drawing on existing instruments and administered the ratings exclusively for feedback purposes (i.e., there were no stakes attached to them). This setup allows us to understand the reliability of non-test measures of teaching effectiveness as they are frequently used by organizations in the education sector in an understudied setting. Further, given that ExA is part of a global network of 60 organizations using similar measures and procedures (Teach for All), we see our study as potentially relevant to this broader group.

Our study 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" (Lord and Novick, 1968; Nunnally and Bernstein, 1978; Allen and Yen, 1979), is increasingly used in teacher feedback systems in the U.S. (Hill et al., 2012; Kane and Staiger, 2012; Ho and Kane, 2013). 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 can reach high levels of 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. During clinical practice, the generalizability coefficient for relative error—a measure of reliability for relative distinctions that ranges from 0 (perfectly unreliable) to 1 (perfectly reliable)—was as high as 0.79 on some years, and the coefficient for absolute error—a metric for absolute judgments that also ranges from 0 to 1—reached 0.76. These figures indicate that almost 80% of the variation in observation scores reflects actual differences in measured teaching effectiveness, as opposed to measurement error, which is encouraging. Based on our review of prior G-studies, observations conducted for research settings have average reliabilities of 0.65 in pre-primary education, 0.33 in primary education, and 0.51 in secondary education, with the highest values reaching 0.94, 0.64, and 0.94, respectively (see next section and Appendix A).

Second, the reliability of these observations varies widely depending on their context, the way in which raters are assigned to lessons, and the year in which they are conducted. During the school year, the generalizability coefficients for relative and absolute error reached lower levels (0.66 and 0.57, respectively) than in clinical practice (reported above). And even within clinical practice, reliability varied based on who acted as raters (coaches or peers) and how they were assigned (whether a teacher was observed by the same person or different people). Observations scored by coaches were not always more reliable than those rated by peers. Rather, when a teacher was observed multiple times by the same person, peers were more reliable (with coefficients for relative and absolute error of 0.79 and about 0.75, respectively) than coaches (with coefficients of 0.53-0.55 and 0.38-0.47). Yet, when each lesson delivered by a teacher was observed by a different peer, reliability was both lowest and most variable from one year to the next (with coefficients of 0.44 and 0.41 in 2014 and of 0.64 and 0.61 in 2015). These figures indicate that, depending on how observations are conducted, the variation in scores reflecting differences in effectiveness can be as low as 41% (almost half the figure above).

Third, it is possible to improve the reliability of observations by increasing the number of times teachers are scored. During clinical practice, observing each teacher three times instead of two would improve the generalizability coefficient for relative error by 5-10 percentage points (pp.) and the one for absolute error by 4-9 pp., depending on the rater type (coaches or peers) and whether all lessons delivered by a teacher are observed by the same or a different rater. Adding an observation during the school year would improve the coefficient for relative error by 8 pp. and the one for absolute error by 7-8 pp., depending on the year used as reference. Further increases in the number of observations would improve reliability by a small margin.

Fourth, student surveys administered by practitioners can also reach high reliability levels. During clinical practice, the generalizability coefficient for relative error was as high as 0.76 on some years, and the one for absolute error reached 0.65. In the school year, the corresponding figures were 0.62 and 0.63, respectively. Despite concerns about the reliability of student surveys (see English et al., 2015, for a review), these results indicate that somewhere between 60 and 70% of variation in observed scores reflects actual differences in measured teaching effectiveness. Notably, this reliability was achieved surveying only a sample of 10 students each time, and it varied relatively little between clinical practice and the school year.

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 five more students would improve the generalizability coefficient for relative error by 9 pp. and the one for absolute error by 8 pp. Adding five students during the school year would improve the coefficients for relative and absolute error by 5-9 pp. and 4-8 pp.

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

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 (Lord and Novick, 1968; Nunnally and Bernstein, 1978; Allen and Yen, 1979). 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). 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" (Cronbach et al., 1972; Brennan, 2001), 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 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 and Staiger, 2012) and subsequently identified multiple approaches to reliable observations (MET Project, 2013). Many policy-makers and practitioners around the world have relied on these guidelines when designing their own systems (e.g., Pouezevara et al., 2016; Cruz-Aguayo et al., 2020).

Much of what we know about the reliability of alternative metrics of teaching effectiveness, however, stems from a relatively small set of measures and contexts. We searched for G-studies of classroom observations and student surveys from pre-primary to secondary education in both low-/middle-income and high-income countries. We did not find any studies of student surveys, but we found 12 studies of classroom observations (see Table A.1 in Appendix A). Most focused on three instruments: the Classroom Assessment Scoring System (CLASS, Mashburn et al., 2008; Pianta et al., 2008; Hamre et al., 2013); Framework for Teaching (FfT, Danielson, 2011); and Mathematical Quality of Instruction (MQI, Hill et al., 2011, 2012). Three-fourths were set in the United States and all of them were in high-income countries. These patterns raise question about the external validity of the guidance from these G-studies.

The classroom observations in these studies were conducted for research purposes and they incorporate 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 and Mesalles, 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 and Mesalles, 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. ExA provides teachers with reports on both measures to help them improve their instruction. 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]).

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. 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 × 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 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.

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). We randomly sampled 10 student surveys per teacher to keep the number of raters constant. 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.

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 and report the average result across raters.

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 randomly sampled 10 students per occasion to keep the number of raters constant. We run a separate analysis per occasion—instead of crossing occasions with everything else—because surveys were anonymous, so we cannot ensure that the 10 students that we sampled 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

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, Mashburn et al., 2008; Pianta et al., 2008; Hamre et al., 2013); the Framework for Teaching (FfT, Danielson, 2011); Teaching As Leadership (TAL, Farr, 2010); the Protocol for Language Arts Teaching Observation (PLATO, Grossman et al., 2013, 2015); and Mathematical Quality of Instruction (MQI, Hill et al., 2011, 2012). 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 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, bar graphs of the domain-level ratings, and tables with correlations among them in Appendix A. We describe the domains and provide translated example items for the protocol, and link to the full protocol, in Appendix B.

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, and the descriptions of domains and example items in Appendix B.

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 sections 3.2.1 and 3.2.2, 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.

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:

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

or

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

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

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

or

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

where $\hat{\sigma}^2_{l:t}$ and $\hat{\sigma}^2_{r:t}$ are the estimated variances from lessons and raters, nested within teachers; $\hat{\sigma}^2_{dt}$ is the variance from the domain-by-teacher interaction; $\hat{\sigma}^2_{dl:t,e}$ or $\hat{\sigma}^2_{dr:t,e}$ 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 $\hat{\sigma}^2_{\Delta}$ (i.e., variation in scores from the facets of error that affect not only rankings, but also teachers' locations on the score scale) as:

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

where $\hat{\sigma}^2_d$ 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:

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

and

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

where $\hat{\sigma}^2_t$ 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:

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

where $\hat{\sigma}^2_{to}$ and $\hat{\sigma}^2_{tdo,e}$ 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:

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

where $\hat{\sigma}^2_d$, $\hat{\sigma}^2_o$, and $\hat{\sigma}^2_{do}$ 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 $d \times (l : t)$ and $d \times (r : t)$ 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:

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

or

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

and also in the calculation of absolute error variance:

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

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:

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

and absolute error variance:

$$\hat{\sigma}^2_\Delta = \frac{\hat{\sigma}^2_d}{n_d} + \frac{\hat{\sigma}^2_o}{n'_o} + \frac{\hat{\sigma}^2_{do}}{n_d n'_o} + \hat{\sigma}^2_\delta, \tag{15}$$

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 can reach high levels of 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 0.53 to 0.79 and those for absolute error (on the second-to-last row) ranged from 0.38 to 0.76. The fact that the former are slightly larger than the latter should not be surprising, as relative error includes sources of variability that only change teachers' relative standing, whereas absolute error includes those that change both their relative and absolute positions (as section 4.1 shows, the calculation of absolute error variance includes, and is thus always equal to or larger than, relative error). The highest values in both sets of coefficients indicate that these observations could be used to both identify which teachers are in greatest need of support and to assign all teachers scoring below a threshold to an intervention (e.g., resources, training, coaching).

