Teach

GUIDANCE NOTE:
How do we know teacher professional development is working?
Measuring changes in teaching practices in the classroom

Authors: Diego Luna-Bazaldua, Ana Teresa del Toro Mijares, Ezequiel Molina, and Adelle Pushparatnam

WORLD BANK GROUP
Education

1818 H Street NW, Washington, DC 20433

Telephone: 202-473-1000; Internet: www.worldbank.org

Cover and interior design: Danielle Willis, Washington, DC, USA

CONTENTS

Acknowledgments

The *Guidance Note: How do we know teacher professional development is working? Measuring changes in teaching practices in the classroom* was led by Diego Luna-Bazaldua, Ana Teresa del Toro Mijares, Ezequiel Molina, and Adelle Pushparatnam.

The team received reviews and comments from Fadila Caillaud (Lead Economist, HMNED), Helena Rovner (Senior Education Specialist, HLCED), Koen Martijn Geven (Economist, HSAED), and Marina Bassi (Senior Economist, HAEE1) from the World Bank Group, and Jacobus Cilliers (Professor, McCourt School of Public Policy - Georgetown University).

This package is part of a series of products by the *Teach* Team. Overall guidance for the development and preparation of the package was provided by Omar Arias, Practice Manager for the Global Knowledge and Innovation Team. The package was designed by Danielle Willis. Patrick Biribonwa provided administrative support.

Objective: To provide guidance on how to (1) establish a numerical indicator to measure changes in teaching practices through the use of classroom observation tools for use in education projects and (2) produce a benchmark to compare changes in teaching practices through this indicator.

Introduction

The teaching that students receive in the classroom is the most important school-based determinant of student learning. Thus, improving teaching within an education system is a necessary endeavor to improve student learning outcomes and address the global learning crisis (Hanushek & Rivkin 2010; World Bank 2018). Moreover, meaningful interactions between the teacher and his or her students are at the center of the learning process. The way that teachers interact with their students in the classroom makes all the difference in ensuring students' academic and socioemotional learning (Curby, Brock, & Hamre, 2013; Hatfield, Hestenes, Kintner-Duffy, & O'Brien, 2013; Kane, Taylor, Tyler, & Wooten, 2011; Muijs et al., 2014).

For this reason, education projects that seek to improve student learning frequently include components focused on improving teaching practices through interventions such as modifying the curriculum, improving pre-service or in-service teacher training, and integrating additional instructional support to the classroom through the use of structured instructional material or technology.[1] These different interventions rely on an underlying theory of change, namely that *teachers improve their teaching practice to improve student achievement.*

Many programs that focus on building teacher capacity today collect data on ***inputs***, or activities, related to program implementation—for example, how many teachers attended trainings, how many teachers received regular coaching or mentoring support, or how many teachers received program materials for the classroom. However, not as many programs also collect data on the ***results*** of these interventions; that is, on whether the programs are leading to improvements in teaching practices in the classroom. Without these data, it is impossible to know whether educational interventions are producing meaningful changes in the learning experience that students receive. Ultimately, programs that are not effective in improving teaching practices in the classroom and improving the learning experience for students—no matter how many teachers they train—will not lead to improved student learning.

In this context, indicators that measure changes in teaching practices can be particularly valuable to understand whether educational interventions will lead to the expected improvements in student achievement (see Figure 1). For example, in the context of interventions focused on improving teachers' in-service professional development (TPD), an indicator that measures changes in teaching practices in the classroom can serve as an intermediate measure of the link between output indicators such as the number of teachers that attended training(s) or received coaching support, and impact indicators related to student achievement (World Bank, forthcoming).

[1] For example, as of April 2021, the WB EDU GP had 166 active projects (excluding small grant RE projects). Out of a sample of 56 active projects in a total of 12 countries, 49 projects have interventions related to pre-service or in-service teacher training and instructional support such as scripted lesson plans, and more than 75% of them have PDO- or IRI-level indicators related to teachers or teaching practices. For more information, please consult the World Bank Teacher Portfolio Repository (World Bank, 2021).

Figure 1. Logic Framework for a Sample Program Focused on Improving Early Grade Reading Outcomes Through an In-service TPD Intervention

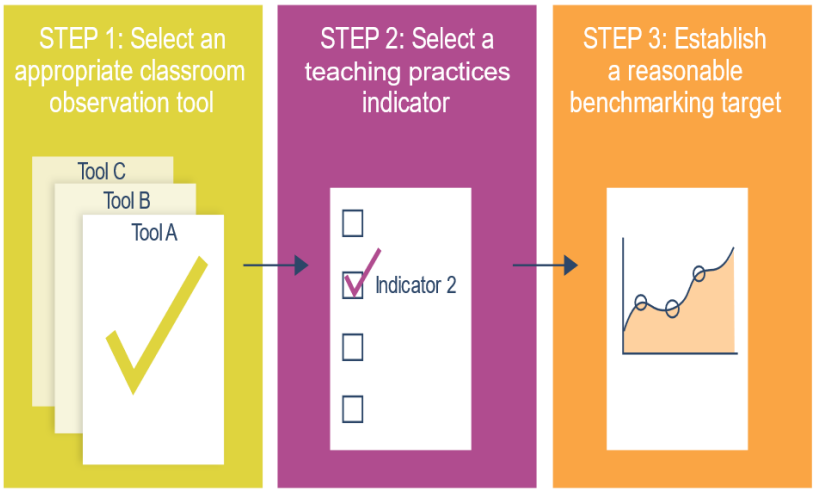
ACTIVITIES	OUTPUTS	OUTCOMES	IMPACT
Initial cluster-based intensive training on strategies for teaching reading Monthly school-based training to reinforce key concepts Monthly classroom observations Monthly 1:1 follow-up coaching sessions with an experienced instructional coach	Number of training hours delivered to each teacher Number of classroom observations conducted Number of 1:1 follow-up coaching sessions provided to each teacher	Improvement in teaching practices by classroom teachers in Grades 1–3	Improvement in student learning reading outcomes for Grades 1–3

Source: Developed based on World Bank (forthcoming).

The rest of this note provides guidance on how to establish a numerical indicator to measure teaching practices through classroom observation tools, and how to benchmark this indicator to track changes in teaching practices over time for use in educational interventions. This guidance is structured through a three-step process (Figure 2):

- ***Step 1:*** Selecting an appropriate classroom observation tool;
- ***Step 2:*** Selecting an indicator to track teaching practices using that tool; and
- ***Step 3:*** Establishing a reasonable benchmarking target for the chosen indicator.

Figure 2. Three-step Guidance Process



The following sections of this document provide guidance on each of these steps, in sequence. Annexes A through F provide additional information on each step, as well as resources for further reading.

A Note on Indicators to Measure Teaching Practices

Research has shown that what drives student learning most directly within the school environment is the quality of interactions between the teacher and students in the classroom, also known as ***process quality*** (Curby, Brock, & Hamre, 2013; Hatfield, Hestenes, Kintner-Duffy, & O'Brien, 2013; Kane, Taylor, Tyler, & Wooten, 2011; Muijs et al., 2014). Therefore, interventions that seek to improve teacher capacity with the goal of improving student learning must necessarily focus on capturing the extent to which the intervention has shifted the quality of the interactions between the teacher and the student(s), that is, the **teaching practices** used by the teacher in the classroom with his or her students.

There are multiple methods of measuring and tracking teaching practices in the classroom. These include classroom observations, student test scores and value-added models, principal and peer evaluations, self-reports of teaching practice, teaching portfolios of evidence, and student evaluations, among others. Each measurement method and source of evidence presents advantages as well as limitations (Little, Goe & Bell 2009).

This guidance note focuses on *measuring changes in teaching practices through the use of classroom observation tools*. Classroom observations have been recognized as one of the most direct ways to measure teaching practices, since they focus on the observable behaviors exhibited by the teacher within the classroom (Little, Goe & Bell 2009), therefore providing a direct measure of the interaction between teachers and students.

Importantly, teaching practice indicators as described and conceptualized in this note are meant to be used to assess progress and outcomes *at the project level*. While the classroom observation tools discussed in this note gather information on teaching practices at the individual teacher level, the aggregated classroom observation data are meant to be analyzed, reported, and shared only at the project level with the goal of assessing project progress and outcomes. The guidance in this note is not meant to be used for any kind of high-stakes decisions (including retention, promotion, and advancement of individual teachers) within a teacher career or professional development system.

