education sciences

MDPI

Article

School Assessment Policy, Teacher Assessment Practice and Training, and Reading Achievement: A Multi-Level Analysis of PISA 2018 Data

Zi Yan [1,*], Ming Ming Chiu [2], Jiahe Gu [1], Lan Yang [1] and Ying Zhan [1]

[1] Department of Curriculum and Instruction, The Education University of Hong Kong, Hong Kong, China; s1147986@s.eduhk.hk (J.G.); yanglan@eduhk.hk (L.Y.); zhanying@eduhk.hk (Y.Z.)
[2] Department of Special Education and Counselling, The Education University of Hong Kong, Hong Kong, China; mingchiu@eduhk.hk
* Correspondence: zyan@eduhk.hk; Tel.: +852-2948-6367

Abstract

Grounded in the assessment ecology framework, we examine how assessment components (school assessment policies, teacher assessment practices and training) are linked to the reading achievements of 151,969 students from 19 countries. Analyses of the 2018 PISA survey and test data yielded these results. Schools that posted assessment results for accountability, or teachers who often clarified learning goals, tracked student progress or accordingly adapted their teaching had students with higher reading scores. By contrast, schools mostly using assessment data to evaluate, teachers trained in reading comprehension assessment, or giving more feedback had students with lower reading scores. Students in richer countries or with better relationships with their teachers had higher reading scores. These findings show the complexity and interactions within assessment ecologies that shape learning outcomes.

Keywords: formative assessment; assessment policy; reading achievement; PISA 2018

1. Introduction

1.1. Background

Beyond student attributes (e.g., gender, socio-economic status), country and school attributes are linked to learning outcomes, but no scholar has developed an evidenced ecological model covering all of them. Assessment policies, teacher education, and professional development can shape teachers' assessment attitudes and practices (G. T. Brown & Harris, 2009; Nadelson et al., 2016). When school policies and resources systematically support teachers, formative assessment activities can enhance student learning (Hopfenbeck et al., 2015; Oo et al., 2024). A teacher who adapts teaching to students' assessment results helps them boost their reading outcomes (Konstantopoulos et al., 2013).

Grounded in Chong and Isaacs's (2023) assessment ecology framework, we determine whether assessment policies, teacher assessment practices and training are related to students' reading achievement. Specifically, we analyse the survey responses and test scores of 151,969 students in 19 countries from the 2018 Programme for International Student Assessment (PISA; OECD, 2018). Our findings can evidence a comprehensive model of assessment and learning.

Academic Editor: Vasilis Grammatikopoulos

Received: 5 March 2026
Revised: 12 April 2026
Accepted: 15 April 2026
Published: 20 April 2026

1.2. Assessment Ecology

Informed by Ecological Systems Theory (Mao & Lee, 2022), Chong and Isaacs (2023) proposed the assessment ecology framework to capture the dynamic interplay between learners and their assessment environments. This framework conceptualises the assessment ecology across three interacting dimensions: engagement, context, and learner. The present study focuses on the contextual dimension, exploring how socio-cultural, instructional, and interpersonal factors are linked to student learning outcomes.

1.3. School Assessment Policy Regarding the Uses of Assessment Data

School assessment policies typically use data to improve teaching priority areas, support and improve students' learning, judge their level of achievement in relation to learning outcomes, and hold themselves accountable. *Assessment for learning* is any evaluative process that aims to improve student learning. Teachers and students gather evidence of student learning, interpret it, and accordingly adapt their teaching and learning activities (Black et al., 2004). School policies can support desirable teacher behaviours that improve their students' learning (Cumming et al., 2016). For example, policies that highlight the formative and diagnostic uses of assessment push teachers toward valuing assessment for boosting student learning (in New Zealand; G. T. Brown, 2011). When a school shared assessments of students' progress and exam scores with their teachers, they used them to monitor student learning, adapt their teaching to their students' needs, and offer individualised help, so their students outperformed their counterparts (e.g., in Germany, Förster & Souvignier, 2014; in the United States, Konstantopoulos et al., 2013). Hence, we propose the following hypothesis:

H-1a: *Among schools, those whose teachers use assessment for learning show higher student achievement.*

Unlike assessment for learning (often *formative assessment*, Sortwell et al., 2024), *assessment for evaluative purposes* is using data to make evaluative judgments about a student's level of attainment (often *summative assessment*; Glazer, 2014). As these are often simple numerical scores or grades for learning outcomes (Glazer, 2014), they facilitate comparisons among schoolmates and encourage them to compete (Gerber et al., 2018), which can interfere with student learning (e.g., via sabotage; J. O. Torres, 2019). For example, high-stakes tests that determine retention increase the likelihood of students dropping out and damage their careers (Andrew, 2014).