On average, observations conducted during clinical practice had similar levels of reliability to those during the school year. The mean generalizability coefficient for relative error for clinical practice was 0.62 and the one for the school year was 0.64. The mean coefficient for absolute error for clinical practice was 0.56 and the one for the school year was 0.47. Yet, the reliability of clinical-practice observations varied more than that of school-year observations. The coefficients for relative and absolute error in clinical practice ranged from 0.44 to 0.79 and from 0.38 to 0.76. Those for the school year ranged from 0.62 to 0.66 and from 0.44 to 0.51. These results suggest that one context was not inherently more conducive for reliability and that there are other factors that explain the variability during clinical practice.

One factor that might explain the similarities clinical practice and the school year and the variability in reliability during clinical practice is the way in which raters were assigned. If we compare observations in which the same coach scored both lessons across clinical practice and the school year, both sets of observations had similar reliability. The generalizability coefficients for relative error ranged from 0.53 to 0.55 during clinical practice and from 0.62 to 0.66 during the school year, and the ones for absolute error ranged from 0.38 to 0.47 during clinical practice and from 0.43 to 0.57 during the school year. Further, within clinical practice, observations in which the same rater (coach or peer) scored both lessons delivered by a teacher had higher reliability than those in which a different person scored each lesson. The coefficients for relative error for the former ranged from 0.53 to 0.79 in the first case and from 0.44 to 0.64 in the second case, and those for absolute error ranged from 0.38 to 0.76 in the first case and from 0.41 to 0.61 in the second. These results indicate that rater assignment matters more than the context in which observations are conducted or even who acts as rater.

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 of relative judgments from observations seems feasible during both clinical practice and the school year. As Figure 1 shows, the generalizability coefficient for relative error in clinical-practice observations is between 0.44 and 0.79 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). Adding a lesson would considerably improve this coefficient by 5-10 pp. (see the y-coordinate of the same lines at 3 lessons). Further increases in the number of lessons would only marginally improve reliability by 3-7 pp., despite making such observations more logistically complex (notice the increasingly flat slopes of these lines beyond 3 lessons). The coefficient for relative error in school-year observations is around 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 in the number of occasions would improve it further by only 5 pp.

Adding lessons or occasions would have a slightly smaller impact on the reliability of absolute judgments from observations. As Figure 1 shows, the generalizability coefficient for absolute error is between 0.38 and 0.76 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 4-9 pp. (see the y-coordinates of the same lines at 4 lessons in panels A, E-F). Further increases in the number of lessons would achieve smaller improvements in reliability by 2-6 pp. The pattern in similar for the school year. The coefficient for absolute error is between 0.43 and 0.57 when each teacher is rated twice (see panels G-H). Adding an occasion would improve this coefficient by 7-8 pp., but further increases would improve it further by only 4-5 pp.

5.2 Student surveys

Student surveys can also reach high levels of reliability. As Table 3 shows, the coefficients for relative error ranged from 0.5 to 0.76 and those for absolute error from 0.37 to 0.65. As in the case of classroom observations, the highest values in both sets of coefficients indicate that surveys can help make relative distinctions between and absolute judgments about teachers based on only 10 students per teacher.

There was relatively little variation in the reliability of surveys across contexts and years. The mean generalizability coefficient for relative error for clinical practice was 0.6 and the one for the school year was 0.63. The mean coefficient for absolute error for clinical practice was 0.58 and the one for the school year was 0.51. These results suggest that the reliability of surveys is stable across contexts and rater assignments. One important caveat, however, is that reliability was lowest in the 2015 school year, 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. As Figure 2 shows, the generalizability coefficient for relative error in clinical- practice surveys is between 0.56 and 0.63 with 10 students (see the y-coordinates of the blue lines at 10 students in panels A-B). Adding 5 students would improve reliability by 9 pp. (see y-coordinates of these lines at 15 students), but adding 5 more would only do so by 7 to 9% (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 0.35 and 0.64 with 10 students. Adding 5 students would increase it by 5-9 pp., and adding 5 more students would only do so by 3-5 pp.

Adding raters would have a similar impact on the reliability of absolute judgments from surveys. As Figure 2 shows, the generalizability coefficient for absolute error is between 0.53 and 0.62 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 8 pp. (see y-coordinates at 15 students), but adding 5 more would only do so by 5 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 0.37 and 0.65 with 10 students. Adding 5 students would increase it by 4-8 pp., but 5 more would only do so by 2-5 pp.

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 focused on a small set of measures and contexts, so we were evaluated whether they are representative of the realities of less established instruments administered in LMICs. We obtained data from an education non-profit and examined the reliability of their metrics.

We found that both classroom observations and student surveys administered in practice settings can achieve high levels of reliability. We believe that this finding is important because it demonstrates that practitioners do not always need to adopt costly quality-assurance mechanisms to produce reliable ratings of teachers. We also found, however, that the reliability of classroom observations varied widely depending on how raters are assigned. We see this as a good reason for practitioners to conduct their own studies of the reliability of their instruments, instead of relying on our estimates for deciding how to design their teacher feedback systems. To support them on this endeavor, we have explained in great detail both how to understand the design of each measurement procedure and how to analyze the data that each produces. We have also made the datasets and code from our analyses available with this paper.

As we illustrated using our own data, G-studies can be helpful to understand not only the current reliability of a measurement procedure, but also the optimal approach to improve it. We showed that simply adding a lesson or occasion in classroom observations or five raters in student surveys achieved meaningful improvements in reliability. Equally importantly, we also demonstrated the diminishing marginal returns of further expansions in the number of lessons, occasions, or raters. We see this approach as particularly helpful for practitioners. First, it enables them to evaluate the tradeoff between the potential improvements in reliability from changes to their existing systems against their cost. Second, it allows them to use their resources as efficiently as possible, changing only what is needed to achieve reliable measures.

In using G-studies to design their teacher feedback systems, it is important for practitioners to 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 and Gilbert, 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.

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. Therefore, we anticipate that the processes and insights from our study will be relevant to these other programs, impacting thousands of teachers every year and the students they serve. Yet, to appropriately 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. We see this shift as akin to the one that has taken place in the impact-evaluation literature, which has transitioned from assessing small-scale programs run by highly capable organizations in a narrow set of contexts to seeking to understand the effect of initiatives when implemented at scale.

References

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

Allen, M. J. and W. M. Yen (1979). Introduction to measurement theory. Prospect Heights, IL: Waveland Press.

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.

Brennan, R. L. (2001). *Generalizability Theory.* New York: Springer-Verlag.