The scope of this note is limited to establishing an indicator to measure changes in teaching practices. For guidance on how to operationalize and implement a classroom observation tool in the field, please consult the resources listed in Annex F of this document, as well as the publicly available *Teach Implementation Guide.* For detailed case studies on the use of the *Teach* classroom observation tool within projects to date, including how the tool was deployed in the field and integrated within project operations, please consult the *Teach in Action* brochure available on the *Teach* website.[2]

Step 1: Selecting an Appropriate Classroom Observation Tool

The first step in this process consists of selecting an appropriate classroom observation tool. Classroom observation tools vary according to the teaching practices they capture and how they do

[2] This guidance note is not *Teach* specific. The *Teach* resources here listed are used for illustrative purposes.

so, as well as other characteristics like their psychometric properties,[3] the extent to which they have been used in different contexts, and the complementary resources and costs associated with each tool. Given these considerations, it is important to carefully evaluate which tool is most appropriate for the project's needs and objectives. This step is critical, as selecting a tool incorrectly at this first stage will be difficult to fix later and will have important cost implications given investments in money, time, and human resources through the project's lifetime.

It is important to keep in mind these six key criteria to assess, evaluate, and select the most appropriate classroom observation tool for a project:

1. **Does the tool measure the appropriate domains of teaching practice, for the appropriate education level(s)?** The most important consideration is ensuring that the selected tool measures the domains of teaching practice (e.g., effective instruction, promotion of a specific learning culture in the classroom, provision of constructive feedback to students) targeted by the intervention. It is also important to also ensure that the tool has been designed for use within the appropriate education levels.

2. **Has the tool been designed for the role it will play within the project?** Classroom observation tools are developed with different objectives in mind; therefore, it is important to use measurement instruments for activities aligned with their original purpose. When using teaching practices indicators to measure the effectiveness of an educational intervention at the aggregate level, it is recommended to use classroom observation tools developed to assess changes in overall teaching practices and not those of individual teachers.[4]

3. **Has the tool been used in similar low- and middle-income (LMIC) country contexts?** Many classroom observation tools have not been designed for their use in LMIC countries, so it is important to consider the extent to which a specific tool has been adapted to and effectively used in these contexts.

4. **Is the tool valid and reliable?** In order to get accurate information about teaching practices, it is critical to evaluate a classroom observation tool's psychometric properties, to ensure that the tool is able to accurately capture changes in the teaching practices outcomes of interest. In this domain, it is important to ensure that the tool presents both **reliability** (i.e., the measurement tool produces accurate and consistent scores of teaching practices) and **validity** evidence (i.e., the measurement tool scores are correctly interpreted and appropriately used for specific purposes) (Ladics et al. 2018).

5. **What kinds of guidelines and materials are available to support implementation?** The extent to which a classroom observation tool includes complementary materials to support in every step of its implementation can make deploying the tool substantially easier. When comparing different classroom observation tools, it is important to consider the extent to which they are complemented by resources such as training guidebooks, scoresheets, data

[3] *Psychometric properties* refer to the technical characteristics of any measurement tool or instrument, such as its accuracy in measuring a specific construct, the consistency of its scores, or the pertinence of its use in a particular context.
[4] For more on the different purposes of classroom observation tools, consult Annex A of this document.

management tools, software to analyze and consolidate results, terms of references, materials to support in the contracting of key project members, and so on.

6. **What are the costs involved in the use of the tool?** It is important to consider the costs (both fixed and recurring) associated with the use of the tool, including licensing, initial and refresher trainings, data collection, and the use of complementary materials such as scoresheets and software to process, analyze data, and report and disseminate results, among other activities.

For more resources linked to step 1, consult:

- **Annex A** — *for more information on what each of the criteria above entail*
- **Annex B** — *for a checklist to aid in comparing and selecting appropriate classroom observation tools*
- **Annex F**— *for additional reading on how to select and choose a classroom observation tool*

Step 2: Selecting an Indicator to Measure Teaching Practices Based on the Chosen Tool

Once a classroom observation tool has been selected, the next step is to establish a teaching practices indicator. The indicator is a numerical value linked to a measure of teaching practices. The first key decision is to establish the type of indicator that will be utilized.

This paper identifies four types of indicators used to date in World Bank education projects to compare temporal changes in teaching practices measures, although a range of other types of indicators exists. These four indicators use the same input information to capture changes in teaching practices over time with variations depending on how each indicator is calculated:

- **Indicator 1**: Changes in the average of composite teaching scores across a population of teachers
- **Indicator 2:** Percentage of a population of teachers who show improvements in a composite teaching score
- **Indicator 3**: Percentage of population of teachers who surpass a minimum threshold in terms of a teaching score
- **Indicator 4**: Percentage of population of teachers who move across tiers of scores.

Some classroom observation tools capture more than one relevant domain linked to teaching practices, such as effective instruction, fostering students' socioemotional skills, or promoting a learning culture in the classroom. The four indicators assume that the project team has chosen to report a composite score, pulled from all domains measured by the classroom observation tool. In some cases, teams may choose to report scores for each domain separately, or a set of subdomains, instead of or in addition to reporting the full composite score. This may be the case for an intervention focused on building teachers' skills within the specific domain. In this case, the four indicators still apply to each domain score.

In the text that follows, each of the approaches is briefly described, including a summary of its advantages and disadvantages. These descriptions assume a baseline and endline data collection of the same sample of teachers.

Indicator 1. Changes in the average of composite teaching scores across a sample of teachers

In this approach, a classroom observation tool yields a teaching score across one or several domains of teaching practice for each teacher. These domain-specific teaching scores are averaged or weighted accordingly to arrive at a composite teaching score for each teacher observed. An average composite teaching score is then produced for a sample of teachers.

After a second data collection, the two average composite teaching scores are compared to assess the change in average composite teaching scores over time. This change can be benchmarked and reported as a change in the score, as a percentage, or as an effect size.

> **Example:** "Increase in the average composite teaching score as measured by a classroom observation tool for the group of teachers participating in the program in X province."

The *advantage* of this indicator is that it provides precise information about teaching practices scores within a sample population of teachers, and it provides the most precise information about the magnitude of changes in teaching practices scores—large or small—over time as a result of an intervention. Moreover, the score changes between baseline and endline can be expressed as an effect-size measure.

The *disadvantage* of this indicator is that it does not capture a minimum threshold of improvement for specific teachers, but rather relies on an aggregate trend of improvement over time. For example, an increase in this indicator may be driven by a subgroup of teachers that is showing meaningful improvement, while obscuring the fact that another subgroup is showing no improvements or is even decreasing in teaching practices scores over time.

Indicator 2. Percentage of a population of teachers who show improvements in a composite teaching score

In this approach, a classroom observation tool yields a teaching score across one or several domains of teaching practice for each teacher observed. These domain-specific teaching scores are averaged or weighted accordingly to arrive at a composite teaching score for each teacher observed.

Upon a second data collection, the number of teachers who show an increase in their endline composite teaching scores in comparison to baseline is captured as a percentage of the sample population.

> **Example:** "Percentage of primary school teachers in X province who show an improvement in teaching endline scores versus baseline scores as measured through a classroom observation tool".

The *advantage* of this indicator is that it provides information on the proportion of teachers who have shown some improvement, highlighting whether the intervention has been able to affect the teaching practices in the intended population. It is an indicator that is easy to understand and communicate effectively.

The *disadvantage* of this indicator is that it does not capture the magnitude of the changes in teaching practices. Teachers who show small increases in teaching practices are not differentiated from those who show large increases through this indicator, so even small positive changes in endline would contribute to increasing the indicator. Therefore, an increase in this indicator would only demonstrate the proportion of teachers who have shown some improvements, and not the magnitude of that change.

Indicator 3. Percentage of teachers who surpass a minimum threshold of teaching practices score

In this approach, a minimum cut-off score threshold is established for a composite teaching score. A classroom observation tool yields a teaching score across one or several domains of teaching practice for each teacher observed. These domain-specific teaching scores are averaged or weighted accordingly to arrive at a composite teaching score for each teacher observed. At baseline, the indicator captures the percentage of teachers whose composite teaching scores surpass the established cut-off score threshold.

