Received: 19 October 2021 | Accepted: 25 August 2022

DOI: 10.1002/berj.3837

ORIGINAL ARTICLE

BERJ BERA

The relationship between formative assessment and reading achievement: A multilevel analysis of students in 19 countries/regions

Zi Yan[1] | **Ming Ming Chiu**[2]

[1]Department of Curriculum and Instruction, The Education University of Hong Kong, Hong Kong, China

[2]Department of Special Education and Counselling, The Education University of Hong Kong, Hong Kong, China

Correspondence

Zi Yan, Department of Curriculum and Instruction, The Education University of Hong Kong, 10 Lo Ping Road, Tai Po, N.T., Hong Kong.
Email: zyan@eduhk.hk

Funding information

None

Abstract

Despite the general consensus on the positive impact of formative assessment on student learning, researchers have not shown the underlying mechanisms between specific formative assessment strategies and academic performance on an international sample. This study examines the link between student and teacher reports of teachers' formative assessment strategies (i.e. clarifying goals and monitoring progress, providing feedback, and instructional adjustments) and students' reading achievement, based on data from 151,969 fifteen-year-olds in 5,225 schools in 19 countries/regions in PISA 2018 via multilevel analysis of plausible values. The results show that clarifying goals and monitoring progress, and instruction adjustments are positively linked to reading achievement, but providing feedback alone has no significant impact. These findings highlight the complexity of formative assessment as a multifaceted concept and the different impacts of formative assessment strategies on student learning. Implications for researchers and practitioners are discussed.

KEYWORDS

formative assessment, instructional adjustments, reading achievement, PISA 2018

Key insights

What is the main issue that the paper addresses?

This study tests whether teachers' reports of their formative assessment strategies (i.e. clarifying goals and monitoring progress, providing feedback and instructional adjustments) or students' reports of them are linked to students' reading achievement, based on data from 151,969 fifteen-year-olds in 5,225 schools in 19 countries/regions in PISA 2018.

What are the main insights that the paper provides?

Clarifying goals and monitoring progress, and instruction adjustments are positively linked to reading achievement (but providing feedback alone is not). This study shows that formative assessment is a multifaceted concept, and its component strategies have different impacts on student learning.

INTRODUCTION

Formative assessment can provide evidence about student learning and inform adjustment of instruction to student needs, which has yielded superior learning outcomes (Black & Wiliam, 1998; Dunn & Mulvenon, 2009). As formative assessment is an umbrella concept without a set of defined strategies, studies operationalise it in different ways, focus on different formative assessment strategies (Bennett, 2011; Yan et al., 2021), and often have small sample sizes; thus, their results are difficult to interpret and might not be generalisable (Dunn & Mulvenon, 2009; McMillan et al., 2013). Hence, this study focuses on three specific formative assessment strategies, i.e. clarifying goals and monitoring progress, providing feedback and instructional adjustments, and examines whether teacher uses of these strategies are linked to students' reading test scores, using data from 151,969 fifteen-year-olds in 5,225 schools in 19 countries/regions from the 2018 Programme for International Student Assessment (PISA; OECD, 2018).

Theoretical framework of formative assessment

Both conceptualisations and classroom practices of formative assessment differ widely (Bennett, 2011; McMillan et al., 2013), although researchers generally recognise that formative assessment involves collecting evidence about student learning through various activities and using that evidence to adapt future teaching to students' learning needs (Black & Wiliam, 2009; Clark, 2011). As a reliable evaluation of formative assessment effects requires a clear conceptualisation of specific formative assessment strategies (McMillan et al., 2013; Rakoczy et al., 2019; Yan & Pastore, 2022), we define formative assessment strategies based on Wiliam and Thompson's (2008) framework of five key strategies and one 'big idea'. The five formative assessment strategies are:

1. clarifying and sharing learning intentions and criteria for success;
2. engineering effective classroom discussions, questions and learning tasks;
3. providing feedback that moves learners forward;

4. activating students as instructional resources for one another; and
5. activating students as the owners of their own learning. (p. 64)

Clearly stated learning intentions and criteria for success help students understand the expected learning outcomes, thereby supporting their focus on their learning target (Reed, 2002). Effective classroom discussions, questions and learning tasks engage students to participate, thereby eliciting evidence about their current learning progress. Both strategies reflect and provide information about the learning gap between students' current and expected performance levels. Hence, they could be labelled as clarifying goals and monitoring progress that is a basis for further formative assessment activities (Torrance & Pryor, 2001).

Hypothesis 1 *Clarifying goals and monitoring progress is positively correlated with reading achievement.*

Using the information provided by learning goals and progress, especially the evidence of the gaps between students' actual vs. target knowledge/skills, a teacher can provide suitable feedback to help them take progressive steps toward their learning goals (Black & William, 2018).

Hypothesis 2 *Providing feedback is positively correlated with reading achievement.*

The 'big idea' is adjusting instruction based on diagnostic information about students' learning progress. Diagnostic information enables teachers to adapt their teaching methods and pace according to students' needs so that the teaching is more likely to move each student in their zone of proximal development (Vygotsky, 1978) and enhance their likelihood of success (Yeh, 2010).

Hypothesis 3 *Instructional adjustments are positively correlated with reading achievement.*

Although other formative assessment strategies and instructional adjustments can occur in sequence or simultaneously, the diagnostic information about students' learning progress and feedback (a) are prerequisites for teachers' meaningful decision-making during instructional adjustments (Connor et al., 2004) and (b) help teachers adapt their teaching to their students' needs, which yields better student learning outcomes (Zhai et al., 2018).

Hypothesis 4 *The relationships between clarifying goals and monitoring progress/providing feedback and reading achievement are mediated by instructional adjustments.*

Formative assessment and academic performance

Review studies show that formative assessment often improves learning outcomes (e.g. Black & Wiliam, 1998; Kingston & Nash, 2011). However, McMillan et al. (2013) criticised previous review studies, such as Kingston and Nash's (2011) review, for their (a) inadequate attention to the methodological quality of each study and (b) insufficient identification of the specific type of formative assessment in each study. As inadequate descriptions of the formative assessment designs and interventions hinder the proper evaluation of them, researchers should clearly conceptualise formative assessment strategies in their studies (McMillan et al., 2013; Yan et al., 2021).

Among the five key formative assessment strategies in Wiliam and Thompson's (2008) framework, researchers studied feedback most intensively. Although feedback can increase students' learning outcomes, its specific type of feedback and how it is given affect its effectiveness (Golke et al., 2015; Hattie & Timperley, 2007). According to Graham et al.'s (2015) meta-analysis of formative feedback on children's writing in grades 2–8, the effect sizes of different types of feedback varied (from adults, 0.87; self, 0.62; peer, 0.58; computer, 0.38). Compared with simple corrective feedback, elaborative feedback (which contains new instruction in addition to correctness) in computer-based environments led to higher learning gains, with effect sizes ranging from −0.78 to 2.29 (in van der Kleij et al.'s, 2015 meta-analysis).

There is also evidence of the positive effect of clarifying goals and monitoring progress. For example, Rust et al. (2003) developed a two-year intervention for 615 undergraduate students aiming to increase their knowledge of learning goals and assessment criteria. The results showed that students receiving the intervention had significantly better performance in their assessed coursework than their peers. Furthermore, the information provided by students' learning progress helped teachers refine their teaching which, in turn, resulted in better student learning outcomes (Zhai et al., 2018).

Past studies also show that formative assessment increases students' reading achievement. Feedback on learning from text showed a positive effect ($g = 0.35$) on reading achievement from primary students to university students (Swart et al., 2019). Likewise, controlled experiments showed that real-time formative assessment of learners' vocabulary knowledge and reading comprehension, supported via software, yielded higher vocabulary test scores and reading comprehension test scores ($d > 1$ in experiments 1 and 2 of Ponce et al., 2018). Another controlled experiment showed that students receiving formative assessment intervention through a collaborative reading annotation system outperformed students without formative assessment on reading comprehension, with larger effects for lower achievers (Chen et al., 2019).