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

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

Cronbach, L. J., G. C. Gleser, H. Nanda, and N. Rajaratnam (1972). *The dependability of behavioral measurements: Theory of generalizability for scores and profiles.* New York: Wiley.

Cruz-Aguayo, Y., D. Hincapié, and C. Rodríguez (2020). Testing our teachers: Keys to a successful teacher evaluation system.

Danielson, C. (2011). *Enhancing professional practice: A framework for teaching.* Association for Supervision and Curriculum Development (ASCD).

Dee, T. S. and J. Wyckoff (2015). Incentives, selection, and teacher performance: Evidence from impact. *Journal of Policy Analysis and Management 34*(2), 267–297.

English, D., J. Burniske, D. Meibaum, and L. Lachlan-Haché (2015). Uncommon measures: Student surveys and their use in measuring teaching effectiveness. Washington, DC: American Institutes for Research (AIR).

Farr, S. (2010). *Teaching as leadership: The highly effective teacher's guide to closing the achievement gap.* San Francisco, CA: Teach for America.

Ferguson, R. F. (2010). Student perceptions of teaching effectiveness. Boston, MA: The National Center for Teacher Effectiveness and the Achievement Gap Initiative.

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

Gage, N. A., H. Han, A. S. MacSuga-Gage, D. Prykanowski, and A. Harvey (2018). *A generalizability study of a direct observation screening tool of teachers' classroom management skills*, Volume Emerging research and issues in behavioral disabilities, pp. 29–50. Emerald Publishing.

Ganimian, A. J. and V. Mesalles (2025). ¿Qué aprendimos de Aprender? Informe sobre el desempeño de las 24 jurisdicciones argentinas en las evaluaciones nacionales. Ciudad Autónoma de Buenos Aires, Argentina: Educar 2050 and Argentinos por la Educación.

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

Grossman, P., J. Cohen, and L. Brown (2015). *Understanding instructional quality in English language arts: Variations in PLATO scores by content and context*, pp. 303–331. Wiley Online Library.

Grossman, P., S. Loeb, J. Cohen, and J. Wyckoff (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

Hamre, B. K., R. C. Pianta, J. T. Downer, J. DeCoster, A. J. Mashburn, S. M. Jones, J. L. Brown, E. Cappella, M. Atkins, S. E. Rivers, M. Atkins, S. E. Rivers, M. A. Brackett, and A. Hamagami (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., C. Y. Charalambous, and M. A. Kraft (2012). 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., L. Kapitula, and K. Umland (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal 48*(3), 794–831.

Himmelsbach, Z. and J. Gilbert (2025). The case for bayesian estimation of the d-study. Presentation at the Measurement Lab. Cambridge, MA: Harvard Graduate School of Education (HGSE).

Ho, A. D. and T. J. Kane (2013). The reliability of classroom observations by school personnel. Seattle, WA: Bill and Melinda Gates Foundation.

INDEC (2024). Encuesta Permanente de Hogares (EPH) total urbano. Evolución de la distribución del ingreso. Tercer trimestre de 2023. (Trabajo e ingresos, Vol. 8, no. 2). Buenos Aires, Argentina: Instituto Nacional de Estadística y Censos (INDEC).

Jackson, C. K. (2013). Match quality, worker productivity, and worker mobility: Direct evidence from teachers. *Review of Economics and Statistics 95*(4), 1096–1116.

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. and E. Bruegmann (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. *Perspectives on Practive.* Seattle, WA: Bill and Melinda Gates Foundation.

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

Johnson, S. M., M. A. Kraft, and J. P. Papay (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., K. Kerr, and R. C. Pianta (2014). *Designing teacher evaluation systems: New guidance from the measures of effective teaching project.* John Wiley & Sons.

Kane, T. J., D. F. McCaffrey, T. Miller, and D. O. Staiger (2013). Have we identified effective teachers? Validating measures of effective teaching using random assignment. *Measures of Effective Teaching Project.* Seattle, WA: Bill and Melinda Gates Foundation.

Kane, T. J., J. E. Rockoff, and D. O. Staiger (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. and D. O. Staiger (2011). Learning about teaching: Initial findings from the measures of effective teaching project. Bill and Melinda Gates Foundation. Seattle, WA.

Kane, T. J. and D. O. Staiger (2012). Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains. Seattle, WA: Bill and Melinda Gates Foundation.

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

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

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

Ley de Educación Nacional (2006). Ley de Educación Nacional No. 26.206. Buenos Aires, Argentina: Honorable Congreso de la Nación Argentina.

Lord, F. M. and M. R. Novick (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mantzicopoulos, P., B. F. French, and H. Patrick (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., B. F. French, H. Patrick, J. S. Watson, and I. Ahn (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., J. T. Downer, S. E. Rivers, M. A. Brackett, and A. Martinez (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., J. P. Meyer, J. P. Allen, and R. C. Pianta (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.

Mashburn, A. J., R. C. Pianta, B. K. Hamre, J. T. Downer, O. A. Barbarin, D. Bryant, M. Burchinal, D. M. Early, and C. Howes (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. *Policy and Practice Brief*. Seattle, WA: Bill and Melinda Gates Foundation.

MdCH (2024). Anuario estadístico 2023. Buenos Aires, Argentina: Secretaría de Educación, Ministerio de Capital Humano.

MET Project (2010). Validation engine for observational protocols. Seattle, WA: 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. *Policy and Practice Brief*. Seattle, WA: Bill and Melinda Gates Foundation.

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

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

Nunnally, J. C. and I. H. Bernstein (1978). Psychometric theory (2nd edition). New York: McGraw-Hill.

Nye, B., S. Konstantopoulos, and L. V. Hedges (2004). How large are teacher effects? *Educational evaluation and policy analysis 26*(3), 237–257.

OECD (2023). PISA 2022 results (Vol. I): The state of learning and equity in education. Paris, France: 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., E. S. Taylor, J. H. Tyler, and M. E. Laski (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., B. F. French, and P. Mantzicopoulos (2020). The reliability of framework for teaching scores in kindergarten. *Journal of Psychoeducational Assessment 38*(7), 831–845.

Pianta, R. C., K. M. La Paro, B. K. Hamre, and P. H. (2008). Classroom assessment scoring system (class) manual: K-3. Baltimore, MD: Paul Brookes Publishing Co.

Pouezevara, S., A. Pflepsen, L. Nordstrum, S. King, and A. Gove (2016). Measures of quality through classroom observation for the sustainable development goals: Lessons from low- and middle-income countries. Paris, France: United Nations Educational, Scientific, and Cultural Organization (UNESCO).

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

Rivkin, S. G., E. A. Hanushek, and J. F. Kain (2005). Teachers, schools, and academic achievement. *Econometrica*, 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., B. A. Jacob, T. J. Kane, and D. O. Staiger (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.