After a second data collection, the percentage of teachers whose second teaching scores surpass the established threshold is compared to the original percentage.

> **Example:** "Percentage of primary school teachers in X province who are meeting standards in student-centered teaching practices as measured through a classroom observation tool".

The *advantage* of this indicator is that it provides information on the proportion of teaches who have met a minimum standard of teaching practices, which may be useful in ensuring that *all* teachers in a given population have reached an established standard. This indicator is easy to understand for many stakeholders.

The *disadvantage* of this indicator is that it only captures the proportion of teachers that surpassed the cut-off score, and it does not capture the magnitude of changes in teaching practices below or above the set threshold. For example, if the threshold is set at score X, and many teachers improve substantially but still fall below score X, their improvement will not be captured by this indicator. Similarly, teachers who already score above score X and improve beyond this score, will not be captured through this indicator. Additionally, the use of a specific threshold can incentivize behaviors to surpass this minimum score without motivating meaningful changes in instruction and classroom practice, potentially incentivizing support to be focused on teachers who are close to the cut-off points, rather than all teachers or those who most need support.

Indicator 4. Percentage of teachers who move across tiers of teaching practices scores

In this approach, tiers of performance are established with minimum and maximum composite teaching scores. A classroom observation tool yields a teaching score across one or several domains of teaching practice for each teacher observed. These domain-specific teaching scores are averaged or weighted accordingly to arrive at a composite teaching score for each teacher observed. At baseline, the indicator measures the percentage of teachers whose composite scores lie within the boundaries of each of the performance tiers established.

Upon a second data collection, the percentage of teachers who have improved across tiers is captured.

> **Example:** "Percentage of primary school teachers in X province who fall under Tier 3 as measured through a classroom observation tool, compared to baseline."

The *advantage* of this indicator is that it provides information on the proportion of teachers who fall in representative tiers of teaching practices (for example, the proportion of teachers who meet a minimum standard of teaching practices can be captured), while still tracking changes beyond this minimum standard through the use of tiers. The inclusion of clear explanations of expected teaching practices in each tier can facilitate the observation tool score interpretation.

The *disadvantage* of this indicator is that it does not fully capture the magnitude of changes in teaching practices. For example, if a given teacher first scores in Tier 2, and then increases in her score but still within the range of Tier 2, the indicator would not capture this increase in teaching practices. Additionally, the use of tiers can incentivize specific behaviors to only surpass the cut-off score for a higher tier without motivating meaningful changes in instruction and classroom practice towards reaching better performance.

Recommended indicator

All things equal, the indicator that provides the most unbiased information regarding teaching practices is **Indicator 1**, and its use is recommended for project teams.

This indicator may be complemented by any of the other three, depending on the context and objectives of the project. For simplicity, it is recommended that these indicators not be combined into one to permit an easier interpretation of the progress over time; instead, project leaders are encouraged to select one or more indicators and calculate them separately to capture different aspects linked to improvements in teaching practices over time.

For more resources linked to step 2, and for in-depth guidance on how to construct each of the four indicators, consult Annex C.

Step 3: Establishing a Reasonable Benchmarking Target for the Chosen Indicator

Once the type of indicator has been selected, the final step is *benchmarking* the indicator by establishing a reasonable target for improvement over a baseline.[5] In general, the expected improvements in classroom teaching practices will vary by intervention (Kraft, Blazar & Hogan 2018), according to design and implementation factors like the following, among others:

- **Type of intervention**: What are the components of the intervention? And to what extent is the intervention focused on building and consolidating teachers' instructional skills? Research has shown that some interventions (e.g., coaching) are more effective at improving teachers' instructional skills than others.
- **Duration of the intervention**: Is the intervention comparatively short? That is, will it last less than a year, or is it medium-term (e.g., 2 to 4 years) or of longer duration? All things equal, longer interventions should lead to larger increases in teaching practices, up to a point.
- **Dosage of intervention**: How intensive is the intervention? All things equal, interventions with higher dosages as measured through output indicators (such as hours of group training or hours of individual coaching provided) should lead to larger increases in teaching practices, up to a point.
- **Fidelity with which the intervention is deployed in the field**: Is the intervention implemented as designed and intended? Assuming an appropriate intervention design, an intervention that is implemented with low fidelity should be expected to yield smaller increases in teaching practices.
- **Extent to which teacher training specifically targets teaching domains captured in the observation tool**: To what extent is the classroom observation tool aligned to the content and learning provided to teachers? In general, the more aligned the tool is to the specific practices and behaviors targeted during teacher training, the better the tool will be able to capture changes in those domains. This consideration highlights the need to focus on an appropriate classroom observation tool (see criteria listed under Step 1 and described in more detail in Annex A).

Each project should assess how its approach aligns within the criteria above and set a benchmark target informed by the intervention's specific design and unique context.

[5] In this note, the focus is on indicators that are easy to calculate and interpret. Potential hybrid indicators that combine some of the properties of the four indicators here described could also be developed. In addition, the use of some advanced psychometric methods, including factor analysis and item response theory models, can be explored to construct indicators. While outside the scope of this note, if more advanced methods will be used, it is important to consider whether the classroom observation tool includes comprehensive documentation regarding its psychometric properties and whether the sample in the study design is of an appropriate size for an advanced analysis. It is also important to anticipate consultations with technical experts in psychometrics to conduct the analytical process, and to develop communication and dissemination strategies that present easy-to-understand results, given the complexity in the analysis.

Establishing a target for improvement

Indicator 1 has an important advantage over other indicators: It permits the calculation of a corresponding effect size[6] when comparing baseline and endline scores. This allows teams utilizing this indicator to leverage evidence from empirical research and past projects that have used effect sizes to estimate an appropriate target. For indicator 3, baseline and endline measures in a comparison group are necessary to calculate effect size estimates. Effect sizes cannot be calculated for indicators 2 and 4.

In the context of classroom observation tools that capture better teaching practices by higher scores, *positive effect size measures* indicate greater average scores of teaching practices in endline compared to baseline. *Negative effect size measures* are not expected but, if they happen, will indicate lower average scores of teaching practices in endline compared to baseline. Effect sizes *close to zero* indicate no or little average difference between baseline and endline measures (Cohen 1988).

Most research on the interpretation of effect size magnitudes within field-based education interventions has focused on providing guidance for interpreting effect sizes for outcomes measured by student achievement. Under this schema, small effect sizes (below 0.05 effect size) indicate minor average changes in learning outcomes over time, medium effect sizes (between 0.05 and 0.2 effect size) indicate some average changes in outcomes, and large effect sizes (above 0.20 effect size) indicate a considerable average score change over time (Kraft, Blazar & Hogan 2018).

While there has been less research to date on interpreting effect size magnitudes in interventions with outcomes measured by improvements on some measure of teaching practices, recent publications provide helpful guidance. A recent meta-analysis of the effect of teacher coaching programs on instructional practice shows importance variance in effect size values, ranging from 0.17 to 0.92 SDs, with an overall pooled effect size of 0.49 SDs (and 0.42 SDs when excluding studies in the United States) (Kraft, Blazar & Hogan 2018).

It is important to note that the teacher coaching programs included in this meta-analysis were mostly of one to two years in total duration. It is also important to note that research on the magnitude of effect sizes for educational interventions that seek to improve teaching practices show that effect sizes are smaller for studies with larger numbers of teachers, indicating fade-out effects as interventions scale up (Kraft, Blazar & Hogan, 2018). This is an important consideration for large-scale projects in which system-level teaching practices indicators will be benchmarked over time.

In light of this meta-analysis and past education projects, teams setting and benchmarking project teaching practices indicators should weight what would be a reasonable expected effect size given the considerations and empirical evidence here described.

[6] In statistics, an effect size is a standardized measure that captures the magnitude of an intervention. Larger effect sizes indicate a stronger impact of the intervention (e.g., in-service teacher training programs) on a measured outcome (e.g., positive changes in classroom observation tool scores).