Furthermore, formative assessment was both directly linked to reading achievement ($\beta = 0.085$; $p < 0.01$) and indirectly related to it ($\beta = 0.105$; $p < 0.01$) via the teacher–student relationship and attitude toward reading, according to a study of US data from the 2009 PISA (Li, 2016). However, Li (2016) combined all three formative assessment strategies (classroom questioning, providing feedback, and sharing criteria with learners) into one latent trait (from nine survey questions) rather than testing for their different links to reading achievement—as McMillan et al. (2013) advocated. Second, PISA 2009 omitted a critical component of formative assessment: instructional adjustments (Heritage, 2012; Wiliam & Leahy, 2007). Third, PISA 2009 only had student data with no teacher survey data. Fourth, Li (2016) only used US data, ignoring the PISA 2009 data from all other countries/regions.

However, differences in students' vs. teachers' perceptions of teacher use of formative assessment might affect students' changes in their learning motivations, processes or outcomes (e.g. Jónsson et al., 2018; van der Kleij, 2019). In the simple case, the teacher only makes minimal, ineffective changes to her teaching and does not substantially improve students' learning motivations, processes or outcomes; although the teacher believes she is exerting great effort to apply formative assessment, her students correctly evaluate her as not substantially changing her teaching or their learning activities. In the more complicated case, the teacher successfully applies formative assessment, and students learn more, but she does not communicate this effectively. In this case, the poor communication might reflect a poor teacher–student relationship that hinders the impact of formative assessment on learning outcomes (Li, 2016).

Controlling for other factors that might affect reading performance

As past studies have shown that several explanatory variables are linked to reading test scores, we include them to reduce *omitted variables bias* (Kennedy, 2008). Students in countries with more national income (real gross domestic product, GDP, per capita) or less household inequality than others were both linked to reading test scores (Chiu et al., 2017). Also, girls, native language speakers or students in higher socioeconomic status families often have higher reading test scores than other students do (Chiu & McBride-Chang, 2010). Students in schools with smaller class sizes (Ding & Lehrer, 2011) or higher socioeconomic status (SES) schoolmates (Chiu & Klassen, 2009) than others have higher reading test scores on average. Lastly, students taught by female teachers (Lam et al., 2010) or teachers with more years of experience (Munoz et al., 2011) outperformed other students in reading.

This study

Based on the existing literature and above discussion, the present study tests the above four hypotheses to examine the relationship between formative assessment and reading achievement. Specifically, we use PISA 2018 student and teacher data from 19 countries/regions, focusing on three formative assessment strategies: clarifying goals and monitoring progress, providing feedback, and instructional adjustments.

METHODS

This study examines whether formative assessment practices are related to reading achievement among 151,969 fifteen-year-olds in 5,225 schools in 19 countries/regions: Albania, Baku (Azerbaijan), Brazil, Chile, Chinese Taipei, Dominican Republic, Germany, Hong Kong, Korea, Macau, Malaysia, Morocco, Panama, Peru, Portugal, Spain, United Arab Emirates, the UK and the USA (OECD, 2018). (Other countries did not collect these data.)

Data

In each country/economy, the Organisation for Economic Co-operation and Development (OECD, 2018) chose at least 150 representative schools based on neighbourhood SES and student intake, and sampled at least 35 fifteen-year-olds from each school (stratified sampling) and 10 teachers eligible to teach these students reading (for schools with fewer than 10 such teachers, all such teachers were included). The OECD excluded students who were mentally incapable, refused to take the exam, could not physically take it or did not speak the test language (less than 5% of the sample). Participating students completed a 2-hour assessment booklet and then a 30–40-min questionnaire. In addition to OECD-PISA data (2018), we also used economic data (national income and inequality) for each country (World Bank, 2018).

Statistical power differs across levels

For $\alpha = 0.05$ and a small effect size of 0.1, statistical power for both 151,969 students and 5,225 schools exceeds 0.99 but is very low for the 19 countries/regions (Konstantopoulos, 2008).

For country-level explanatory variables, the likelihood that a non-significant result is a false negative is high, but we retain our usual confidence in our significant results (Kennedy, 2008).

Outcome, explanatory and control variables

International experts from OECD and non-OECD countries/regions defined *reading achievement*, built assessment frameworks, created 140 test items, forward- and backward-translated them and pilot tested them to check their validity and reliability (for details and sample items, see OECD, 2018 and www.pisa.oecd.org). OECD (2018) defined reading achievement as the ability to understand, use and reflect on written texts. The reading tasks on the test varied along three dimensions: text form, aspects of reading and usage. The text forms included both continuous prose (such as narration, exposition and argumentation) and other texts (such as lists, forms, graphs and diagrams). The test items assessed five aspects of reading: retrieving information, general understanding of the texts, interpreting them, evaluating their content and evaluating their form. These test items varied according to their context of use: private (e.g. letter, biography), public (e.g. announcements), work (e.g. manuals) and educational (e.g. textbooks).

PISA 2018 used *adaptive testing* for reading assessment. Students answered the same first set of questions, and then if they performed well, they often received harder questions. If they performed poorly, they often received easier questions. This adaptive testing increases the accuracy of both lower scores and higher scores.

Country

Country measures include economic production and inequality. *Log (gross domestic product per capita)* is the log of the market value of all final goods and services made within the borders of a country in a year divided by its total population (World Bank, 2018). (Linear GDP per capita did not fit the data as well as log GDP per capita.)

GDP Gini (a measure of inequality) is defined as the integral of the cumulative distribution function of a perfectly equal income country minus the integral of the cumulative distribution function of the actual country's income (World Bank, 2018). Scores range from 0 (perfect equality; everyone has equal income) to 100 (perfect inequality; one person has all the income and everyone else's income is zero).

Student

Student demographics include gender, native language speaker and SES. *Girl* has a value of one for a female student and zero otherwise. *Native language speaker* is a dichotomous variable with a value of one for a native speaker of the nation's official language and zero otherwise.

Socioeconomic status is the standardised index from a congeneric *multilevel confirmatory factor analysis* (ML-CFA, Muthén & Muthén, 2018) of mother's years of schooling, father's years of schooling and highest parent job status (Ganzeboom et al., 1992). Socioeconomic status is standardised (m = 0; SD = 1).

School

School variables capture teacher gender, class size and mean SES. *Percentage of female teachers* is the proportion of female teachers in a school. *Class size* is the mean number of students in a class for a school. *School mean SES* is the mean SES of all sampled students within a school. When controlling for student SES, *School mean SES* can be interpreted as essentially schoolmates' mean SES. (As the OECD only collected the SES of students in their sample rather than the entire school, we calculated the mean SES of the sample of students in each school [rather than that of students in the entire school]).

All language teacher-related variables are school-level variables, indicating the mean values of language teacher responses (Bonneville-Roussy et al., 2019). These variables include *percentage of female teachers*, *mean years of school teaching experience* and *percentage of teachers receiving any formative assessment training*.

The following language teaching practices within a school are standardised indices created from the congeneric ML-CFA (Muthén & Muthén, 2018) of language teacher responses to survey questions (these standardised indices have an absolute minimum of 1 and an absolute maximum of 4, corresponding to their survey extremes, to facilitate reader interpretation). *Clarifying goals and monitoring progress (teacher)* is from responses to four questions about learning goals and progress for each lesson (see Appendix Table A1 for questions). *Providing feedback (teacher)* is from responses to three questions about giving feedback to students. *Instructional adjustments (teacher)* is from responses to three questions about adapting teaching to student needs.