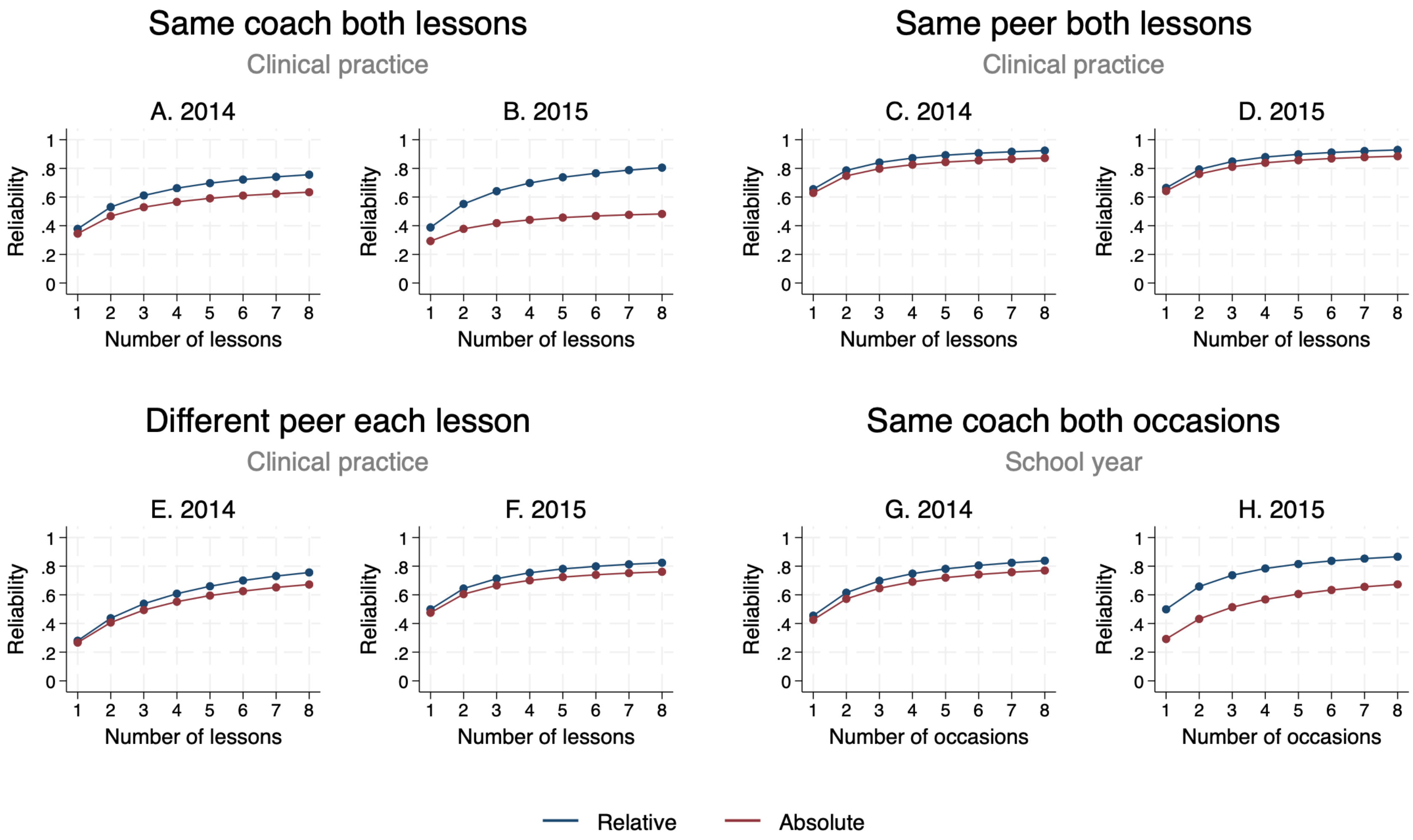
van der Lans, R. M., W. J. C. M. van de Grift, K. van Veen, and M. Fokkens-Bruinsma (2016). Once is not enough: Establishing reliability criteria for feedback and evaluation decisions based on classroom observations. *Studies in Educational Evaluation 50*, 88–95.

World Bank (2024a). DataBank - Education Statistics. Retrieved from `https://databank.worldbank.org/`.

World Bank (2024b). Education statistics (Edstats). `https://datatopics.worldbank.org/education/` Retrieved: June 8, 2023.

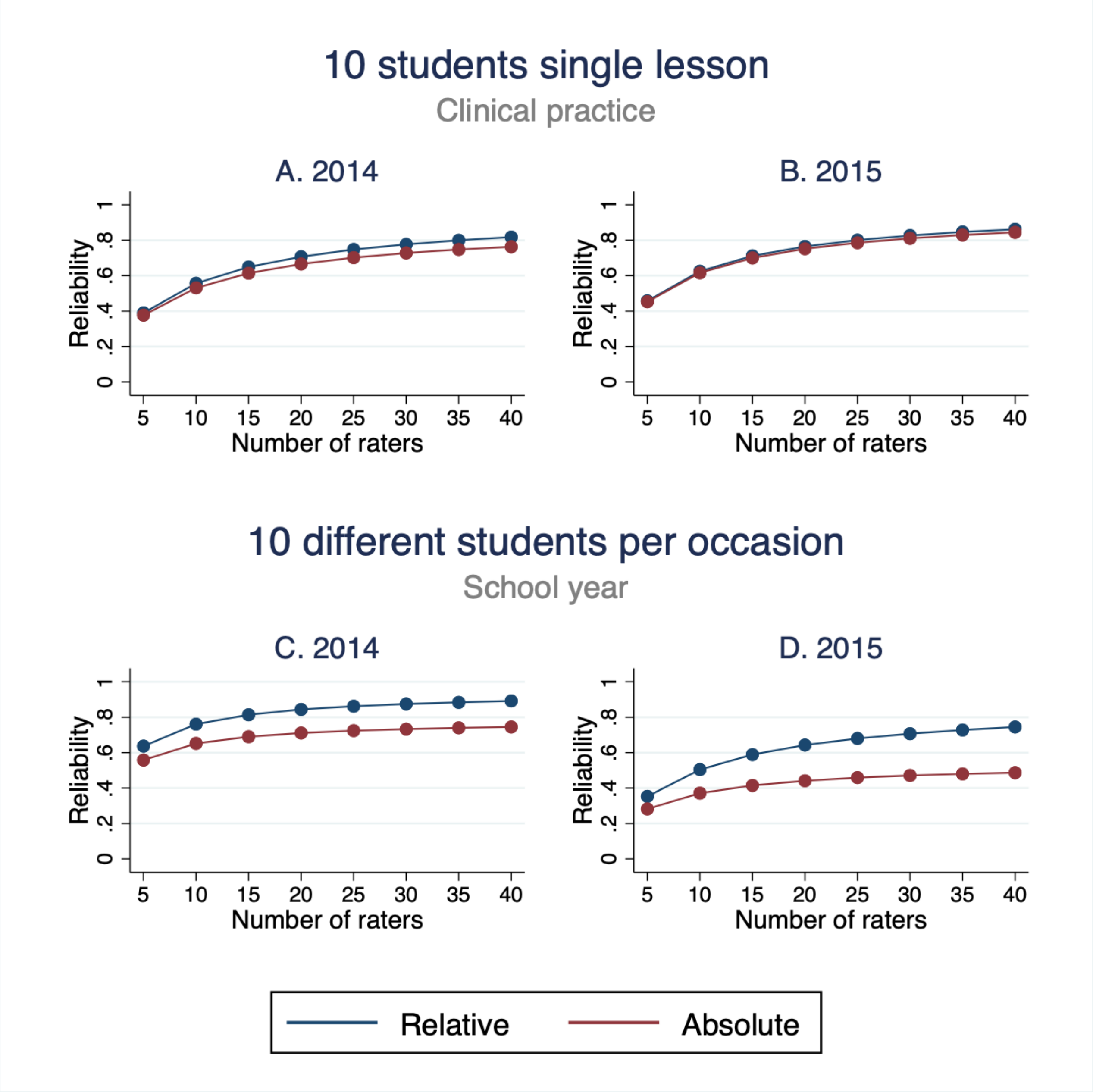
World Bank (2024c). Poverty and inequality platform: Argentina country profile. Washington, DC: The World Bank. `https://pip.worldbank.org/country-profiles/ARG`.