In constructing a benchmark target for **Indicator 1**, it is recommended that expected targets ***should be set within a range of 0.2 and 0.5 effect size units, depending on the type of intervention, duration, dosage, and other relevant factors that can have an impact on improvements in teaching practices.*** **Box 1 offers two examples of the modeling used to measure interventions.**

Box 1. Examples of Modeling Used to Measure Intervention Targeting and Improvement Outcomes

Example 1: A high-intensity coaching intervention

Assume that a team is helping develop a high-intensity coaching intervention in a district in country Y. The coaching intervention is being designed according to best practice, and the program is highly intensive: teachers are expected to receive weekly personalized coaching visits and support to improve their teaching practice.

At baseline, teaching practices are assessed on a scale of 1-5 using the *Teach* tool; the baseline aggregate score is 2.7 with a within-group standard deviation (SD) of 0.5.

The team is estimating a high-impact intervention, and assessing the literature on coaching programs, sets a target of a desired effect size of 0.5 SD. Using the baseline aggregate score, and an SD_{within} of 0.5, the team's calculation yields a target $Score_{Endline}$ of 2.95, that is, an improvement of 9% over baseline teaching practices.

$$0.5 = \frac{Score_{Endline} - 2.7}{0.5\ SD}$$

$$[Target\ for]\ Score_{Endline} = 2.95$$

$$[Target\ for]\ Score_{Endline} = 9\%\ improvement\ over\ baseline$$

Example 2: A curriculum reform and teacher training

Assume that a team is supporting a country that has recently implemented substantial curriculum reform, and is designing a program to support teachers in shifting their teaching to the new pedagogical model that accompanies the curricular reform. Teachers will receive intensive initial training at the start of the school year and further training halfway through the school year.

At baseline, teaching practices are assessed on a scale of 1-5 using the *Teach* tool; the baseline aggregate score is 2.7 with a within-group SD of 0.5.

The team is estimating a medium-impact intervention, and assessing the literature on similar group-training programs, sets a target of a desired effect size of 0.2 SD. Using the baseline aggregate score, and an SD_{within} of 0.5, the team's calculation yields a target $Score_{Endline}$ of 2.8, that is, an improvement of 3.7% over baseline teaching practices:

$$0.2 = \frac{Score_{\,Endline} - 2.7}{0.5\ SD}$$

$$[Target\ for]\ Score_{\,Endline} = 2.8$$

$$[Target\ for]\ Score_{\,Endline} = 3.7\%\ improvement\ over\ baseline$$

For more resources linked to step 3, consult:

- ***Annex D*** — for an example on setting a target for Indicator 1 by following this guidance
- ***Annex E*** — for an example on how to benchmark targets expressed in composite teaching scores and percentage increases, consult Annex E and Table A-5 which provides projections of different endline scores given a baseline teaching score, under different improvement scenarios
- ***Annex F*** — for more on how to interpret effect sizes in the field of education
- ***Endline Score Projection Table*** for support in calculating projected changes in teaching practices under different scenarios

References

Azam, M., and G.G. Kingdon. 2015. “Assessing Teacher Quality in India.” *Journal of Development Economics 117*, 74-83.

Borenstein, M., L.V. Hedges, J.P. Higgins, and H.R. Rothstein. 2011. *Introduction to Meta-Analysis.* Hoboken, NJ: Wiley.

Cilliers, J., B. Fleisch, C. Prinsloo, and S. Taylor. 2020. “How to Improve Teaching Practice? An Experimental Comparison of Centralized Training and In-Classroom Coaching.” *Journal of Human Resources 55*(3), 926-962.

Cohen, J. 1988. Statistical Power Analysis for the Behavioral Sciences. Hillsdale, NJ: Erlbaum.

Curby, T. W., Brock, L. L., & Hamre, B. K. 2013. “Teachers' emotional support consistency predicts children's achievement gains and social skills.” *Early Education & Development*, *24*(3), 292-309.

Dobbie, W., and R.G. Fryer Jr. 2013. “Getting Beneath the Veil of Effective Schools: Evidence from New York City.” *American Economic Journal: Applied Economics*, *5*(4), 28-60.

Hatfield, B. E., Hestenes, L. L., Kintner-Duffy, V. L., & O’Brien, M. 2013. “Classroom Emotional Support predicts differences in preschool children's cortisol and alpha-amylase levels.” *Early Childhood Research Quarterly*, *28*(2), 347-356.

Hanushek, E.A., and S.G. Rivkin. 2010. “The Quality and Distribution of Teachers Under the No Child Left Behind Act. *Journal of Economic Perspectives 24*(3), 133-50.

Hunter, S.B. 2020. “The Unintended Effects of Policy-Assigned Teacher Observations: Examining the Validity of Observation Scores.” *AERA Open* *6*(2), 2332858420929276.

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

Kraft, M.A. 2020. “Interpreting Effect Sizes of Education Interventions.” *Educational Researcher 49*(4), 241-253.

Kraft M.A., D. Blazar, and D. Hogan. 2018. “The Effect of Teacher Coaching on Instruction and Achievement: A Meta-Analysis of the Causal Evidence.” *Review of Educational Research, 88*(4), 547-588.

Ladics, J., E. Molina, T. Wilichowski, and N. Yarrow. 2018. *The Measurement Crisis: An Assessment of How Countries Measure Classroom Practices*. Paper presented at the 2018 Research on Improving Systems of Education (RISE) Annual Conference, Oxford, UK.

Little, O., L. Goe, and C. Bell. 2009. A Practical Guide to Evaluating Teacher Effectiveness. National Comprehensive Center for Teacher Quality.

Muijs, D., Kyriakides, L., Van der Werf, G., Creemers, B., Timperley, H., & Earl, L. 2014. “State of the art–teacher effectiveness and professional learning.” *School effectiveness and school improvement*, *25*(2), 231-256.

Piper, B., S. Zuilkowski, M. Dubeck, E. Jepkemei, and S.J. King. 2018. “Identifying the Essential Ingredients to Literacy and Numeracy Improvement: Teacher Professional Development and Coaching, Student Textbooks, and Structured Teachers’ Guides.” *World Development, 106*, 324-336.

Snilstveit, B., J. Stevenson, R. Menon, D. Phillips, E. Gallagher, M. Geleen ... and E. Jimenez. 2016. *The Impact of Education Programmes on Learning and School Participation in Low-and Middle-Income Countries.* Systematic Review Summary 7. London, U.K: International Initiative for Impact Evaluation.

World Bank. 2018. *Ending Learning Poverty: What will it take?* Washington, D.C.

World Bank. (Forthcoming). *Monitoring and Evaluation for In-Service Teacher Professional Development Programs.* Coach Program Guidance Note. Washington, DC.

Annex A: Criteria in the Selection of a Classroom Observation Tool

It is important to keep in mind these key considerations to assess, evaluate, and select the most appropriate classroom observation tool for a project:

1. **Does it measure the appropriate domains of teaching practice, for the appropriate education level(s)?** *The most important consideration is ensuring that the selected tool measures the domains of teaching practice targeted by the intervention.* Depending on a project's ultimate goals, the selected tool may have to measure specific domains of classroom teaching practices. These domains may include, for instance, exhibition of effective pedagogical strategies during instruction, facilitation of a positive learning environment in the classroom, or the promotion of socioemotional skills in students. It is important to also ensure that the tool has been designed for use within the appropriate education levels.

2. **Has the tool been designed for the purpose it will play within the project?** Classroom observation tools are developed with different objectives in mind, including population monitoring, formative feedback provision, impact evaluation, research and hypothesis-testing, and screening for high-stakes decision-making about teachers and their careers (see more information on uses of classroom observation tools under Table A-1, Annex A). As a best practice, it is important to use instruments for activities aligned with their original purpose. In this sense, it is recommended to use classroom observation tools developed for *population monitoring* purposes to set indicators to get a sense of changes in teaching practices at the system level over time. These tools provide a snapshot of teaching practices from a representative sample of teachers within the education system to understand the level of instruction quality in the system. If additional information about student achievement is also measured, these tools provide relevant information about the relationship between teaching practices and student achievement at the system level. Exploratory research tools, tools developed for impact evaluations, and those used for high-stakes decisions about teachers (e.g., hiring, promotion or compensation) may not be useful to set these indicators.