Assessments can also stress teachers. When an administrator uses student assessment data to evaluate the teaching effectiveness of teachers (El Helou et al., 2016), they can become anxious, which can weaken their teaching quality and reduce student learning (Wong et al., 2017). Hence, we propose the following hypothesis:

H-1b: *Among schools, those that use assessment data for evaluative purposes show lower student achievement.*

Assessment for accountability is using results to judge the effectiveness of teachers, schools, or education systems (Parsons, 2017). This includes public reporting of aggregated student data (e.g., test scores, graduation rates), often to attract students (Kuroda, 2022) or for public accountability (e.g., school rankings; Maingot & Zeghal, 2008). Public accountability can motivate schools to examine themselves, get feedback, understand it clearly (Hellrung & Hartig, 2013), and adapt teaching accordingly to enhance student learning (Sarrico et al., 2012). Also, it decreases information asymmetry between the two sides: schools and teachers, and parents and authorities. So, it informs parents about their children's learning progress, helping them better decide how to allocate education resources to support their children's learning (Seitsinger et al., 2008).

However, public accountability can harm teachers. Teachers can see school-wide assessments as valuing school rank more than student improvement (G. T. Brown & Harris, 2009). Also, US elementary school teachers facing greater accountability often felt more pressure, burned out, or left their jobs (Berryhill et al., 2009). Such dispirited teachers might teach poorly and their students might learn less.

Studies showed mixed results. Among schools, those comparing their test results to national standards (*soft accountability*) had slightly higher achievement in a PISA 2015 study (Klieme, 2020). Public disclosure of test results reduced the number of schools with low reading scores in a study of PISA 2009 and 2015 data (Teltemann & Schunck, 2020) but not for high-income countries with PISA 2006–2015 data (R. Torres, 2021). So, we test the following hypothesis:

H-1c: *Among schools, those that use assessment for accountability show higher student achievement.*

1.4. Teacher Formative Assessment Strategies

Wiliam and Thompson (2008) identified one big idea and five key formative assessment strategies. The idea is adapting teaching to student needs. The strategies are: (1) clarify learning goals and assessment criteria, (2) engineer effective learning activities and assessment, (3) offer feedback, (4) help students capitalise on classmates as learning resources, and (5) help them take charge of their learning.

A teacher who accurately interprets students' assessments better understands their learning needs and can accordingly adapt their teaching and scaffolding to help them achieve their learning goals (Wiliam & Thompson, 2008; Yeh, 2010; Zhai et al., 2018). For example, students showed higher reading scores when their teachers used assessment data to adjust their reading instruction (in the US; Slavin, 2013). Indeed, among teachers, those who adjusted their teaching after formative assessment improved their students' reading achievement more than those who only monitored students' learning performance without changing their teaching (Xuan et al.'s (2022) meta-analysis). Also, school cultures can encourage teachers to learn from one another, helping adaptive teaching spread across classrooms to improve student reading outcomes (Yan & Chiu, 2023; Yan & King, 2023). We propose the following:

H-2a: *Among students, those whose teachers adjust their instruction using assessment data show higher reading achievement.*

As clarifying and engineering strategies help identify and close the learning gap between students' current and expected performance levels, we can operationalise them as *clarifying goals and monitoring progress* (Torrance & Pryor, 2001). Students who know the goals and criteria know what is expected, which helps them choose suitable strategies and resources and direct their time and effort to improve their learning (Reed, 2002; Rust et al., 2003). Indeed, students taught goal-setting and reading strategies outperformed those receiving traditional instruction (in Taiwan; Shih & Reynolds, 2018). Also, teachers who track their students' progress help all of them see the gap between current and target performance, diagnose difficulties, and adapt their teaching and learning to improve their learning outcomes (Förster & Souvignier, 2014; Harris et al.'s (2022) systematic review). So, we hypothesise the following:

H-2b: *Among students, those whose teachers clarify goals and monitor students' progress show higher reading achievement.*

Feedback helps students grasp assessment criteria, check their progress, recognise learning gaps, and consider better ways to learn, which improves their learning outcomes (Winstone & Carless, 2019). For example, teacher feedback helped students apply their knowledge to learn new reading methods, which improved their reading outcomes (in

Sweden; Mežek et al., 2022). Indeed, feedback has one of the largest effects on student achievement (Hattie's (2017) meta-analyses of 80,000 studies), including reading achievement (e.g., China, Khine et al., 2023). Specifically, feedback for text comprehension increases uses of reading strategies and reading comprehension (Swart et al.'s (2022) meta-analysis). Note that results can differ across feedback type, content or implementation (Hattie & Timperley, 2007; Hogan & Payne, 2024; Wisniewski et al., 2020). So, we propose the following:

H-2c: *Among students, those who receive more teacher feedback show higher reading achievement.*

Although helping students take charge and capitalise on classmates are key formative assessment strategies (Wiliam & Thompson, 2008), PISA 2018 did not collect such data, so we could not assess their effects on reading achievement. So, we do not discuss them further.