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

Context	Year	Teachers	Lessons/ occasions	Raters per lesson/occasion	Domains	Study design	Mean	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	10 students single lesson	7	$d \times (r : t)$	4.47	0.75
	2015	31	1	10 students single lesson	7	$d \times (r : t)$	4.44	0.88
School year	2014	33	2	10 different students per occasion	7	$2[d \times (r : t)]$	3.75	0.94
	2015	28	2	10 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).

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	.075	.049	.258	.226	.093	.131	.11	.076
Domain	.117	.245	.1	.075	.093	.079	.065	.08
Lesson : Teacher	.089	.046	.097	.084	.201	.087		
Occasion							0	.09
Domain × Teacher	.06	.015	.03	.024	0	.079	.033	.015
Occasion × Teacher							.088	.032
Domain × Occasion							.032	.026
Residual	.145	.172	.203	.188	.232	.188	.231	.251
SD of teacher effect	.274	.221	.508	.475	.305	.362	.332	.276
SEM of a single observation	.258	.2	.265	.248	.346	.269	.262	.199
Reliability of a single observation								
Relative standing of teachers	.53	.55	.79	.79	.44	.64	.62	.66
Absolute scores of teachers	.47	.38	.75	.75	.41	.61	.57	.43
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. 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	
	10 students single lesson		10 different students per occasion	
	2014	2015	2014	2015
Variance component	Var.	Var.	Var.	Var.
Teacher	.025	.116	.113	.04
Domain	.017	.012	.166	.156
Rater : Teacher	.223	.458	.232	.218
Domain × Teacher	.014	.009	.036	.034
Residual	.265	.246	.409	.37
SD of teacher effect	.158	.341	.336	.2
SEM of a single observation	.168	.225	.185	.179
Reliability of a single observation				
Relative standing of teachers	.47	.7	.77	.56
Absolute scores of teachers	.45	.69	.66	.42
Number of teachers	23	31	33	25

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. 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

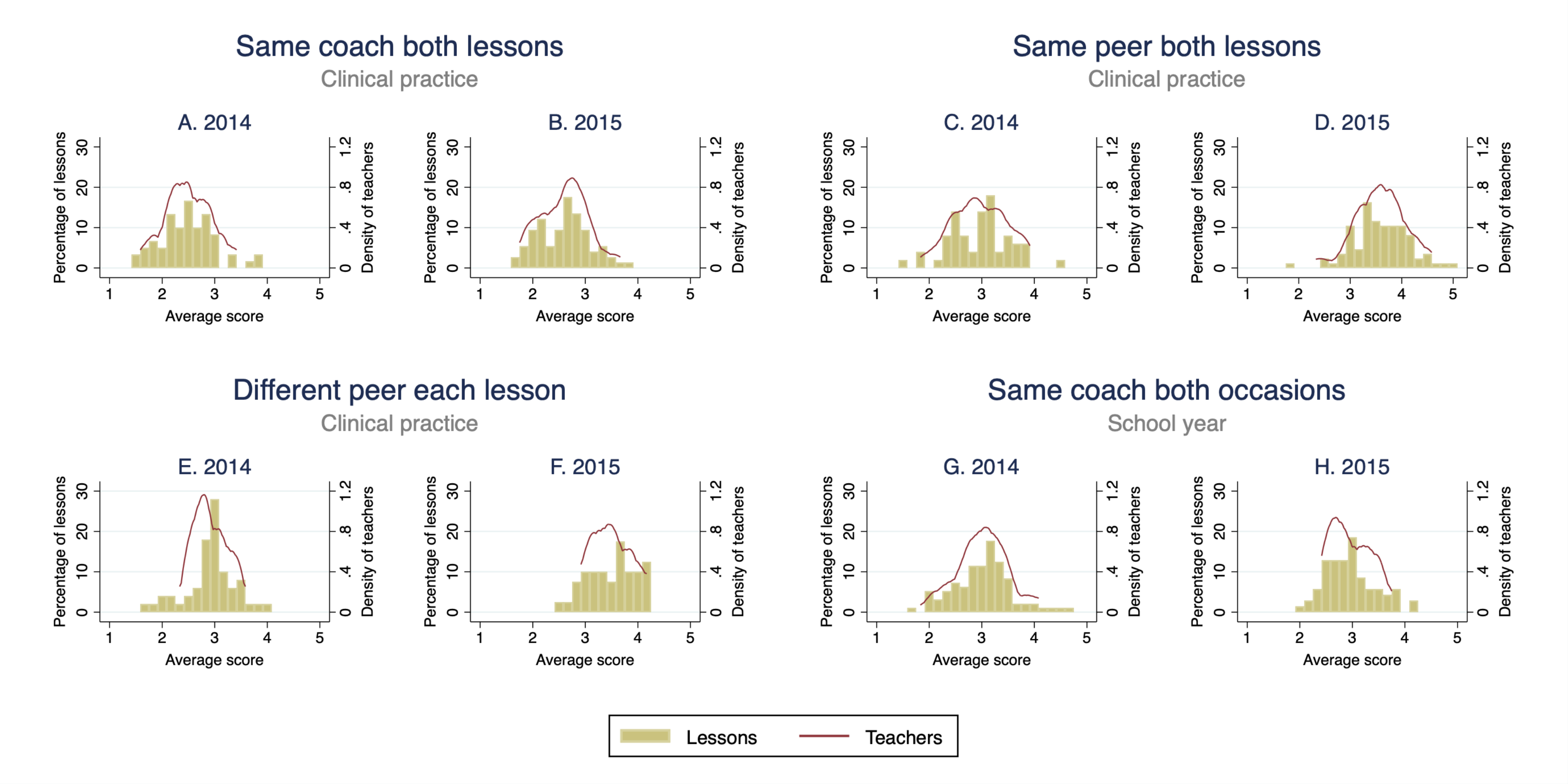
Table A.1: Generalizability studies of classroom observations

Study	Context	Instrument	Teachers	Obs. per teacher	Mean score	SD of teacher effect	SEM of single obs.	Reliability of single obs.
A. Pre-primary								
Mantzicopoulos et al. (2018)	Midwestern U.S.	CLASS K-3-EMSUP	10	4	4.75/7	0.43	0.23	0.78
		CLASS K-3-CLORG			5.19/7	0.23	0.24	0.47
		CLASS K-3-INSUP			3.04/7	0.41	0.34	0.61
		FfT Classroom environment			2.36/4	0.26	0.18	0.68
		FfT Classroom instruction			1.87/4	0.19	0.22	0.44
Mantzicopoulos et al. (2018)	Midwestern U.S.	MQI-R	20	5		0.17	0.10	0.80
		MQI-WWSM				0.13	0.10	0.60
		MQI-EI				0.00	0.10	0.10
		MQI-CCASP				0.09	0.10	0.58
		MQI-CWCM				0.13	0.14	0.82
		Whole lesson				0.29	0.14	0.70
Patrick et al. (2020)	Indiana, IN	FfT Reading	20	10	2.47/4	0.37	0.09	0.94
		FfT Math			2.37/4	0.35	0.10	0.93
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
Gage et al. (2018)	Southeastern U.S.	CMS Praise	11		0.28/1	0.13	0.08	0.19
		CMS BSP			0.61/1	0.45	0.25	0.45
		CMS OTR			2.57/7	0.32	0.42	0.20
		CMS PE			0.26/1	0.10	0.05	0.19
C. Secondary								
Hill et al. (2012)	Southwestern U.S.	MQI-R	24	1				0.45
		MQI-EI						0.37
		MQI-SPMMR						0.46
Kane and Staiger (2012)	Charlotte-Mecklenburg, NC;	FfT	1333	4		0.29	0.38	0.37
	Dallas, TX; Denver, CO;	CLASS						0.31
	Hillsborough Co., FL;	PLATO						0.34
	New York City, NY;	MQI						0.14
	Memphis, TN	UTOP	1000					0.30
Mashburn et al. (2014)	Southeastern U.S.	CLASS-EMSUP	47	3	4.11/7	0.46	0.32	0.67
		CLASS-INSUP			3.21/7	0.43	0.39	0.54