Student views of teaching practices

Similarly, we created three parallel indices based on student responses to the questions in Appendix A, in which 'I' was replaced with 'my teacher': *clarifying goals and monitoring progress (student)*, *providing feedback (student)* and *instructional adjustments (student)*.

To capture the students' perceived prevalence of such classroom practices within a school, we computed the mean for each school of these student indices: *school mean clarifying goals and monitoring progress (students)*, *school mean providing feedback (students)* and *school mean instructional adjustments (students)*. When entering both a student variable and its school mean variable into a regression, the latter's regression coefficient indicates how this mean attribute of schoolmates is linked to the outcome.

To indicate the students' perceived differences among such classroom practices within a school, we computed the standard deviation (SD) for each school of these student indices: *SD clarifying goals and monitoring progress*, *SD providing feedback* and *SD instructional adjustments*.

To capture perceived differences in these classroom practices between teachers and students, we computed the student value minus the teacher value for (a) *clarifying goals and monitoring progress difference*, (b) *providing feedback difference* and (c) *instructional adjustments difference*.

Data analysis

Analytical issues and statistics strategies

Suitable analyses of these data must address issues involving data, outcomes and explanatory variables (see Table 1). Data issues include representative data, sampling error, missing data, many questions to cover extensive reading content and survey measurement

TABLE 1 Statistics strategies to address each analytic difficulty

Analytic difficulty	Statistics strategy
Dataset	
• Representative data	• Weight sample to reflect population (Kennedy, 2008)
• Sampling error	• Plausible values (Monseur & Adams, 2009)
• Missing data (01??10,011)	• Markov Chain Monte Carlo multiple imputation (Peugh & Enders, 2004)
• Many questions to cover extensive reading content	• Overlapping subtests with anchor items (Embretson & Reise, 2013)
• Measurement errors on surveys	• Factor analysis (Muthen & Muthen, 2018) • Multigroup Rasch (Embretson & Reise, 2013)
Outcome variables	
• Nested data (students within schools within countries …)	• Multilevel analysis (aka Hierarchical linear modelling, Goldstein, 2011)
Explanatory variables	
• Indirect, multilevel mediation effects ($X \rightarrow \mathbf{M} \rightarrow Y$)	• Multilevel M-test (MacKinnon et al., 2004)
• Cross-level interactions (school × student)	• Random effects model (Goldstein, 2011)
• Cross-level correlations	• Center explanatory variables around country mean
• Many hypotheses' false positives	• Two-stage linear step-up procedure (Benjamini et al., 2006)
• Compare effect sizes ($\beta_1 > \beta_2$?)	• Lagrange multiplier tests (Bertsekas, 2014)
• Consistency of results across datasets (robustness)	• Analyses of subsets of the data (Kennedy, 2008) • Original (not estimated) data

error. To create a *representative* sample, we *weight* the data to reflect the student population (Kennedy, 2008). Analysis of *plausible values* reduces *sampling error* (Monseur & Adams, 2009).

As *missing data* (4%) can bias results, reduce estimation efficiency or complicate data analyses, we estimate missing data with *Markov Chain Monte Carlo multiple imputation*, which outperforms *listwise deletion, pairwise deletion*, *mean substitution* and *simple imputation* according to computer simulations (Peugh & Enders, 2004). The Little (1988) test result ($p = 0.88$) suggests that the data was *missing completely at random* (MCAR). (A true MCAR test requiring follow-up interviews of respondents was too costly.)

Students received *subtests* (overlapping subsets of all multiple-choice and open-ended questions) for wider coverage of reading skills while reducing student fatigue and test-learning effects (*balanced incomplete block* test, Baker & Kim, 2004; OECD, 2018). A *graded response Rasch* model of these subtests measured the difficulty of each test item to estimate each student's reading competence more precisely (Embretson & Reise, 2013).

To reduce survey measurement error, we used multiple questions (e.g. parents' educations and jobs) for each construct (e.g. SES) to create a precise index via an ML-CFA (Muthén & Muthén, 2018) and a *multigroup Rasch model* (Embretson & Reise, 2013). To assess the fit of the ML-CFA, we used the comparative fit index (CFI), Tucker–Lewis index (TLI), standardised root mean square residual (SRMR) and root mean square error approximation (RMSEA), which minimised type I and type II errors under many conditions in Hu and Bentler's (1999) simulations. The fit thresholds are: *good* (CFI and TLI > 0.95; SRMR < 0.08; RMSEA < 0.06) and *moderate* (0.90 < CFI and TLI < 0.95; 0.08 < SRMR < 0.10; 0.06 < RMSEA < 0.10).

The multigroup Rasch models for each item in each country yielded similar parameters, indicating measurement equivalence across countries/regions (May, 2006). (Unlike factor

analysis, a multigroup Rasch model has the advantages of requiring only one invariant anchor item across countries/regions and modelling heterogeneous use of the ordinal rating scale; Rossi et al., 2001.) All anchor items showed acceptable discrimination ($\alpha > 0.50$), with threshold parameters within the expected range ($-3 < \beta < 3$). Other studies also showed consistent questionnaire responses and participant understandings across countries/regions (Brown et al., 2005; Schulz, 2003).

Regarding outcomes, students in the same school within the same country probably resemble one another more than those in different schools in different countries/regions (*nested data*). As an ordinary least squares regression underestimates the *standard errors*, we use a *multilevel analysis* (Goldstein, 2011; also known as *hierarchical linear modelling*, Bryk & Raudenbush, 1992).

Explanatory variable issues include indirect effects, cross-level interactions, cross-level correlations, many hypotheses' false positives, effect size comparisons and robustness. Separate, single-level tests of indirect mediation effects on nested data can bias results, so we test for multilevel mediation effects with a *multilevel M-test* (MacKinnon et al., 2004).

With nested data, incorrectly modelling interaction effects across levels (e.g. school size × student gender) can bias the results, so we use a *random effects* model (Goldstein, 2011). If the regression coefficient of an explanatory variable (e.g. $\beta_{yvj} = \beta_{yv0} + f_{yvj}$) differs significantly across levels ($f_{yvj} \neq 0$?), then *cross-level moderation* might exist, and we model the regression coefficient with structural variables (e.g. class size). As cross-level correlations between continuous student-level explanatory variables and country-level random effects might yield biased regression coefficients, we remove this bias by *centring* such variables around their country means (Kennedy, 2008).

As testing many hypotheses increases the possibility of a false positive, we reduce its likelihood via the *two-stage linear step-up procedure*, which outperformed 13 other methods in computer simulations (Benjamini et al., 2006).

When testing whether the effect sizes of explanatory variables differ, *Wald* and *likelihood ratio* tests do not apply at boundary points. Hence, we use *Lagrange multiplier tests* which apply to the entire dataset and show greater statistical power than Wald or likelihood ratio tests for small deviations from the null hypothesis (Bertsekas, 2014).

Lastly, we test whether the results remain stable despite minor changes in the data or analyses (*robustness*, Kennedy, 2008). First, we run subsets of the data separately. Then, we repeat the analyses for the original, unestimated data.

Explanatory model

We model each student's reading test score with a *multilevel analysis variance components* model to test for significant differences at each level: student, school, and country (Goldstein, 2011).

$$\mathbf{Reading}_{ijk} = \beta_0 + e_{ijk} + f_{jk} + g_k \quad (1)$$

The **Reading** test score of student *i* in school *j* in country *k* has a grand mean intercept β_0, with unexplained components (*residuals*) at the student, school and country levels (e_{ijk}, f_{jk}, g_k).