1.5. Teacher Assessment Training

Teachers who learn to use student assessment in their teacher education or training often grasp and apply assessment goals and processes better in their teaching (Andersson & Palm, 2017a, 2017b; van Kuijk et al., 2016). Among teachers, those trained to use a learning tracking tool and interpret student performance data for formative assessment had students who learned more (in the Netherlands: van Kuijk et al., 2016; in Sweden: Andersson & Palm, 2017b). Hence, we hypothesise the following:

H-3a: *Among teachers, those whose teacher education or training included student assessment have students with higher reading achievement.*

Academic content and learning processes differ across subjects. Whereas math algorithms yield reliable, correct answers to closed problems (e.g., 4x = 16. Find x), reading comprehension relies on heuristics that evoke personal experiences and can yield multiple, suitable interpretations of open texts. When asked a reading comprehension question, "In this story, why did Ana spill the milk?" a student might answer, "Sometimes, I'm looking out the window when I reach for my milk."

Hence, compared to a math teacher, a reading teacher might use more background knowledge of their students' recent experiences to track their learning toward goals that are less clearly defined. So, teachers trained in reading assessment, rather than only general assessment, might assess and teach reading better for better student learning outcomes. We test this hypothesis:

H-3b: *Among teachers, those whose teacher education or training included reading assessment have students with higher reading achievement.*

1.6. This Study

Past studies showed that some socio-cultural or classroom assessment practices are linked to learning, but no study has examined them together across levels. So, we do so in this study. Building on the assessment ecology framework, we examine how school assessment policies, teacher assessment practice, or teacher assessment training are related to student reading performance. We used PISA 2018 because it has many survey items about school and teacher assessment practices, as well as students' reading achievement. To reduce *omitted variable bias*, we controlled for other factors linked to student reading achievement: country (education-emphasising culture, C. Y. Tan & Liu, 2018; national income, income inequality, R. B. King et al., 2024), classroom (class size, Shen & Konstantopoulos, 2017; classroom gender composition, S. Lee et al., 2014; school mean socio-economic status, Perry & McConney, 2010; student–teacher relationship, Chen et al., 2013; classroom discipline, Decristan et al., 2015; use of digital devices, A. L. Tan & Towndrow, 2009), teacher (teacher gender, Chudgar & Sankar, 2008; teaching experience, Huang & Moon, 2009), and

student (socio-economic status, Nguyen & Griffin, 2010; gender, Reilly et al., 2019; native language Kim et al., 2017; grade level, Konstantopoulos et al., 2013).

2. Method

In this study, we determine whether schools' and teachers' assessment practices are linked to reading achievement for 151,969 fifteen-year-olds within 5225 schools within 19 regions/countries: the United States, Brazil, Chile, Dominican Republic, Panama, Peru, Chinese Taipei, Hong Kong, Korea, Macau, Malaysia, Germany, Portugal, Spain, the United Kingdom, Morocco, Albania, Baku (Azerbaijan), and United Arab Emirates (OECD, 2018). Countries without assessment-related data were excluded because assessment practice is the key investigation in this study. The remaining 19 regions/countries represented a varied sample, including both Western and Eastern countries, thereby enhancing the generalizability of the findings.

2.1. Data

OECD (2018) employed a two-stage stratified sampling design: in each country/region, it selected 150 or more schools reflecting neighbourhood SES and student intake; and within each sampled school, it sampled 35 or more fifteen-year-olds and up to ten teachers who could teach them reading (*stratified sampling*). Students who did not take the exam (linguistically, mentally or physically incapable; less than 5% of the sample) were excluded by the participating countries and schools based on the international guidelines. Participating students completed a 2 h assessment and a 30 to 40 min survey. We also used World Bank (2018) data (national income and inequality).

Statistical power ($\alpha = 0.05$ and a small effect size of 0.1) differs across levels; it exceeds 0.99 for 151,969 students and 5225 schools (Konstantopoulos, 2008). However, it is low for 19 regions/countries, so the likelihood of a non-significant country/region false negative is high, but we retain confidence in the significant results (Kennedy, 2008).