3. **Has the tool been used in similar LMIC contexts?** Many classroom observation tools have not been designed for low- and middle-income countries, and it is important to consider the extent to which a specific tool has been adapted to and used in these contexts. Has the tool been used in the project country before? If not, has it been used in similar contexts and if so, what kind of adaptations to the local context were made? Has the tool already been translated and adapted to the specific context where it will be used? Has the tool been implemented reliably by non-expert observers in this context?

4. **Is the tool valid and reliable?** In order to get accurate information about teaching practices, it is critical to evaluate a classroom observation tool's psychometric properties, to ensure that data collected and analyzed with the tool will provide a reliable snapshot of what's truly happening in the classroom and that the tool is able to capture changes in the

teaching practices outcomes of interest. In this domain, it is important to ensure that the tool meets both reliability and validity evidence for its intended uses (Ladics et al, 2018).

In terms of *reliability*, a classroom observation tool should include:

a) Evidence of high internal consistency (e.g., with reported Cronbach's alpha statistics above 0.70);
b) Evidence of consistency in the use of the tool between trained enumerators and master enumerators to decrease biases during the classroom observation process (e.g., with inter-rater score agreement coefficients above 80 percent at the aggregate level). Concerning this aspect, it is recommended to use a tool that incorporates procedures to ensure and increase consistency between enumerators (e.g., enumerator training checks, guidelines and training tests), even if these are trained but non-expert observers.

In terms of *validity,* a classroom observation tool should have documented validity evidence to ensure the appropriate interpretation and use of its scores. Accumulated validity evidence varies from one tool to another depending on each tool's intended use, but it is usually classified into five broad domains reported in the instrument manuals:

a) Items measure relevant behaviors and features linked to instructional quality. Moreover, the tool content does not include irrelevant factors unrelated to the teaching process.
b) To facilitate understanding of the item scoring process for enumerators, the scoring rubrics and scoring guidelines include examples of actions and behaviors of teachers with different levels of instructional quality.
c) Items are empirically consistent with each other. Positive but non-perfect inter-item correlations provide evidence of this internal coherence.
d) The observation tool scores predict or are positively correlated with student learning outcomes
e) Related to the purposes of the measurement tool previously described, the instrument scores are used for its intended purposes and promote instructional quality improvement over time.

Importantly, both validity and reliability are context-specific constructs. Tools may have documented robust psychometric evidence in one context that may or may not hold in a different one. Therefore, if possible, the selected tool should show both reliability and validity evidence in the same or similar contexts.

5. **What kinds of guidelines and materials are available to support implementation?** Deploying a classroom observation tool effectively will require training master trainers and classroom observers, having guidelines for the instrument translation and adaptation, and implementing a set of quality assurance checks to ensure that the data collected are reliable and valid. Once data are collected, they must be processed, analyzed, consolidated, and reported out to identified stakeholders. The extent to which a classroom observation tool includes complementary materials to support in every step of its implementation can make the evaluation process substantially easier. When comparing different classroom observation tools, consider the extent to which they come with complementary resources such as training guidebooks, scoresheets, data management

tools, software to analyze and consolidate results, ToRs and other materials to support in the contracting of key project members, and so on.

6. **What are the costs involved in the use of the tool?** Finally, it is important to consider the costs associated with the use of the tool. These can include licensing, initial and refresher trainings, data collection, and the use of complementary materials such as scoresheets and software to process, analyze and report results, among other items. Both start-up and recurring costs throughout the project lifetime should be considered in the tool selection process.

Table A-1. Definition of the Frequent Uses of Classroom Observation Tools

Criteria	Item
Population monitoring	The tool is used to monitor teaching practices at the system level. Results are commonly used to inform policy decisions and reforms to improve quality in the education system.
Formative feedback provision	The tool is used to provide formative feedback to teachers based on their performance and provide additional opportunities for professional development.
Impact evaluation	The tool is used in the context of a randomized-control trial to determine the impact of an intervention on teaching practices outcomes or as a mediator between an intervention and student achievement outcomes.
Research and hypothesis testing	The tool is used in the context of a research project to test the significance of correlations between teaching practices and other constructs, or comparisons of teaching practices among relevant groups.
High-stakes decisions	The tool is used for making decisions that will have substantial impact on teachers or schools. For instance, high-stakes decisions for teachers may include certification, promotion, or denial of tenure. High-stakes decisions for schools include allocation or denial of resources to improve the quality of education service delivery.

Annex B: Checklist for the Selection of a Classroom Observation Tool

The following checklist can be utilized by project leaders, policymakers and teams to help compare and assess a classroom observation tool for use in a project across the range of criteria described in Annex A.

Criteria	Item	Tool 1: _____
Domain	1a. Does the tool measure the appropriate domains of teaching practice targeted by the intervention?	☐ Yes ☐ No
	1b. Is it focused at the appropriate education level(s) (pre-primary, primary, secondary)?	☐ Yes ☐ No
Purpose	2a. Is the tool's purpose aligned to the way it will be used within the project's design and development?	☐ Yes ☐ No
Psychometric properties	3a. Does the tool present reliability evidence? See below: 1) Evidence of high internal consistency (e.g., with reported Cronbach's alpha statistics above 0.70); 2) Evidence of consistency between trained enumerators and master enumerators (e.g., with inter-rater score agreement above 80 percent); 3) Publication of reporting procedures (e.g., enumerator training checks, guidelines and training tests) to ensure consistency (recommended).	☐ Yes, 3/3 ☐ Yes, 2/3 ☐ Yes, 1/3 ☐ None
	3b. Does the tool present validity evidence?	☐ Yes ☐ No
Contextualization and adaptation	4a. Has the tool been designed with LMICs in mind?	☐ Yes ☐ No
	4b. Has the tool been used in similar contexts before?	☐ Yes ☐ No
	4c. Does the tool include guidelines for translation and adaptation?	☐ Yes ☐ No
Implementation support	5a. Does the tool come with implementation support and guidelines?	☐ Yes ☐ No
Costs	6a. What are the total initial startup costs (such as licensing, initial trainings) of using the tool?	$_____
	6b. What are the total estimated recurring costs (such as refresher trainings, use of software or materials to collect, process, analyze or report results, and/or updates to the tool as needed) of using the tool during the project's lifetime?	$_____

Annex C: Building an Indicator to Track Teaching Practices (Types 1–4)

The next examples use the score range of the *Teach* classroom observation tool, which takes values between 1 and to 5 using information from nine elements or items. Some adjustments may have to be made for classroom observation tools with a different score range and different number of items.

For all indicators, the process starts with the calculation of an average score for each observation in baseline and endline by adding the score in each item and dividing it by the total number of items. Here, scores for each of the nine elements captured by *Teach* range from 1 to 5. These nine scores can be aggregated to calculate an unweighted average *Teach* score for each case or observation, which consequently also takes values between 1 and 5 (see Equation 1):

$$Score_{\ average} = \sum_{1}^{9} \frac{Score_{\ Element\ Scores}}{9} \tag{1}$$

Indicator 1: Changes in average scores

This indicator is used to capture the effect size of changes in average *Teach* scores in a sample of teachers measured with respect to their classroom performance in baseline and endline time points (see Borenstein et al., 2011). Note that this formula does not consider the inclusion of a control or comparison group of teachers measured in both time points. It is recommended to get the support of a statistician or quantitative social scientist to compute this indicator.

This indicator is calculated following the next steps:

Step 1.1. Calculate *average endline Teach score* for the whole sample presented in equation 2. For the total sample *n* of observed teachers, this average score is obtained by adding the average *Teach* scores (defined in Equation 1) measured during the endline data collection.

$$\overline{Score}_{\ end\ line} = \sum_{i=1}^{n} \frac{Score_{\ average\ (in\ end\ line)\ i}}{n} \tag{2}$$

Step 1.2. Calculate *average baseline Teach score* for the whole sample presented in equation 3. Similar to equation 2, this average score is obtained by adding the average *Teach* scores (defined in Equation 1) measured during the baseline data collection.

$$\overline{Score}_{\ baseline} = \sum_{i=1}^{n} \frac{Score_{\ average\ (in\ baseline)\ i}}{n} \tag{3}$$

Step 1.3. Calculate *difference in average Teach scores* in endline minus baseline using the values computed in Equations 2 and 3 (see Equation 4). If the value of this difference is positive, it indicates that greater average *Teach* scores were observed at the end of the study compared to the beginning of it. Nevertheless, this difference is still not enough to capture the effect size for this *Teach* indicator.

$$\overline{Score}_{difference} = \overline{Score}_{end\ line} - \overline{Score}_{baseline} \quad (4)$$

Step 1.4. Calculate *within-group standard deviation of baseline and endline measures*. This standard deviation is calculated following the formula in equation 5.

$$SD_{within} = \frac{\sqrt{SD^2_{end\ line} + SD^2_{baseline} - (2 \times r \times SD_{end\ line} \times SD_{baseline})}}{\sqrt{2(1-r)}} \quad (5)$$

where $SD_{endline}$ and $SD_{baseline}$ correspond to the standard deviations for the average *Teach* scores in both time points, and *r* is the Pearson correlation between endline and baseline average *Teach* scores for all observed teachers.