We enter explanatory variables in sequential sets to estimate the variance explained by each set and to test for mediation effects (Kennedy, 2008).

$$\mathbf{Reading}_{ijk} = \beta_0 + e_{ijk} + f_{jk} + g_k + \beta_t \mathbf{Country}_k + \beta_{ujk} \mathbf{Student}_{ijk} + \beta_{vk} \mathbf{School}_{ijk} + \beta_{xjk} \mathbf{Student\_views\_of\_Teaching\_practice}_{ijk} + \beta_{zjk} \mathbf{Interactions}_{ijk} \quad (2)$$

Country variables (real GDP per capita, GDP GINI) can affect all of the people, so we enter them first. A *nested hypothesis test* (χ^2 log likelihood) indicates whether each set of explanatory variables is significant (Kennedy, 2008). As omitting *non-significant* variables does not cause *omitted variable bias*, we safely remove them to increase precision and reduce *multicollinearity* (Kennedy, 2008).

Next, we enter **Student** variables (SES, girl, native language speaker). As families and students often select their schools, we enter **School** variables afterwards (school mean SES, class size, percentage of female teachers, mean years teaching experience, percentage of formative assessment training, learning goals and progress [teacher], feedback [teacher] and instructional adjustments [teacher]). As teachers' perceptions and implementations of teaching practices affect student perceptions of them, we enter **Student_views_of_Teaching practice** (learning goals and progress [student], feedback [student] and instructional adjustments [student]; school mean learning goals and progress [student], school mean feedback [student] and school mean instructional adjustments [student]; SD learning goals and progress [student], SD feedback [student] and SD instructional adjustments [student]; and learning goals and progress difference, feedback difference and instructional adjustments difference). We also analyse residuals for influential outliers.

RESULTS

First, we report the confirmatory factor analyses results, followed by the summary statistics. Then, we discuss the results of the explanatory model.

Indices

The ML-CFA of SES and both student's and teacher's responses to the three sets of formative assessment questions (learning goals and progress, feedback, and instructional adjustments) all showed (a) acceptable reliabilities at both school and student levels and (b) good fits to the data (see Table 2, for factor loadings, see Appendix Table A2).

Summary statistics

This sample's countries ranged from poor, unequal nations (e.g. Morocco) to richer, more equal ones (e.g. Germany). See Table 3 for summary statistics (see Appendix Table A3 for correlation–variance–covariance matrices). Nearly 95% of the teachers in this study received formative assessment training. Most teachers reported clarifying goals and monitoring progress in some lessons (m = 1.499) but gave feedback or adapted their teaching in many lessons (m = 3.193 and 3.233, respectively). However, students reported that their teachers used less feedback or instructional adjustments (means of 2.422 and 2.588; differences of −0.771 and −0.645, respectively).

Explanatory model

Reading test scores mostly differed across students (50%) rather than across schools (23%) or across countries/regions (28%, see Table 3). All of the results discussed below describe first entry into the regression, controlling for all previously included explanatory variables. Ancillary regressions and statistical tests are available upon request.

Country, student, school and teaching practice attributes were linked to students' reading scores. Students in countries/regions with higher real GDP per capita had higher reading scores (see Table 4, model 1, top). This variable accounted for over 17% of the differences

TABLE 2 Goodness of fit statistics for confirmatory factor analyses

Factor	Level	R_c	α	SRMR	CFI	TLI	RMSEA	χ^2	d.f.	p
SES	Student	0.689	0.762	0.000	1.000	1.000	0.002	5.5	4	0.237
	School	0.980		0.031						
Teacher										
Learning goals and progress	Student	0.809	0.813	0.014	0.990	0.971	0.050	280.4	4	0.000
	School	0.979		0.061						
Feedback	Student	0.863	0.852	0.001	1.000	1.000	0.000	1.2	4	0.876
	School	0.962		0.013						
Instructional adjustments	Student	0.681	0.714	0.003	1.000	1.000	0.002	4.3	4	0.364
	School	0.933		0.023						
Student										
Learning goals and progress	Student	0.779	0.801	0.012	0.992	0.975	0.037	828.0	4	0.000
	School	0.958		0.025						
Feedback	Student	0.865	0.831	0.000	1.000	1.000	0.001	4.7	4	0.320
	School	0.964		030						
Instructional adjustments	Student	0.767	0.776	0.000	1.000	1.000	0.000	2.6	4	0.619
	School	0.952		0.038						

Note: R_c, reliability coefficient; α, Cronbach's alpha; SRMR, standardised root mean square residual; CFI, comparative fit index; TLI, Tucker–Lewis index; RMSEA, root mean square error approximation; d.f., degrees of freedom.

in reading test scores, including over 63% of the differences across countries/regions (see Table 4, model 1, bottom).

Student gender, SES, and native language were all linked to reading scores. Girls outscored boys (β = 0.096; see Table 4, model 2, top). Students with higher SES than other students had higher reading scores (β = 0.052). Also, native speakers of the national language outscored other students (β = 0.073). These variables accounted for about 5% of the variance in reading scores (see Table 4, model 2, bottom).

School mean SES and proportion of female teachers were also linked to reading scores. Students with higher SES schoolmates than others had far higher reading scores (β = 1.280; see Table 4, model 3, middle). Also, students in schools with a higher proportion of female teachers had higher reading scores (β = 0.052). These variables accounted for over 5% of the variance in reading scores (see Table 4, model 3, bottom).

The formative assessment practices of clarifying goals and monitoring progress and instructional adjustments were also linked to students' reading scores. When schoolmates reported greater teacher clarifying goals and monitoring progress, students had higher reading scores, suggesting that schoolmates' peer culture (rather than each individual student's view) drives this result (β = 0.158; see Table 4, model 4, middle). Also, students who perceived that their teachers used more instructional adjustments outperformed other students in reading (β = 0.060). Likewise, in schools whose students reported greater instructional adjustments, a student had a higher reading score (β = 0.139). (Note that the school mean of instructional adjustment is substantially negatively correlated with the school mean of learning goals and progress [r = −0.62, see Appendix Table A3], suggesting that having all teachers in a school consistently doing both is very difficult.) Notably, when teachers and students' reports of instructional adjustments differed, students had far lower reading scores

TABLE 3 Summary statistics (N = 151,696)

Variable	Mean	SD	Minimum	Maximum
Reading test score	453.461	107.482	84.05	868.87
Country				
Real GDP per capita	26112.169	15264.377	3361.2	58,641.6
GDP GINI	38.059	7.019	26.6	53.9
Student				
Girl	0.497	0.500	0.000	1.000
Socioeconomic status (SES)	0.000	1.000	−2.735	1.738
Native language speaker	0.822	0.383	0.000	1.000
School				
School mean SES	−0.003	0.580	−2.673	1.385
Class size	30.290	10.208	13.000	53.000
Percentage of female teachers	0.747	0.270	0.000	1.000
Teacher				
Female teacher	0.747	0.270	0.000	1.000
Years' teaching experience	9.716	5.961	0.000	50.000
Training on formative assessment	0.947	0.127	0.000	1.000
Teaching practice				
Learning goals and progress (teacher reported)	1.499	0.351	1.000	4.000
Feedback (teacher)	3.193	0.383	1.274	4.000
Instructional adjustments (teacher)	3.233	0.364	1.000	4.000
Learning goals and progress (student reported)	1.846	0.730	1.000	4.000
Feedback (student)	2.422	0.856	1.000	4.000
Instructional adjustments (student)	2.588	0.792	1.000	4.000
School mean learning goals and progress (students)	1.845	0.303	1.000	2.980
School mean feedback (students)	2.424	0.326	1.125	4.000
School mean instructional adjustments (students)	2.587	0.272	1.225	4.000
Standard deviation (SD) learning goals and progress (among students)	0.661	0.136	0.000	2.053
SD feedback	0.796	0.119	0.000	2.121
SD instructional adjustments	0.749	0.110	0.000	2.121
Learning goals and progress difference (student vs. teacher)	0.347	0.773	−3.000	3.000
Feedback difference	−0.771	0.909	−3.000	2.411
Instructional adjustments difference	−0.645	0.865	−3.000	2.205

(β = −1.238), an exploratory finding worthy of further study. These variables accounted for about 5% of the variance in reading scores (see Table 4, model 4, bottom).