2.2. Operationalising Assessment Ecology

To operationalise the contextual dimension of the assessment ecology framework (Chong & Isaacs, 2023), we mapped the available variables from the PISA 2018 dataset to three distinct levels (see Table 1). Socio-cultural context covers school (e.g., school assessment policy and school attributes) and country factors (e.g., national income and culture). Instructional context includes teacher attributes and assessment practices (e.g., formative assessment), while interpersonal context includes relationships (e.g., between student and teacher). While the broader assessment ecology framework also identifies textual (e.g., feedback features) and temporal (e.g., changes in activities over time) contexts, the PISA 2018 dataset lacks corresponding metrics; therefore, we could not include them in our analyses.

Table 1. Language assessment ecology theoretical framework.

Context	Explanatory Variables	Assessment-Relevant Indicators in PISA Data	Other Indicators in PISA Data
Socio-cultural	• School • Country	School assessment policy on the use of assessment data: • Assessment for learning; • Assessment for accountability; • Assessment for evaluative purposes; • Percentage of teachers receiving any formative assessment training.	• Percentage of female students • Percentage of female teachers • School mean SES • Class size • Mean years of teaching experience • National income (Real GDP per capita) • Income inequality (GDP Gini) • Culture (Confucian and non-Confucian)

Table 1. *Cont.*

Context	Explanatory Variables	Assessment-Relevant Indicators in PISA Data	Other Indicators in PISA Data
Instructional	• Teaching processes	Teacher formative assessment practice: • Clarify goals and monitor progress; • Provide feedback; • Instructional adjustments.	• Classroom discipline • Use of digital devices
Interpersonal	• Interpersonal attributes		• Student–teacher relationship
Student	• Student attributes		• Girl • SES • Native language speaker • Grade level

2.3. Variables

2.3.1. Outcome

Experts from participating countries/economies defined *reading achievement*, created assessment frameworks, designed 140 test questions, forward- and backward-translated them, and tested them for validity and reliability (see www.pisa.oecd.org; OECD, 2018). According to OECD (2018), reading achievement is the competence to understand, use, and reflect on written texts. Reading test questions varied across three dimensions: text form, reading attributes, and use. Along with continuous prose (such as narration, exposition, and argumentation), text forms also include lists, forms, graphs, and diagrams. Test items assessed five reading aspects: retrieve information, understand texts, interpret them, evaluate content, and evaluate form. Test questions varied according to contexts: public (e.g., announcement), school (e.g., book), work (e.g., manual), and private (e.g., letter).

PISA (2018) used adaptive testing to increase accuracy. All participants answered the same first set of questions. If they performed well, they often received more difficult or challenging questions; otherwise, they received easier questions (*adaptive testing*).

2.3.2. Country

Country measures include national income (*real per capita gross domestic product* [GDP]) and income inequality (*GDP Gini*, World Bank, 2018). See Table 1. *GDP Gini* is 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 (everyone has equal income; perfect equality) to 100 (one person has all income; perfect inequality). *Confucian* indicates a country whose culture was heavily influenced by Confucius's teachings: Chinese Taipei, Hong Kong, Korea, or Macau.

2.3.3. Student

The student demographics were girl, native language speaker, grade level, and socio-economic status. *Girl* has a value of one for a female; otherwise, zero. *Native language speaker* has a value of one for a native speaker of the official national language; otherwise, zero. *Grade level* is a student's years of schooling.

Socio-economic status (*SES*) is the standardised index ($m = 0$; SD = 1) of the highest parent job status (Meraviglia et al., 2018), father's years of schooling, and mother's years of schooling, created from a congeneric *multi-level confirmatory factor analysis* (ML-CFA, Hox et al., 2017).

2.3.4. School

The school variables comprised student demographics, school practices, and teacher means. Demographics were the school's *percentage of female students*, *class size* (mean number of students per class), *and school mean SES* (each school's mean SES of all sampled participants). As we control for student *SES*, the regression coefficient of *school mean SES* indicates the relation of schoolmates' mean SES to the outcome. (OECD collected the SES of only students in their sample (not the whole school), so we only compute each school's mean SES of the sample of students in each school [not all students].)