		CLASS-CLORG			5.18/7	0.65	0.34	0.78
Mashburn et al. (2014)	Brooklyn and Queens, NY	CLASS-EMSUP	48	6		0.48		0.48
		CLASS-INSUP				0.49		0.51
		CLASS-CLORG				0.56		0.44
Praetorius et al. (2014)	Germany and Switzerland	CLASS Classroom management	38	5	3.64/7	0.10		0.92
		Personal learning support			2.60/7	0.00		0.94
		Cognitive activation			1.93/7	0.02		0.63
D. Multiple levels								
Ho and Kane (2013)	Hillsborough Co., FL	FfT	67	46	2.58/4	0.27	0.34	0.39
van der Lans et al. (2016)	Netherlands	ICALT3	69	3		1.14		0.51

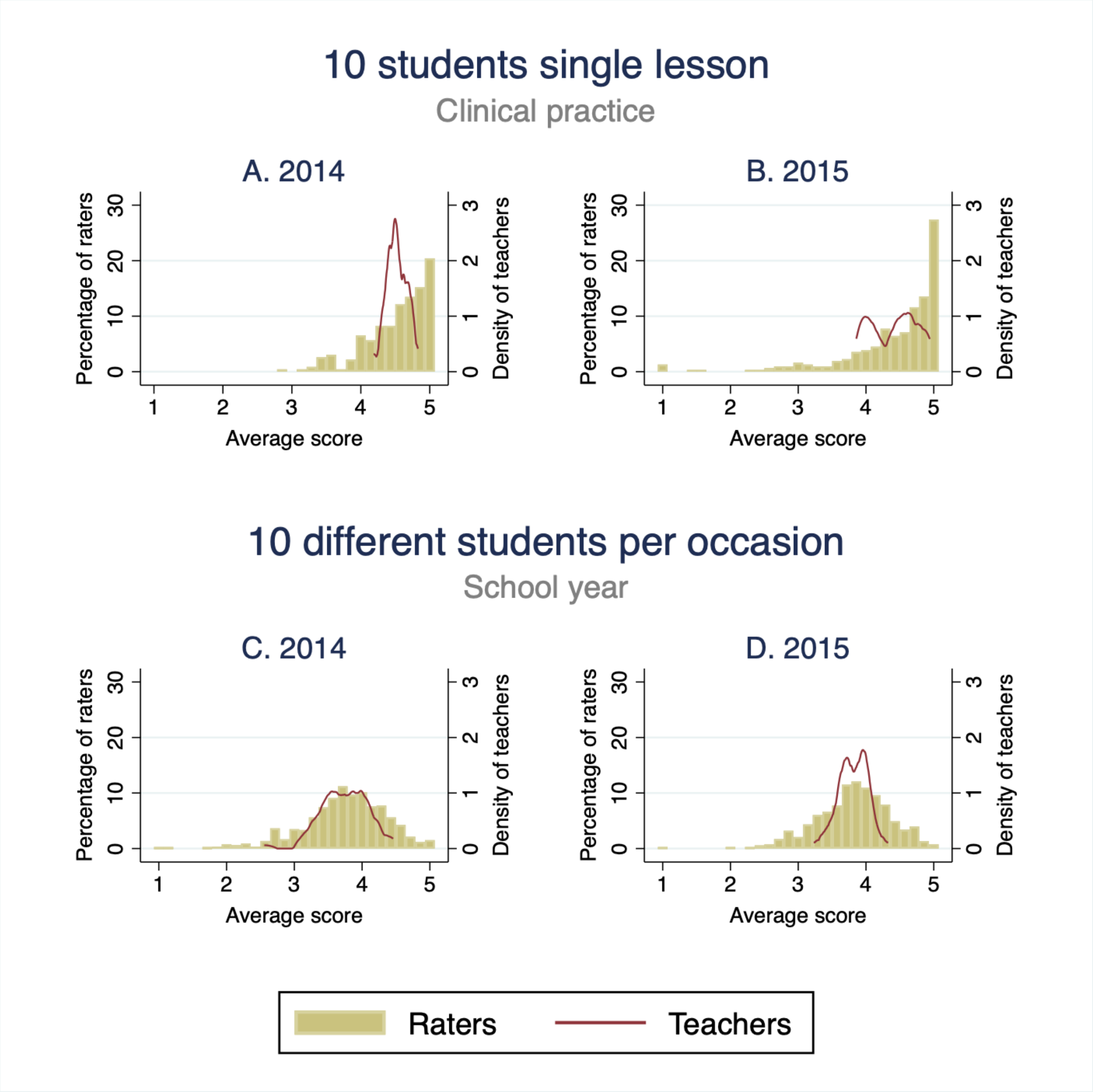
Notes: This table provides an overview of prior studies on the reliability of classroom observations and student surveys. The standard deviation (SD) of the teacher effect is the square root of true-score variance. The standard error of measurement (SEM) of a single observation is the square root of relative-error variance. 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, EI, CCASP, WWSM, CWCM, and SPMMR are its domains: Richness, Errors and Imprecision, Common Core-aligned Student Practices, Working with Student and Mathematics, Classroom Work is Connected to Mathematics, and Student Participation in Meaning Making and Reasoning. 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. NSSE stands for The National Survey of Student Engagement. Mantzicopoulos et al. (2018) reports reliability coefficient for five observations. In Ho and Kane (2013), each teacher was observed on average 46 observations per teacher by different observers and lessons.