The final model accounted for nearly 33% of the variance in reading scores. All other explanatory variables were not significant. Furthermore, all mediation effects and moderation effects were not significant. Analysis of the residuals showed no significant outliers.

TABLE 4 Summary of unstandardised regression coefficients (with standard errors in parentheses) and standardised regression coefficients of three-level analyses of students' reading test scores

	Reading test score							
Explanatory variable	**Model 1**		**Model 2**		**Model 3**		**Model 4**	
Real GDP per capita	0.003	***	0.003	***	0.002	***	0.002	***
	(0.000)		(0.000)		(0.000)		(0.000)	
	0.380		0.366		0.288		0.251	
Girl			20.590	***	20.590	***	19.400	***
			(0.439)		(0.438)		(0.431)	
			0.096		0.096		0.090	
SES			8.746	***	7.191	***	6.265	***
			(0.243)		(0.247)		(0.244)	
			0.081		0.067		0.058	
Native language speaker			20.600	***	20.800	***	19.820	***
			(0.741)		(0.734)		(0.723)	
			0.073		0.074		0.071	
School mean SES					58.930	***	53.060	***
					(1.280)		(1.289)	
					0.318		0.286	
Percentage of female teachers					20.540	***	20.420	***
					(2.316)		(2.224)	
					0.052		0.051	
Clarifying goals and monitoring progress:							55.940	***
School mean of students							(3.439)	
							0.158	
Instructional adjustments:							8.198	***
Student view							(0.266)	
							0.060	
Instructional adjustments:							54.790	***
School mean of students							(3.084)	
							0.139	
Instructional adjustments:							−153.800	***
Differences in students'							(2.577)	
and teachers' views							−1.238	
Variance at each level	Explained variance at each level							
Country (27%)	0.633		0.670		0.619		0.706	
School (23%)	0.000		0.124		0.415		0.465	
Student (50%)	0.000		0.024		0.024		0.053	
Total variance explained	0.174		0.224		0.277		0.327	

DISCUSSION

The effect of background factors

Background factors at the country/economy, school and student levels were linked to students' reading scores. Students in countries/regions with higher national incomes (real GDP per capita) had higher reading scores, concurring with international studies of primary and secondary school students (e.g. Chiu et al., 2017; Chiu & Chow, 2010). Also, students from higher SES families or with higher SES schoolmates outscored other students. These results support the view that students with access to more resources, especially education resources, have more learning opportunities, capitalise on them more often to learn more and show higher reading test scores (*resource provider* hypothesis; Chiu & McBride-Chang, 2010). Girls outperformed boys in reading, consistent with previous PISA results (e.g. Brozo et al., 2014). Not surprisingly, native speakers of the testing language had higher reading scores than other students, similar to past studies (e.g. Chiu & McBride-Chang, 2010). Also, students in schools with a higher proportion of female teachers had higher reading scores, consistent with some studies showing that female teachers teach differently from male teachers to yield higher student learning outcomes (e.g. Lam et al., 2010); however, other studies showed no teacher gender effect on student learning outcomes (e.g. Carrington et al., 2008; Holmlund & Sund, 2008), suggesting instruction-specific or context-specific effects. As PISA includes several countries (e.g. Albany, Azerbaijan, Qatar, UAE) that run single-gender schools with teachers of matching gender, an alternative explanation for female teachers' superior teaching results is that they teach girls. However, studies examining the effects of same-gender teachers on student achievement generate mixed findings: some revealed positive effects of having female teachers for female students (e.g. Hwang & Fitzpatrick, 2021), while others (e.g. Cho, 2012) did not. Future studies can scrutinise how teaching patterns differ across teacher gender, student gender or their interaction and how they influence students' learning outcomes.

The effect of clarifying goals and monitoring progress

A student had higher reading scores only when schoolmates reported greater teacher clarifying goals and monitoring progress—not when the student reported them—thereby supporting Hypothesis 1 only at the school level. This finding is consistent with the view that goal clarification provides benchmarks for students' learning (Torrance & Pryor, 2001), while continuous progress monitoring helps students identify gaps between their current and expected competence and modify their learning strategies or exert more effort to narrow the gap (Rust et al., 2003; Yan et al., 2021). Furthermore, this result also suggests that clarifying goals and monitoring progress across multiple classes—not simply clarifying and monitoring by a student's teacher—drives effective formative assessment collectively (not individually).

Future studies can determine whether one or more of the following mechanisms account for this collective link: (a) more teacher colleagues engaged in goal clarification and progress monitoring mutually support one another to do so more effectively than doing it alone; (b) a student's greater exposure to more teachers' goal clarification and progress monitoring aids understanding and use of them; or (c) greater schoolmates' exposure to teachers' goal clarification and progress monitoring supports a student's understanding and use of information generated by formative assessment. Also, future intervention studies can determine whether formative assessment training (especially on clarifying goals and monitoring progress) of all of a school's teachers of reading (especially together) is much more effective than

comparable training of some teachers from different schools at improving students' reading performances.

The effect of providing feedback

Providing feedback was not linked to students' reading achievement, showing no support for Hypothesis 2. We have four possible accounts for this result: (a) poor feedback; (b) greater importance of other aspects of formative assessment (clarifying goals and monitoring progress, instructional adjustments); (c) unknown use of feedback by students; and (d) more teacher feedback to weaker students. Although feedback is one of the most powerful influences on learning and achievement, feedback with low-quality content or poor implementation does not aid student learning (Hattie & Timperley, 2007).

Student reports of teacher feedback were negatively correlated with their reports of teacher clarifying goals and monitoring progress (r = −0.42), and the latter was positively linked to superior reading performance. The negative correlation indicates that these students perceived that many teachers either shared learning goals or gave students feedback but not both. Together, these results seem to suggest the greater importance of teacher's goal clarification and progress monitoring over teacher feedback for students' reading performance.

Also, feedback might be necessary but insufficient without other formative assessment aspects (e.g. instructional adjustments). Students' reports of teacher feedback and of instructional adjustments were positively correlated (r = 0.55), indicating that both often occurred in these students' classrooms. This result is consistent with the view that teacher feedback requires accompaniment by instructional adjustments to improve student learning (McMillan et al., 2013).

To capitalise on high-quality feedback, students must make sense of feedback information and actively use it to enhance their learning strategies to learn more and show superior academic performance (Carless & Boud, 2018; Hattie & Timperley, 2007). Students' low commitment to processing and using feedback may reduce the effect of feedback (Golke et al., 2015; Yan & Carless, 2021). As PISA 2018 did not ask students to report their uptake or use of teacher feedback, we could not model this part of the process. These three interpretations of the non-significant feedback suggest that teacher feedback alone does not guarantee improved student learning (Shute, 2008).

As analysis of correlational datasets like PISA cannot identify the direction of causation, the results might also show reverse causation. Specifically, as teachers' formative assessment practice is influenced by students' characteristics (Yan et al., 2021), students with lower reading performance might receive more teacher feedback than students with higher reading performance. This phenomenon might neutralise the positive link between providing feedback and reading achievement.

The effect of instructional adjustments

Instructional adjustments were linked to student reading performance in three ways: a student's report of his or her teacher's instructional adjustment; schoolmates' reports of their teachers' instructional adjustments; and discrepancies between students' and teachers' reports of instructional adjustments. Students who reported greater instructional adjustments by their teachers outperformed other students in reading; this result supports Hypothesis 3. Also, this result is consistent with the view that instruction adjusted according to students learning needs aligns each student with his or her own zone of proximal development (Vygotsky, 1978) to enhance the efficiency of teaching and learning (Wiliam & Leahy, 2007).