Step 1.5. Calculate *effect size for the difference between baseline and endline measures*. This effect size is the *Teach indicator* to report as the difference in average *Teach* scores over time. As shown in equation 6, the indicator is the result of dividing the difference in average *Teach* scores calculated in Step 3 by the within-group standard deviation calculated in Step 4.

$$Score_{Effect\ Size.} = \frac{\overline{Score}_{difference}}{SD_{within}} \quad (6)$$

Note: Effect size measures can be easier to communicate to a broad audience (including policymakers) when expressed in terms of percentage points increases over time. See the example presented in Annex E for more information about the translation of effect size measures into percentages.

Indicator 2: Percentage of teachers who show improvements in a composite teaching score

Step 2.1. Following steps 1.1 to 1.3, calculate the difference in average *Teach* scores presented in equation (4). Use a logic indicator function to score teachers with positive difference in average *Teach* scores between baseline and endline with a value of 1, and with a value of 0 otherwise.

$$\overline{Score}_{difference\ IND} = \begin{cases} 1, & \overline{Score}_{difference}\ is\ positive \\ 0, & Otherwise \end{cases} \quad (7)$$

Step 2.2. Using the logic indicator function from equation (7), calculate the percentage of teachers that show improvement by adding up the logic indicator scores over cases, dividing that by the *n* total number of cases, and multiplying it by 100. The higher the value in this indicator, the higher the percentage of teachers that showed improvement in endline scores compared to baseline.

$$Indicator\ 2 = \left(\sum_{i=1}^{n} \frac{\overline{Score}_{difference\ IND}}{n} \right) \times 100 \quad (8)$$

Indicator 3: Percentage of teachers who surpass a minimum threshold in scores

Step 3.1. Following steps 1.1 to 1.2, calculate average baseline and endline *Teach* scores. Determine the cut-off score that will be used as a threshold of teaching practices. Then, generate two independent logic indicator functions to score teachers with baseline and endline average *Teach* scores above the threshold with a value of 1, and with a value of 0 otherwise.

$$\overline{Score}_{\,end\ line\ IND} = \begin{cases} 1, & \overline{Score}_{\,end\ line} > score\ threshold \\ 0, & Otherwise \end{cases} \tag{9}$$

$$\overline{Score}_{\,baseline\ IND} = \begin{cases} 1, & \overline{Score}_{\,baseline} > score\ threshold \\ 0, & Otherwise \end{cases} \tag{10}$$

Step 3.2. Using the logic indicator functions from equations (9) and (10), calculate the proportion of teachers that are above the threshold in baseline and endline by adding up the respective logic indicator scores over cases, and dividing them by the *n* total number of cases.

$$P_{end\ line} = \sum_{i=1}^{n} \frac{\overline{Score}_{\,end\ line\ IND}}{n} \tag{11}$$

$$P_{baseline} = \sum_{i=1}^{n} \frac{\overline{Score}_{\,baseline\ IND}}{n} \tag{12}$$

Step 3.3. Using the proportions calculated in equations (11) and (12), calculate Indicator 3 by computing the difference in these two proportions. If the indicator is positive, a higher proportion of teachers score above the threshold of teaching practices in endline compared to baseline.

$$Indicator\ 3 = P_{end\ line} - P_{baseline} \tag{13}$$

Indicator 4: Changes in the proportion of teachers in each Teach tier level

This indicator is used to capture changes in terms of the proportion of teachers that change from one tier to another in baseline and endline time points. *An effect size estimate cannot be produced using this indicator*. Note that the formula does not consider the inclusion of a control or comparison group of teachers measured and tiered in both time points. It is recommended to get the support of a statistician or quantitative social scientist to compute this indicator.

Recently, the team behind the development of the *Teach* classroom observation tool analyzed *Teach* data from seven low-, middle-, and high-income countries from around the world to determine the optimal number of tiers that could be produced using average *Teach* scores. Results showed that three tiered groups were optimal across countries (see Table A-2). Based on the information documented in Table A-2, notice that average *Teach* scores (as calculated in Equation 1) to place teachers on each tier level are mutually exclusive to avoid the classification of any teacher in more than one level.

Table A-2. Average *Teach* Score Tier Levels

Tier	Cut-off *Teach* scores	Tier description
3	From 3.01 to 5.00	Teachers in this tier attain higher average *Teach* scores than those in the other groups, as well as in the 9 elements and 3 primary areas measured by this classroom observation tool. Teachers in tier 3 more frequently promote a supportive learning environment in their classroom, deliver effective instruction supplemented with actions to monitor students' understanding of the learning contents, and reinforce socioemotional skills in their students. Despite the higher average *Teach* scores in comparison to the other tier levels, teachers in tier 3 could also benefit from additional support and training related to specific elements of classroom practice.
2	From 2.40 to 3.00	Teachers in this tier reach average *Teach* scores above those from their peers in tier 1, but below teachers in tier 3. They may present strengths in some of 9 elements measured by *Teach*, but require additional support to improve in others. Teachers identified in tier 2 will benefit from targeted support and professional development in the areas of their teaching practice where they score lower, while also getting general coaching to continue improving their overall teaching skills, monitor their classroom environment and promote a space for learning.
1	From 1.00 to 2.39	Teachers in this tier tend to obtain lower average *Teach* scores than those in the other groups; these teachers also show a lower performance on the 9 elements and 3 primary areas measured by this classroom observation tool. Teachers scoring within this tier require more support and training that can help them to foster a positive learning culture in their classroom. They can also benefit from coaching activities that will enable them to improve their instructional skills in the classroom and effective approaches to provide constructive feedback to their students. Teachers may also require additional concentrated training on key actions and strategies they can implement in the classroom to help students strengthen their socioemotional skills.

Using the information from Table A-2, this indicator is calculated following the next steps:

Step 4.1. Calculate *baseline and endline average Teach scores* for each teacher using Equation 1.

Step 4.2. Classify *each case in their corresponding baseline and endline tier using the cut-off Teach scores* listed in Table A-2.

Step 4.3. Produce a contingency table that compares the number of cases in each Teach tier both in baseline and endline. The table would be similar to Table A-3 shown below.

Table A-3. Example of Contingency Table for Teach Tier Levels

		Endline			
		Tier 1	**Tier 2**	**Tier 3**	***Row total (r)***
Baseline	**Tier 3**				
	Tier 2				
	Tier 1				
	Column total (c)				***Grand total***

Note: The cell colors in this table represent changes, as follows:

Blue	Indicates a positive change from a lower tier in baseline to an upper tier in endline.
Yellow	Indicates no change in tier level between baseline and endline.
Red	Indicates a negative change from an upper tier in baseline to a lower tier in endline.

Ideally, if a program is effective, more teachers should be placed in the blue cell, followed by a smaller amount of them in the yellow cell, and very few in the red cell.

Annex D: Setting an Appropriate Target for Indicator 1

In constructing a target for Indicator 1, it is recommended that expected targets should be set within a range of 0.2 and 0.5 effect size units, depending on the type of intervention, duration, dosage, and other relevant factors that can have an impact on improvements in teaching practices.

If a team has a baseline score of teaching practices ($Score_{Baseline}$) and establishes an assumption for SD_{within} , the team can compute a target for ($Score_{Endline}$) to match a desired effect size, where

$$Score_{Effect\ Size.} = \frac{\overline{Score}_{difference}}{SD_{within}} \quad (6)$$

In the following two examples, a within-group standard deviation of 0.5 is assumed. Prior *Teach* implementations to date have shown an average within-group standard deviation of 0.5-0.55.