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



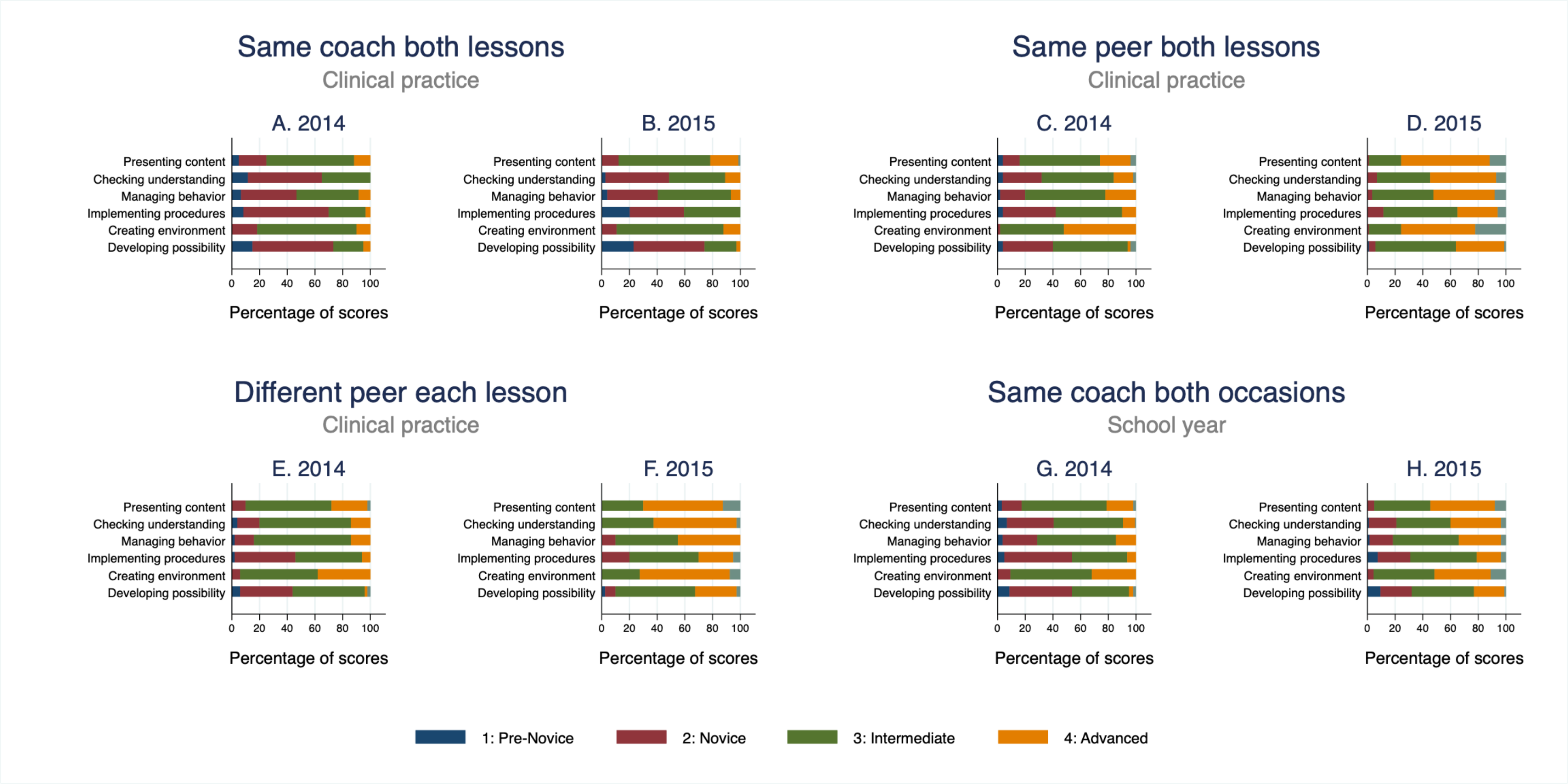
Notes: This figure shows the distribution of lesson- or occasion-level (histogram) and teacher-level (kernel plot) average scores on classroom observations of ExA teachers during clinical practice and the school year in 2014 and 2015.

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



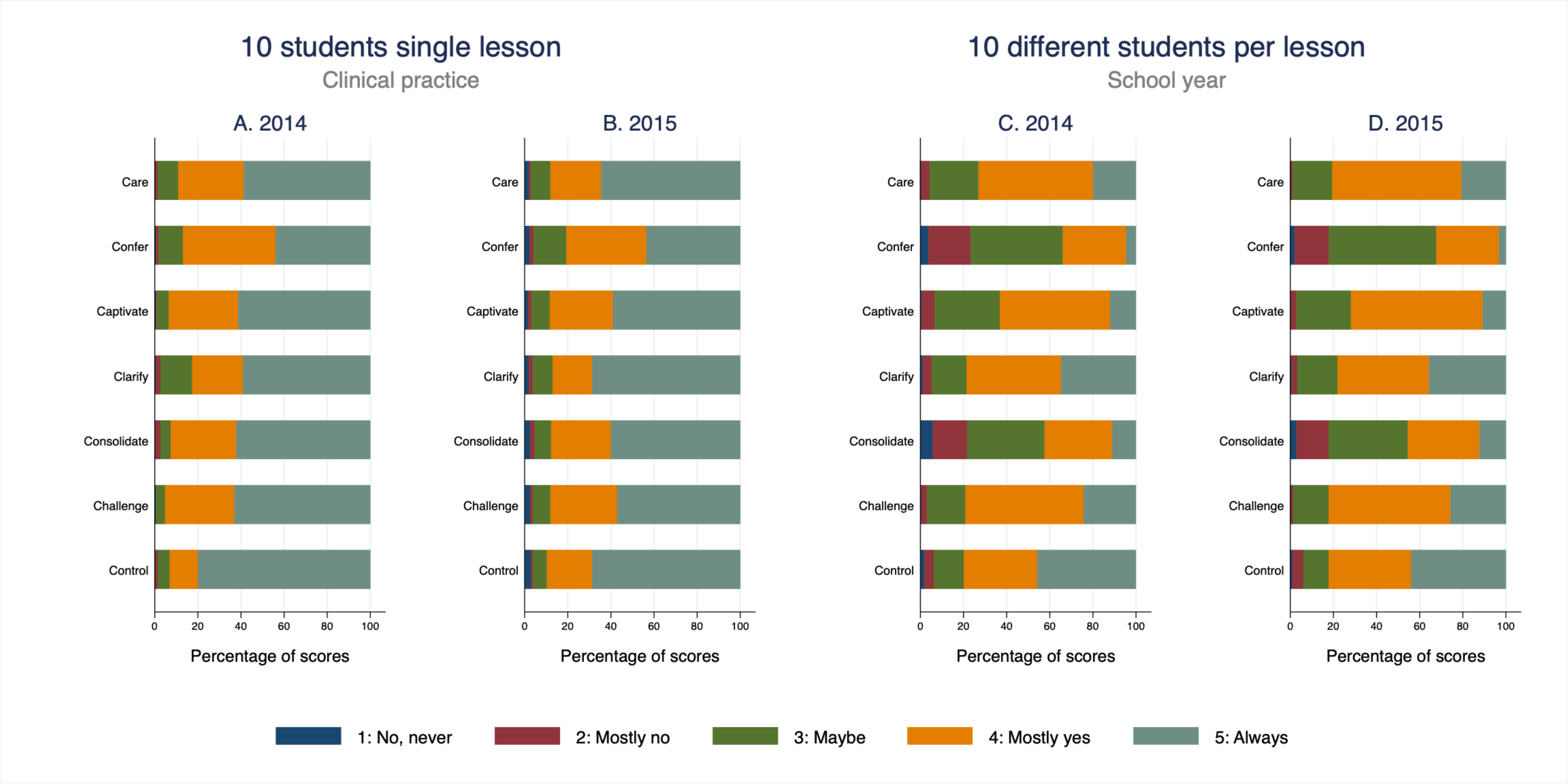
Notes: This figure shows the distribution of rater-level (histogram) and teacher-level (kernel plot) average scores on student surveys of ExA teachers during clinical practice and the school year in 2014 and 2015.