The significant instructional adjustment result also highlights the importance of formative assessment not only for students but also for teachers (Black et al., 2003).

Also, when schoolmates reported more instructional adjustments by their teachers, a student showed greater reading performance, supporting Hypothesis 3 at the school level. Like the learning goals result, this result suggests that teachers' instructional adjustments across multiple classes within a culture that supports it aid effective formative assessment and improve student reading performance. Possibly, more teacher colleagues engaged in instructional adjustments mutually support one another to do so more effectively than doing it alone; future studies can test this hypothesis. Furthermore, the high negative correlation between the school means of learning goals and instructional adjustment (−0.62) suggests the difficulty of having all teachers within a school successfully doing both. Future studies can determine whether schools simply invest in professional development to focus on one aspect or whether teachers eventually learn both after sufficient professional development. In addition, future intervention studies can test whether formative assessment training (including instructional adjustments) of all of a school's teachers (especially together) is much more effective than comparable training of some teachers from different schools.

Furthermore, when students reported fewer instructional adjustments by teachers than their teachers did, students had lower reading scores, consistent with past studies showing (a) differences in student vs. teacher perceptions of teachers' formative assessment (e.g. Jónsson et al., 2018; van der Kleij, 2019) and (b) bigger differences yielding lower reading scores. We offer two possible interpretations: inferior instructional adjustments and poor communication. Some teachers only make superficial or poor instructional adjustments that correspondingly do not change or inadequately change students' learning processes, so students correctly devalue them, are less likely to apply these formative assessment-informed changes, feel less motivated and learn less than otherwise, consistent with past studies (e.g. Gamlem & Smith, 2013; Havnes et al., 2012). Or teachers might communicate poorly with students, so that the students do not appreciate the teacher's efforts (e.g. teacher adapts each lesson based on students' performances in previous lessons, but she does not tell them, and students do not notice it). In this case, the instruction adjustments might be effective, but the poor communication and consequent poor teacher–student relationship might reduce students' motivations and learning outcomes. Future studies can determine the validity of these two speculations.

The mediating effect of instructional adjustments

Ideally, teachers formatively assess students' learning progress to provide feedback to students and then adjust their instruction accordingly to improve student performance. Hence, this study tested for mediation within this sequence but found no significant mediation effect, showing no support for Hypothesis 4. Along with the negative correlation between student reports of teacher clarifying goals and monitoring progress and teacher feedback, this result indicates that teachers do not consistently implement all of these formative assessment processes.

Practical implications

According to this study, teacher feedback alone was not linked to reading achievement, whereas schoolmates' perceived teacher clarifying goals and monitoring progress and adjusting instruction (student perception, schoolmates' perceptions and student–teacher

perception differences) were all linked to reading achievement. These findings suggest several practical implications regarding enacting formative assessment on student learning.

The non-significant association between teacher providing feedback and students' reading achievement highlights the complexity of the feedback process. It suggests a potential gap between feedback theory and teachers' feedback practices in classrooms. Although the literature has long suggested that designing and providing high-quality feedback (e.g. timely, detailed, with concrete instructional guidance, etc.) is effective in enhancing student learning (Hattie & Timperley, 2007), this study showed that feedback alone is insufficient; instead, accompanying it with clarifying goals and monitoring progress and instructional adjustments are vital. These latter two were especially linked to higher student reading scores. Also, past studies showed that students' perceptions of formative assessment matter (Guo & Yan, 2019; Yan & Brown, 2021) and their uptake of feedback (not available in PISA 2018 data) is crucial for improving their learning outcomes (Carless & Boud, 2018), so teachers need to consider how to promote students' willingness and capacities to use feedback to improve their academic performance.

The positive link between perceived instructional adjustments and student reading scores holds at both the individual and school levels. This finding highlights how instructional adjustments might bridge other formative assessment strategies and enhance student learning. Formative assessment informs teachers' decisions as they adjust their instruction. Hence, these results suggest appropriate teacher professional development to improve teachers' accurate interpretation of formative assessment results and appropriately adjust their instruction. As teachers' instructional adjustments within the same school collectively drive effective formative assessment and improve student reading achievement, professional learning communities on formative assessment within a school might mutually benefit its members. Such communities could help develop a school culture fostering professional exchange and collaboration on effective formative assessment practices.

Limitations and future directions

The limitations of the PISA data include cross-sectional data, school-linked student and teacher reports, few formative assessment questions and partial measurement invariance. As PISA data are cross-sectional, we can only determine correlations, not causation (Forestier & Adamson, 2017). This study tested our theoretical model and hypotheses regarding how formative assessment might be linked to reading achievement. However, students with low reading performance might elicit more formative assessment activities from their teacher. For example, teachers might provide more feedback to lower achievers, as discussed earlier. Hence, we echo Li's (2016) call for studies with longitudinal or experimental designs on the causal relationship between formative assessment and reading achievement.

The PISA data only contain school-linked students' and teachers' reports of teachers' formative assessment practices, not their objective actions. While collecting both students' and teachers' reports allows triangulation of their data, this study found discrepancies between student and teacher reports on instructional adjustments. Thus, collecting objective data (such as classroom videotapes) would be preferable for future studies. These students and teachers reported on their school's language teaching practices, so future studies ask for student and teacher views of the same lesson(s) by the same teacher.

The PISA survey covered only three formative assessment strategies. Notably, the PISA survey lacked questions regarding student-directed formative assessment (e.g. self and peer assessment) corresponding to strategies 4 and 5 in Wiliam and Thompson's (2008) framework. These strategies may encourage students to own and self-regulate their learning, which often enhances their learning outcomes (Yan & Brown, 2017). Also, the survey did

not ask teachers or students about how students used teacher feedback. Future large-scale international studies can include such questions in their surveys to capture a more detailed, nuanced picture of teachers' formative assessment strategies.

The analyses indicated that item parameters did not differ across countries/regions, showing partial invariance; however, person parameters differed across countries/regions, which is also a limitation of this study.

CONCLUSION

This study investigated the link between student- and teacher-reported formative assessment-relevant strategies—clarifying goals and monitoring progress, providing feedback and instructional adjustments—and students' reading achievement, using data from PISA 2018. While teacher feedback alone was not linked to student reading achievement, teacher clarifying goals and monitoring progress throughout a school and adjusting instruction both by a teacher and throughout a school were positively linked to reading achievement. These results emphasise that formative assessment is a multifacet concept, so different formative assessment strategies can have different impacts on student learning. This study also reminds researchers and practitioners that 'conducting formative assessment' and 'making formative assessment useful' are not the same thing. Teachers' implementation of some theoretically sound strategies alone (e.g. providing feedback) does not guarantee expected learning outcomes. Also, discrepancies between students' and their teachers' views of teachers' instructional adjustments were negatively linked to student reading achievement. Hence, teachers who monitor the effectiveness of their formative assessment strategies and revise their instructions accordingly can enact the desirable benefits on learning outcomes.

CONFLICT OF INTEREST

There is no conflict of interest associated with this study.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in PISA 2018 Database at https://www.oecd.org/pisa/data/2018database/

ETHICAL APPROVAL STATEMENT

Ethical approval for this study was granted by The Education University of Hong Kong.

ORCID

Zi Yan https://orcid.org/0000-0001-9305-884X

REFERENCES

Baker, F. B., & Kim, S. H. (2004). *Item response theory*. CRC Press.

Benjamini, Y., Krieger, A. M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*, *93*, 491–507.

Bennett, R. E. (2011). Formative assessment. *Assessment in Education*, *18*(1), 5–25.