Example 1.
Assume a desired effect size of 0.5 SD and an SD_{within} of 0.5, a baseline score of 2.7 on a 5-point scale yields the following target for $Score_{Endline}$:

$$0.5 = \frac{Score_{Endline} - 2.7}{0.5\ SD}$$

$$[Target\ for]\ Score_{Endline} = 2.95$$
$$[Target\ for]\ Score_{Endline} = 9\%\ improvement\ over\ baseline$$

Example 2.
Assume a desired effect size of 0.2 SD and an SD_{within} of 0.5, a baseline score of 2.7 on a 5-point scale yields the following target for $Score_{Endline}$:

$$0.2 = \frac{Score_{Endline} - 2.7}{0.5\ SD}$$

$$[Target\ for]\ Score_{Endline} = 2.8$$
$$[Target\ for]\ Score_{Endline} = 3.7\%\ improvement\ over\ baseline$$

Annex E: Numerical Example and Graphical Representation of the Four Indicators

An example based on real *Teach* data from a country in South Asia is included to illustrate (1) the computation of indicators and (2) benchmarking of baseline and endline measures. The aim of the example is both to illustrate the difference in teaching practices increases under different scenarios (small, medium, and large effect sizes), and to show how these increases would be captured differently by each of the 4 types of indicators listed under Step 2.

The example simulated endline data under three scenarios: small effect size (0.02), medium effect size (0.15), and large effect size (0.60) (Kraft et al., 2018). In the three scenarios, a baseline and endline Teach score for each classroom observation is calculated by averaging scores assigned to the nine Teach elements measured by this classroom observation tool.

Table A-4 shows some key *Teach* score statistics and indicators under each simulation scenario. As shown in this table, the average endline *Teach* scores tend to increase in simulation scenarios with a higher effect size. An additional Excel spreadsheet (Endline Score Projection Table) is available to support teams interested in computing Indicator 1 and the corresponding score change percentage under different scenarios.

Specifically for Indicator 1, effect-size estimates increase from 0.02 in the small effect-size scenario, 0.15 in the medium effect-size scenario, and 0.60 in the large effect-size scenario. When expressed in terms of the average score increase (percentage) between baseline and endline, these three scenarios correspond to 1%, 10%, and 13% score increases, respectively.

Indicator 2 increases from 13% of the teachers showing improvements in their *Teach* scores in the small-effect scenario to 97% of them showing improvement in the large-effect scenario.

For Indicator 3, the *Teach* score threshold is set at the value of 3. Teachers scoring below 3 are identified as below the threshold and those above 3 are identified as above the threshold. In the small-effect scenario, only 16% of the teachers are above this minimum threshold, whereas that proportion doubles to 32% of them surpassing this minimum score in the large-effect scenario.

For Indicator 4, only 1% of the teachers show a positive change in their tier level in the small-effect scenario, whereas 39% have a positive tier change in the large-effect scenario.

Notice that, despite being modeled based on the same baseline data and under the same set of assumptions, each indicator provides a different perspective of progress over time. For instance, under the large-effect scenario, the use of Indicator 2 would allow users to conclude that 97% of the teachers showed gains in their scores over time; nevertheless, only 32% of them scored above a minimum score threshold (Indicator 3), and 39% positively changed in their teaching practices tier level (Indicator 4). Thus, it is important to think about the use of a combination of teaching practices indicators to understand progress over time.

Table A-4. Selected *Teach* Score Statistics and Indicators Under Each Simulation Scenario

		Effect-size simulation scenario (n = 1754)		
		Small	Medium	Large
Statistic	**Baseline**	**Endline**	**Endline**	**Endline**
Teach mean	2.48	2.5	2.57	2.81
Teach SD	0.54	0.54	0.54	0.55
Indicators				
Indicator 1		0.02 ES	0.15 ES	0.60 ES
Indicator 2		13%	55%	97%
Indicator 3	15%	16%	19%	32%
Indicator 4		1%	10%	39%
% of mean score increase		1%	4%	13%

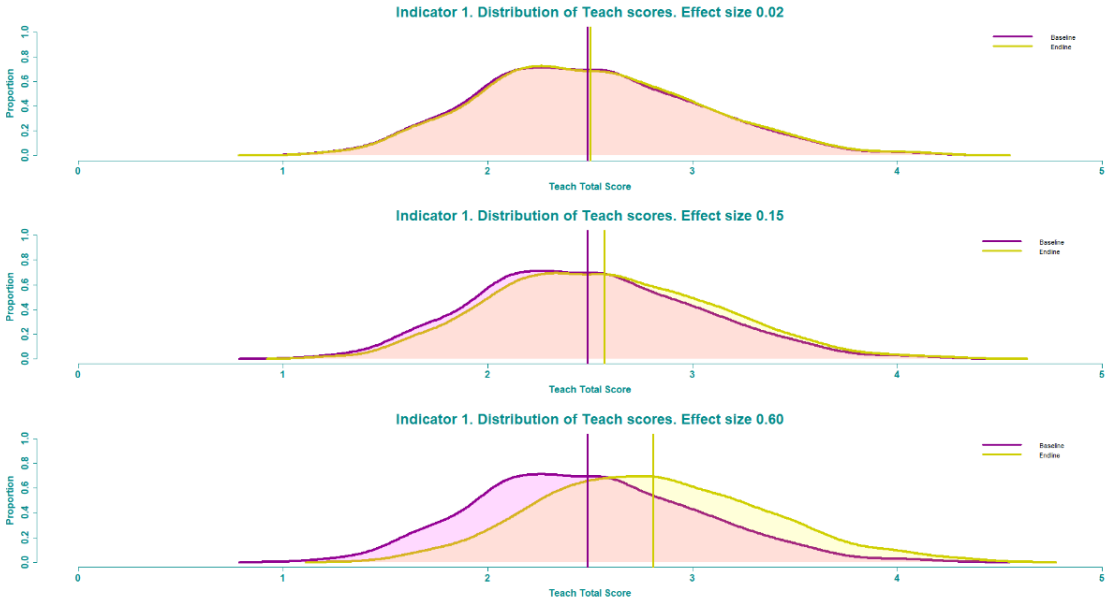
Different visual representations of these four indicators can be created using statistical software to visually convey changes in *Teach* scores. This example uses density plots for Indicator 1, bar plots for Indicators 2 and 3, and jitter plots for Indicator 4.

The preference for density plots in the case of Indicator 1 comes from the fact that the interest is to compare the distribution of numerical scores in two moments. Some statistical software packages permit one to produce density plots with visual representations of descriptive statistics like the mean and standard deviations. Alternative graphical representations for changes in score distributions could include histograms, boxplots, or scatterplots.

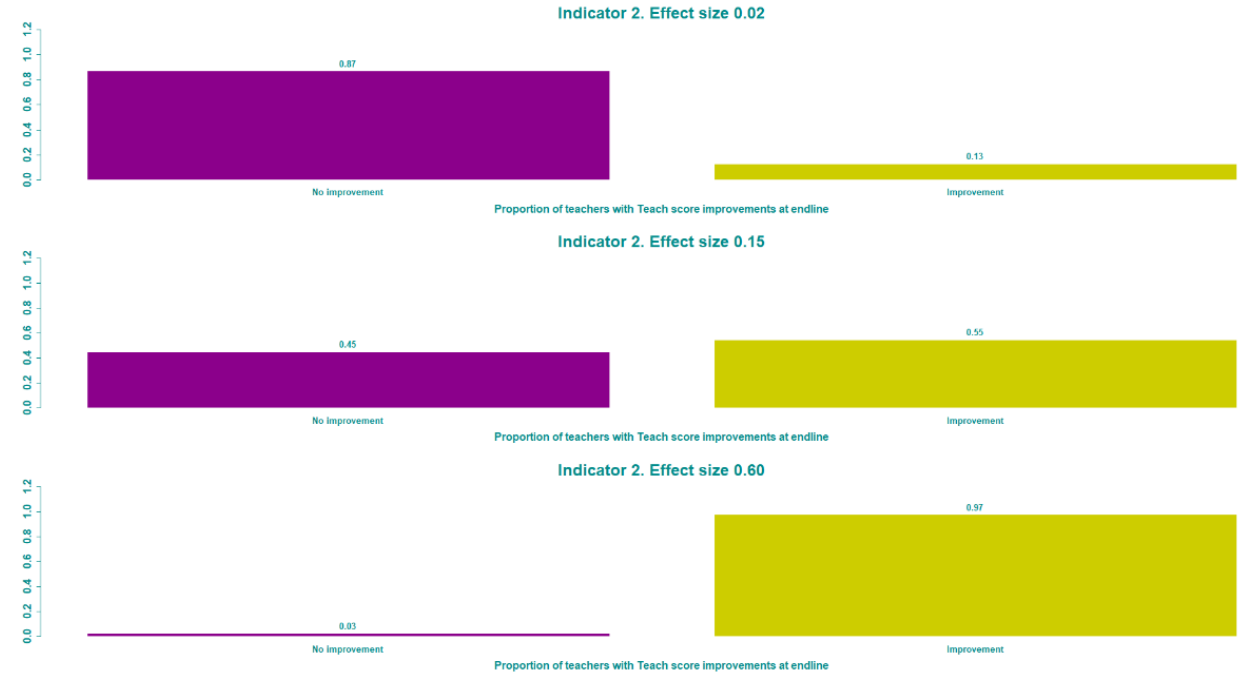
The use of bar plots for Indicators 2 and 3 is preferred to portray changes in proportions or percentages over time. Bar plots are also easy to understand for a wide range of audiences. Alternative plots to describe changes in proportions may include pie charts.