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

	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)

	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)

	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	0.72***	1.00												
Captivate	0.71***	0.88***	1.00											
Clarify	0.45	0.17	0.29	1.00										
Consolidate	0.21	-0.05	0.04	-0.06	1.00									
Challenge	0.64**	0.69***	0.73***	0.10	0.38	1.00								
Control	0.40	0.60**	0.69***	-0.08	0.00	0.59**	1.00							
B. School year														
Care	0.39	0.46	0.70***	0.22	0.08	0.48*	0.58**	1.00						
Confer	0.39	0.51*	0.72***	0.38	-0.21	0.45	0.66**	0.75***	1.00					
Captivate	0.28	0.59**	0.73***	0.21	-0.08	0.41	0.65**	0.84***	0.82***	1.00				
Clarify	0.44	0.66**	0.82***	0.11	0.10	0.56**	0.66**	0.92***	0.77***	0.91***	1.00			
Consolidate	0.46	0.57**	0.66**	0.49*	-0.12	0.43	0.33	0.64**	0.74***	0.79***	0.68**	1.00		
Challenge	0.17	0.44	0.66**	0.19	-0.30	0.13	0.63**	0.78***	0.76***	0.88***	0.79***	0.60**	1.00	
Control	0.37	0.64**	0.74***	-0.00	0.16	0.61**	0.69***	0.88***	0.66**	0.90***	0.96***	0.61**	0.71***	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)

	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	0.73	1.00												
Captivate	0.93***	0.78*	1.00											
Clarify	0.74*	0.56	0.88**	1.00										
Consolidate	0.77*	0.49	0.90**	0.96***	1.00									
Challenge	0.81*	0.51	0.88**	0.96***	0.94***	1.00								
Control	0.92**	0.59	0.93***	0.90**	0.91**	0.97***	1.00							
B. School year														
Care	-0.12	-0.14	-0.25	-0.56	-0.37	-0.59	-0.43	1.00						
Confer	-0.22	-0.08	-0.39	-0.78*	-0.69	-0.67	-0.48	0.66	1.00					
Captivate	0.02	-0.12	-0.16	-0.45	-0.27	-0.47	-0.30	0.97***	0.57	1.00				
Clarify	0.36	0.45	0.19	-0.15	-0.08	-0.21	-0.04	0.75*	0.44	0.79*	1.00			
Consolidate	0.03	-0.23	-0.08	-0.30	-0.07	-0.32	-0.19	0.91**	0.39	0.94***	0.60	1.00		
Challenge	-0.23	-0.16	-0.30	-0.56	-0.39	-0.64	-0.51	0.98***	0.59	0.93***	0.70	0.89**	1.00	
Control	-0.71	-0.36	-0.56	-0.53	-0.46	-0.71	-0.76*	0.53	0.22	0.39	0.13	0.43	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? For example, for item a), the descriptions were: 1) pre-novice: no, they make content mistakes in their explanation and answers to student questions; 2) novice: no, their presentation is correct but too elemental and they cannot answer basic questions; 3) intermediate: more or less, their presentation is correct but they cannot answer advanced questions; 4) advanced: yes, their presentation is correct and comprehensive and they can answer most questions; 5) exemplary: yes, their presentation is correct, comprehensive, and they can answer all questions.

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? For example, for item a), the descriptions were: 1) pre-novice: no, the teacher is speaking during the whole lesson; 2) novice: no, the class includes a lecture and an activity, but there are no student-teacher interactions; 3) intermediate: more or less, the teacher asks only a few questions; 4) advanced: yes, they ask questions in several moments of the lesson; and 5) exemplary: yes, they incorporate questions throughout the lesson.

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? For example, for item a), the descriptions were: 1) pre-novice: no, there are no signs in the classroom and the teacher never alludes to rules; 2) novice: no, there are signs in the classroom, but the teacher never refers to them; 3) intermediate: more or less, there are signs, but the teacher refers to them selectively; 4) advanced: yes, there are signs and the teacher refers to them consistently; 5) exemplary: yes, there are signs and the teacher and students refer to them consistently.

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? For example, for item a), the descriptions were: 1) pre-novice: no, there are no signs in the classroom and the teacher never alludes to routines; 2) novice: no, there are signs in the classroom, but the teacher never refers to them; 3) intermediate: more or less, there are signs in the classroom, but the teacher refers to them selectively; 4) advanced: yes, there are signs in the classroom and the teacher refers to them consistently; 5) exemplary: yes, there are signs in the classroom and the teacher and students refer to them consistently.

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? For example, for item a), the descriptions were: 1) pre-novice: no, they are hostile and offensive towards students; 2) novice: no, they are not hostile/offensive, but they make comments in poor taste; 3) intermediate: more or less, they are not hostile/offensive and their comments are not in poor taste, but they treat students unequally; 4) advanced: yes, they are respectful and treat all students equally; 5) exemplary: yes, they are respectful and treat all students equally and they demonstrate a genuine interest in their students' lives.

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? For example, for item a), the descriptions were: 1) pre-novice: no, they do not congratulate students for performing well in activities, homework, or assessments; 2) novice: no, they praise students generally, but do not indicate what students did well or improved on; 3) intermediate: more or less, they comment on students' performance in general terms; 4) advanced: yes, they comment on individual students' performance or improvements; and 5) exemplary: yes, they comment on individual students' performance or improvements pointing to specific evidence.

B.2 Student surveys

ExA adjusted and translated the Tripod survey (Ferguson, 2010, 2012) to provide feedback to teachers during clinical practice and the school year. It covered 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 (setting high standards for students), and control (managing students' behavior). 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.

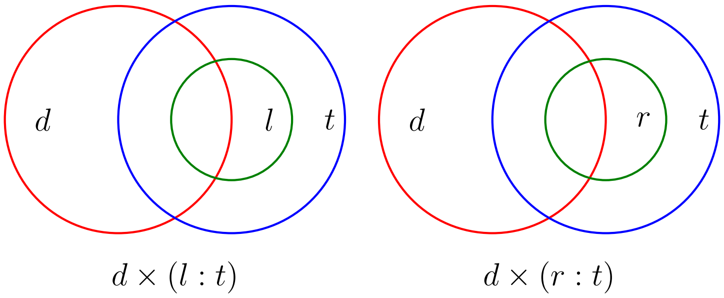
Challenge included dos items: a) my teacher pushes us to think hard about things we do; b) in this class, my teacher accepts nothing less than our full effort.

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:



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

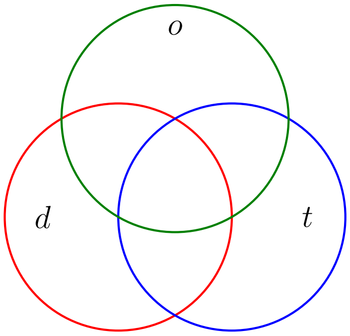
	Teacher 1		Teacher 2		Teacher 3	
	Lesson 1	Lesson 2	Lesson 3	Lesson 4	Lesson 5	Lesson 6
Domain 1	X	X	X	X	X	X
Domain 2	X	X	X	X	X	X
Domain 3	X	X	X	X	X	X
Domain 4	X	X	X	X	X	X
Domain 5	X	X	X	X	X	X
Domain 6	X	X	X	X	X	X

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

	Teacher 1		Teacher 2		Teacher 3	
	Rater 1	Rater 2	Rater 3	Rater 4	Rater 5	Rater 6
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 $t \times d \times o$ designs

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 all domains and occasions.

This means is that our datasets for this study look as follows:

Table C.3: Data segment for $d \times (l : t)$ 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