Bertsekas, D. P. (2014). *Constrained optimisation and Lagrange multiplier methods*. Academic.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning*. Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*, 7–74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, *21*, 5–31.

Black, P., & William, D. (2018). Classroom assessment and pedagogy. *Assessment in Education*, *25*(6), 551–575.

Bonneville-Roussy, A., Bouffard, T., Palikara, O., & Vezeau, C. (2019). The role of cultural values in teacher and student self-efficacy. *Contemporary Educational Psychology*, *59*, 101798. https://doi.org/10.1016/j.cedpsych.2019.101798

Brown, G., Micklewright, J., Schnepf, S. V., & Waldmann, R. (2005). Cross-national surveys of learning achievement –how robust are the findings? *Journal of the Royal Statistical Society (Series A)*, *170*, 623–646.
Brozo, W. G., Sulkunen, S., Shiel, G., Garbe, C., Pandian, A., & Valtin, R. (2014). Reading, gender, and engagement. *Journal of Adolescent & Adult Literacy*, *57*(7), 520–593.
Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models*. Sage.
Carless, D., & Boud, D. (2018). The development of student feedback literacy. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325.
Carrington, B., Tymms, P., & Merrell, C. (2008). Role models, school improvement and the 'gender gap' – Do men bring out the best in boys and women the best in girls? *British Educational Research Journal*, *34*(3), 315–327.
Chen, C. M., Chen, L. C., & Horng, W. J. (2019). A collaborative reading annotation system with formative assessment and feedback mechanisms to promote digital reading performance. *Interactive Learning Environments*, *29*(5), 848–865.
Chiu, M. M., & Chow, B. W. Y. (2010). Culture, motivation, and reading achievement: High school students in 41 countries. *Learning and Individual Differences*, *20*, 579–592.
Chiu, M. M., Chow, B. W. Y., & Joh, S. W. (2017). Streaming, tracking and reading achievement: A multilevel analysis of students in 40 countries. *Journal of Educational Psychology*, *109*(7), 915–934.
Chiu, M. M., & Klassen, R. M. (2009). Calibration of reading self-concept and reading achievement among 15-year-olds: Cultural differences in 34 countries. *Learning and Individual Differences*, *19*, 372–386.
Chiu, M. M., & McBride-Chang, C. (2010). Family and reading in 41 countries: Differences across cultures and students. *Scientific Studies of Reading*, *14*, 514–543.
Cho, I. (2012). The effect of teacher-student gender matching: Evidence from OECD countries. *Economics of Education Review*, *31*(3), 54–67. https://doi.org/10.1016/j.econedurev.2012.02.002
Clark, I. (2011). Formative assessment. *Florida Journal of Educational & Administration Policy*, *4*(2), 158–180.
Connor, C. M., Morrison, F. J., & Petrella, J. N. (2004). Effective reading comprehension instruction. *Journal of Educational Psychology*, *96*(4), 682–698.
Ding, W., & Lehrer, S. F. (2011). Experimental estimates of the impacts of class size on test scores. *Education Economics*, *19*(3), 229–252.
Dunn, K. E., & Mulvenon, S. W. (2009). A critical review of research on formative assessments. *Practical Assessment & Research and Evaluation*, *14*(7), 1–11.
Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.
Forestier, K., & Adamson, B. (2017). A critique of PISA and what Jullien's plan might offer. *Compare*, *47*(3), 359–373.
Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international Socio-Economic Index of occupational status. *Social Science Research*, *21*(1), 1–56. https://doi.org/10.1016/0049-089X(92)90017-B
Gamlem, S. M., & Smith, K. (2013). Student perceptions of classroom feedback. *Assessment in Education: Principles, Policy & Practice*, *20*, 150–169. https://doi.org/10.1080/0969594X.2012.749212
Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons.
Golke, S., Dörfler, T., & Artelt, C. (2015). The impact of elaborated feedback on text comprehension within a computer-based assessment. *Learning and Instruction*, *39*, 123–136.
Graham, S., Hebert, M., & Harris, K. R. (2015). Formative assessment and writing. *The Elementary School Journal*, *115*(4), 523–547.
Guo, W. Y., & Yan, Z. (2019). Formative and summative assessment in Hong Kong primary schools: Students' attitudes matter. *Assessment in Education: Principles, Policy & Practice*, *26*(6), 675–699. https://doi.org/10.1080/0969594X.2019.1571993
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112.
Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation*, *38*, 21–27. https://doi.org/10.1016/j.stueduc.2012.04.001
Heritage, M. (2012). Gathering evidence of student understanding. In J. H. McMillan (Ed.), *Sage handbook of research on classroom assessment* (pp. 179–196). Sage Publications.
Holmlund, H., & Sund, K. (2008). Is the gender gap in school performance affected by the sex of the teacher? *Labour Economics*, *15*, 37–53.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis. *Structural Equation Modeling*, *6*(1), 1–55.
Hwang, N., & Fitzpatrick, B. (2021). Student–teacher gender matching and academic achievement. *AERA Open*, *7*(1), 1–11. https://doi.org/10.1177/23328584211040058
Jónsson, Í. R., Smith, K., & Geirsdóttir, G. (2018). Shared language of feedback and assessment. Perception of teachers and students in three Icelandic secondary schools. *Studies in Educational Evaluation*, *56*, 52–58. https://doi.org/10.1016/j.stueduc.2017.11.003
Kennedy, P. (2008). *Guide to econometrics*. Wiley-Blackwell.
Kingston, N., & Nash, B. (2011). Formative assessment. *Educational. Measurement*, *30*(4), 28–37.
Konstantopoulos, S. (2008). The power of the test in three-level cluster randomised designs. *Journal of Research on Educational Effectiveness*, *1*, 66–88.

Lam, Y. R., Tse, S. K., Lam, J. W., & Loh, E. K. (2010). Does the gender of the teacher matter in the teaching of reading literacy? Teacher gender and pupil attainment in reading literacy in Hong Kong. *Teaching and Teacher Education*, *26*(4), 754–759.
Li, H. (2016). How is formative assessment related to students' reading achievement? Findings from PISA 2009. *Assessment in Education*, *23*(4), 473–494.
MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect. *Multivariate Behavioral Research*, *39*, 99–128.
May, H. (2006). A multilevel Bayesian IRT method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics*, *31*, 63–79.
McMillan, J. H., Venable, J. C., & Varier, D. (2013). Studies of the effect of formative assessment on student achievement. *Practical Assessment, Research & Evaluation*, *18*(2), 1–15.
Monseur, C., & Adams, R. (2009). Plausible values. *Journal of Applied Measurement*, *10*, 320–334.
Munoz, M. A., Prather, J. R., & Stronge, J. H. (2011). Exploring teacher effectiveness using hierarchical linear models: Student- and classroom-level predictors and cross-year stability in elementary school reading. *Planning and Changing*, *42*(3–4), 241–273.
Muthén, L. K., & Muthén, B. O. (2018). *Mplus 8.1*. Muthén & Muthén.
OECD. (2018). *PISA 2018 Technical Report*. Retrieved November 21, 2020 from https://www.oecd.org/pisa/data/pisa2018technicalreport/
Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research. *Review of Educational Research*, *74*, 525–556.
Ponce, H. R., Mayer, R. E., Figueroa, V. A., & López, M. J. (2018). Interactive highlighting for just-in-time formative assessment during whole-class instruction. *Interactive Learning Environments*, *26*(1), 42–60.
Rakoczy, K., Pinger, P., Hochweber, J., Klieme, E., Schütze, B., & Besser, M. (2019). Formative assessment in mathematics. *Learning and Instruction*, *60*, 154–165.
Reed, D. (2002). Clearly communicating the learning objective matters! *Middle School Journal*, *43*(5), 16–24.
Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity. *Journal of the American Statistical Association*, *96*, 20–31.
Rust, C., Price, M., & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment & Evaluation in Higher Education*, *28*(2), 147–164.
Schulz, W. (2003). *Validating questionnaire constructs in international studies*. Australian Council for Educational Research.
Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*, 153–189.
Swart, E. K., Nielen, T. M. J., & Sikkema-de Jong, M. T. (2019). Supporting learning from text. *Educational Research Review*, *28*, 100296.
Torrance, H., & Pryor, J. (2001). Developing formative assessment in the classroom: Using action research to explore and modify theory. *British Educational Research Journal*, *27*(5), 615–631.
van der Kleij, F. M. (2019). Comparison of teacher and student perceptions of formative assessment feedback practices and association with individual student characteristics. *Teaching and Teacher Education*, *85*, 175–189.
van der Kleij, F. M., Feskens, R. C. W., & Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes. *Review of Educational Research*, *85*(4), 475–511.
Vygotsky, L. S. (1978). *Mind in society*. MIT Press.
Wiliam, D., & Leahy, S. (2007). A theoretical foundation for formative assessment. In J. H. McMillan (Ed.), *Formative assessment* (pp. 29–42). Teachers College Press.
Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction. In C. A. Dwyer (Ed.), *The future of assessment* (pp. 53–82). Erlbaum.
World Bank. (2018). *The world development report 2018*. Oxford University Press.
Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education*, *42*(8), 1247–1262.
Yan, Z., & Brown, G. T. L. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. *Studies in Educational Evaluation*, *68*, 100985. https://doi.org/10.1016/j.stueduc.2021.100985
Yan, Z., & Carless, D. (2021). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment & Evaluation in Higher Education*. https://doi.org/10.1080/02602938.2021.2001431
Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice*, *28*(3), 228–260. https://doi.org/10.1080/0969594X.2021.1884042
Yan, Z., & Pastore, S. (2022). Assessing teachers' strategies in formative assessment: The teacher formative assessment practice scale. *Journal of Psychoeducational Assessment*, *40*(5), 592–604. https://doi.org/10.1177/07342829221075121