The use of jitter plots for Indicator 4 is preferred since these plots visually convey changes in group status when two categorical variables are cross-tabulated. Because they are used less frequently compared to density plots or bar plots, jitter plots may require some explanation to the target audience to indicate that each dot represents one or more observations and changes in tier levels over time.

Indicator 1. The following graphs depict the baseline and endline distributions of Indicator 1 under the small, medium, and large effect size scenarios.



Indicator 2. The following bar graphs depict the changes between baseline and endline scores for the low, medium and large effect size scenarios for Indicator 2.



Indicator 3. The following bar graphs depict the changes between baseline and endline scores for the low, medium and large effect size scenarios for Indicator 3.

Indicator 4. The following jitter plots depict the changes between baseline and endline scores for the low, medium and large effect size scenarios for Indicator 4.

Table A-5: Projected proportion of endline average score improvement

Here are different scenarios of the proportion of average endline score increase for different average baseline scores and effect sizes.

	Intervention effect size			
Average baseline score	***0.2***	***0.3***	***0.4***	***0.5***
2.0	*5.4%*	*8.1%*	*10.7%*	*13.4%*
2.1	*5.1%*	*7.7%*	*10.2%*	*12.8%*
2.2	*4.9%*	*7.3%*	*9.8%*	*12.2%*
2.3	*4.7%*	*7.0%*	*9.3%*	*11.7%*
2.4	*4.5%*	*6.7%*	*9.0%*	*11.2%*
2.5	*4.3%*	*6.4%*	*8.6%*	*10.7%*
2.6	*4.1%*	*6.2%*	*8.3%*	*10.3%*
2.7	*4.0%*	*6.0%*	*8.0%*	*9.9%*
2.8	*3.8%*	*5.8%*	*7.7%*	*9.6%*
2.9	*3.7%*	*5.6%*	*7.4%*	*9.3%*
3.0	*3.6%*	*5.4%*	*7.2%*	*9.0%*
3.1	*3.5%*	*5.2%*	*6.9%*	*8.7%*
3.2	*3.4%*	*5.0%*	*6.7%*	*8.4%*
3.3	*3.3%*	*4.9%*	*6.5%*	*8.1%*
3.4	*3.2%*	*4.7%*	*6.3%*	*7.9%*
3.5	*3.1%*	*4.6%*	*6.1%*	*7.7%*
3.6	*3.0%*	*4.5%*	*6.0%*	*7.5%*
3.7	*2.9%*	*4.4%*	*5.8%*	*7.3%*
3.8	*2.8%*	*4.2%*	*5.7%*	*7.1%*
3.9	*2.8%*	*4.1%*	*5.5%*	*6.9%*
4.0	*2.7%*	*4.0%*	*5.4%*	*6.7%*

Annex F: Further Resources

For further reading on the topics in this note, please consult the resources listed below.

Box F.1. Further Resources

To learn more about different classroom observation tools available, please consult:

- B. Bruns, S. De Gregorio & S. Taut (2016), "Measures of effective teaching in developing countries," *Research on Improving Systems of Education (RISE) Working Paper* 16(009).
- D. Filmer, E. Molina & W. Wane (2020). Identifying Effective Teachers: Lessons from Four Classroom Observation Tools (Policy Research Working Paper 9365) (Washington, DC: The World Bank).
- B. Gill, M. Shoji, T. Coen & K. Place (2016), *The Content, Predictive Power, and Potential Bias in Five Widely Used Teacher Observation Instruments* (REL 2017-191) (Regional Educational Laboratory Mid-Atlantic).
- J. Ladics, E. Molina, T. Wilichowski & N. Yarrow (2018, March), *The Measurement Crisis: An Assessment of How Countries Measure Classroom Practices*, paper presented at the 2018 Research on Improving Systems of Education (RISE) Annual Conference, Oxford, UK.
- F. Martinez, S. Taut & K. Schaaf (2016). "Classroom observation for evaluating and improving teaching: An international perspective," Studies in Educational Evaluation, 49, 15-29.
- E. Molina, S.F. Fatima, A.D. Ho, C. Melo, T.M. Wilichowski & A. Pushparatnam (2020), "Measuring the quality of teaching practices in primary schools: Assessing the validity of the Teach observation tool in Punjab, Pakistan," Teaching and Teacher Education 96, 103171.
- S. Pouezevara, A. Pflepsen, L. Nordstrum, S. King & A. Gove (2016). "Measures of quality through classroom observation for the Sustainable Development Goals: Lessons from low-and middle-income countries." Background paper for the 2016 Global Education Monitoring Report, Education for People and Planet: Creating Sustainable Futures for All.

For more on how to select and use a classroom observation tool, including information on validity and reliability, please see below:

- B. Hamre (Unknown), *Using Classroom Observation to Gauge Teacher Effectiveness: Classroom Assessment Scoring System (CLASS)*, presented at CEPR, Harvard University.

- S.B. Hunter (2020), "The unintended effects of policy-assigned teacher observations: Examining the validity of observation scores." *AERA Open* 6(2).

- D.A. Luna-Bazaldúa, E. Molina, & A. Pushparatnam (2021), "A generalizability study of Teach, a global classroom observation tool," in M. Wiberg, D. Molenaar, J. Gonzalez, U. Bockenholt, & J.-S. Kim (Editors), *Quantitative Psychology* (Annual Meeting of the Psychometric Society) (Switzerland: Springer Nature).

- M.W. Stuhlman, B.K. Hamre, J.T. Downer & R.C. Pianta (2010), *A Practitioner's Guide to Conducting Classroom Observations: What the Research Tells Us About Choosing and Using Observational Systems* (Charlottesville, VA: The Center for Advanced Study of Teaching and Learning, University of Virginia).

- The New Teacher Project (2009), Rating a Teacher Observation Tool: Five Ways to Ensure Classroom Observations Are Focused and Rigorous.

For more on effect sizes in the field of education, please consult:

- M.D. Baird & J.F. Pane (2019), "Translating standardized effects of education programs into more interpretable metrics," *Educational Researcher* 48(4), 217-228.

- H.S. Bloom, C.J. Hill, A.R. Black & M.W. Lipsey (2008), "Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions," *Journal of Research on Educational Effectiveness* 1(4), 289-328.

- M.A. Kraft (2020), "Interpreting effect sizes of education interventions," *Educational Researcher,* 49(4), 241-253.

- M.A. Kraft, D. Blazar & D. Hogan (2018), "The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence," *Review of Educational Research* 88(4), 547-588.

- C.J. Hill, H.S. Bloom, A.R. Black & M.W. Lipsey (2008), "Empirical benchmarks for interpreting effect sizes in research," *Child Development Perspectives* 2(3), 172-177.

Teach

WORLD BANK GROUP
Education

Contact us at **teach@worldbank.org** and
visit us at **www.worldbank.org/education/teach**