Yeh, S. S. (2010). Understanding and addressing the achievement gap through individualised instruction and formative assessment. *Assessment in Education*, *17*(2), 169–182.
Zhai, X., Li, M., & Guo, Y. (2018). Teachers' use of learning progression-based formative assessment to inform teachers' instructional adjustment. *International Journal of Science Education*, *40*(15), 1832–1856.

How to cite this article: Yan, Z., & Chiu, M. M. (2023). The relationship between formative assessment and reading achievement: A multilevel analysis of students in 19 countries/regions. *British Educational Research Journal*, *49*, 186–208. https://doi.org/10.1002/berj.3837

APPENDIX A: ANCILLARY ANALYSES

APPENDIX TABLE A1 Questions on formative assessment strategies

	Questions	Response options
Clarifying goals and monitoring progress		
1	I set clear goals for the students' learning	(a) never or hardly ever; (b) some lessons; (c) many lessons; or (d) every lesson
2	I ask questions to check whether students have understood what was taught	
3	At the beginning of a lesson, I present a short summary of the previous lesson	
4	I tell students what they have to learn	
Providing feedback		
1	I give students feedback on their strengths in my course	(a) never or almost never; (b) some lessons; (c) many lessons; or (d) every lesson or almost every lesson
2	I tell students in which areas they can still improve	
3	I tell students how they can improve their performance	
Instructional adjustments		
1	I tailor my teaching to meet the needs of my students	(a) never or almost never; (b) some lessons; (c) many lessons; or (d) every lesson or almost every lesson
2	I provide individual help when a student has difficulties understanding a topic or task	
3	I change the structure of my lesson on a topic that most students find difficult to understand	

APPENDIX TABLE A2 Factor loadings

	School level			Student level		
Variable	**Factor loading**	**SE**	**Unique**	**Factor loading**	**SE**	**Unique**
S_Learn_ Goals and Progress						
ST102Q01TA	0.327	0.005	0.016	0.697	0.003	0.514
ST102Q02TA	0.305	0.004	0.010	0.724	0.003	0.477
ST102Q03TA	0.397	0.007	0.078	0.635	0.003	0.596
ST102Q04TA	0.302	0.006	0.019	0.666	0.003	0.557
S_Feedback						
ST104Q02NA	0.364	0.005	0.016	0.715	0.002	0.489
ST104Q03NA	0.297	0.003	0.009	0.878	0.002	0.230
ST104Q04NA	0.301	0.004	0.010	0.818	0.002	0.330
S_Instruction_Adapt						
ST212Q01HA	0.232	0.010	0.008	0.739	0.003	0.454
ST212Q02HA	0.254	0.009	0.008	0.736	0.003	0.459
ST212Q03HA	0.237	0.010	0.007	0.690	0.003	0.524
T_Learn_ Goals and Progress						
TC171Q01HA	0.358	0.015	0.003	0.708	0.006	0.499
TC171Q02HA	0.291	0.014	0.022	0.764	0.006	0.417
TC171Q03HA	0.198	0.030	0.053	0.615	0.006	0.623
TC171Q04HA	0.250	0.020	0.019	0.743	0.006	0.448
T_Feedback						
TC202Q06HA	0.354	0.012	0.015	0.704	0.005	0.505
TC202Q07HA	0.303	0.012	0.017	0.886	0.004	0.215
TC202Q08HA	0.356	0.011	0.011	0.788	0.005	0.380
T_Instruction_Adapt						
TC202Q01HA	0.383	0.021	0.014	0.638	0.007	0.593
TC202Q02HA	0.333	0.021	0.051	0.665	0.007	0.558
TC202Q03HA	0.226	0.024	0.032	0.629	0.007	0.604
SES (3 variables)						
FYRSCH	0.559	0.003	0.031	0.670	0.004	0.551
MYRSCH	0.562	0.003	0.009	0.722	0.004	0.479
HISEI	0.563	0.004	0.073	0.487	0.004	0.763

APPENDIX TABLE A3 Outcomes' and significant explanatory variables' correlations, *variances* and covariances along the lower-left, *diagonal* and upper-right matrices

	Variable	1	2	3	4	5	6	7	8	9	10	11
1	Reading test score	**11,552**	561,072	6.69	26.01	5.08	21.22	1.64	9.69	6.01	1.32	10.41
2	Real GDP per capita	0.34	**232,999,680**	−49.92	2829	−172	2807	−241	1989	−510	−505	−467
3	Girl	0.12	−0.01	**0.25**	−0.01	0.00	0.00	0.02	0.00	0.01	0.00	0.01
4	SES	0.24	0.19	−0.02	**1.00**	0.02	0.34	0.00	0.02	0.03	0.03	0.02
5	Native language speaker	0.12	−0.03	0.02	0.04	**0.15**	0.01	0.01	−0.01	0.01	0.01	0.02
6	School mean SES	0.34	0.32	0.00	0.58	0.03	**0.34**	0.00	0.02	0.03	0.03	0.02
7	Percentage of female teachers	0.06	−0.06	0.12	−0.01	0.07	−0.01	**0.07**	0.00	0.00	0.00	0.00
8	Learning goals and progress: school mean	0.30	0.43	−0.01	0.07	−0.07	0.12	−0.03	**0.09**	−0.05	−0.05	−0.04
9	Instructional adjustment: student	0.07	−0.04	0.02	0.03	0.04	0.06	0.01	−0.22	**0.63**	0.07	0.62
10	Instructional adjustment: school mean	0.05	−0.12	0.02	0.10	0.09	0.17	0.02	−0.62	0.34	**0.07**	0.07
11	Different views of Instructional adjustment by teacher vs. students	0.11	−0.04	0.01	0.02	0.06	0.03	−0.01	−0.14	0.91	0.29	**0.75**