School practices included standardised indices of assessment for accountability, assessment for evaluative decisions, and assessments for learning, from congeneric CFA (Muthén & Muthén, 2018) of **principals' survey responses** (see Appendix A Table A1 for questions; see Table 2 for reliabilities). *Assessment for accountability* (three questions) indicates the uses of students' achievement data for accountability (for public posting, administrative tracking, or given to parents). All of these were yes/no questions. *Assessment for evaluative purposes* (six questions) captures other uses of student assessments: inform parents, determine retention or promotion, compare with other schools, compare with district or national performance, judge teacher effectiveness, or award student certificates. *Assessment for learning* (five questions) indicates uses of student assessments related to their learning: guide student learning, group students, monitor student progress, identify teaching improvements, or adapt to student needs.

Table 2. Goodness-of-fit statistics for confirmatory factor analyses.

Factor	#	Level	R_c	α	SRMR	CFI	TLI	RMSEA	χ^2	df	*p*
SES	3	Student	0.689	0.762	0.000	1.000	1.000	0.002	5.5	4	0.237
	3	School	0.980		0.031						
School											
Assessment for Accountability	3	School	0.572	0.295	0.007	0.999	0.998	0.005	2	2	0.325
Assessment for Learning	5	School	0.904	0.639	0.034	0.988	0.976	0.033	33	5	0.000
Assessment for Evaluative Purposes	6	School	0.759	0.600	0.007	0.999	0.997	0.025	13	3	0.005
Teacher											
Clarify goals and monitor progress	4	Student	0.809	0.813	0.014	0.990	0.971	0.050	280.4	4	0.000
	4	School	0.979		0.061						
Provide feedback	3	Student	0.863	0.852	0.001	1.000	1.000	0.000	1.2	4	0.876
	3	School	0.962		0.013						
Instructional adjustments	3	Student	0.681	0.714	0.003	1.000	1.000	0.002	4.3	4	0.364
	3	School	0.933		0.023						
Student view of Teacher											
Clarify goals and monitor progress	4	Student	0.779	0.801	0.012	0.992	0.975	0.037	828.0	4	0.000
	4	School	0.958		0.025						
Provide feedback	3	Student	0.865	0.831	0.000	1.000	1.000	0.001	4.7	4	0.320
		School	0.964		030						
Instructional adjustments	3	Student	0.767	0.776	0.000	1.000	1.000	0.000	2.6	4	0.619
	3	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; df = Degrees of freedom.

All teacher variables are school means of language teacher responses (Bonneville-Roussy et al., 2019): *percentage of female teachers*, *mean years of school teaching experience*, and *percentage of teachers receiving any formative assessment training*. All language teaching practices of each school, including clarifying goals and monitoring progress, providing

feedback, and instructional adjustments, are standardised indices from congeneric ML-CFA (Muthén & Muthén, 2018) of these **teachers' survey responses**. Possible responses for all of these survey questions were: (1) never or hardly ever, (2) some lessons, (3) many lessons, or (4) every lesson (see Appendix A Table A1 for questions; see Table 2 for reliabilities). *Clarify goals and monitor progress* is assessed by four questions. *Provide feedback* is assessed by three questions. *Instructional adjustments* is assessed by three questions.

2.3.5. Student Views of Teaching Practices

Likewise, we created three corresponding student indices from **students' survey responses** ("my teacher" replaced "I"): *student view of clarifying goals and monitoring progress*, *student view of providing feedback*, and *student view of instructional adjustments* (see Appendix A). We computed each school's means of these student indices: *school mean of student view of clarifying goals and monitor progress*, *school mean of student view of providing feedback*, and *school mean of student view of instructional adjustments*. 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.

We captured students' perceived differences among such classroom practices within a school. Specifically, we computed each school's standard deviations (SD): *SD clarify goals and monitor progress*, *SD provide feedback*, and *SD instructional adjustments*.

We capture perceived differences in these classroom practices between teachers and students. Specifically, we computed the student value minus the teacher value: *clarify goals and monitor progress difference, provide feedback difference*, and *instructional adjustments difference.*

2.4. Data Analysis

To accurately analyse these data, we address the following issues with specific statistics strategies (for details, see R. King et al., 2022): (a) unrepresentative data with *weighting* (Hansen, 2022), (b) sampling error with *plausible values analysis* (Monseur & Adams, 2009), (c) missing data (4% in these data with *Markov Chain Monte Carlo multiple imputation* (Peugh & Enders, 2004) and Little (1988) test (p = 0.88, suggesting *missing completely at random* (MCAR) data; true MCAR tests requiring follow-up interviews of respondents were too costly), (d) many test questions with *overlapping subtests* and *anchor items* (Embretson & Reise, 2013), (e) nested structure of the data across students across schools across countries with *multi-level (ML) analysis* (Hox et al., 2017), (f) survey measurement errors (e.g., for *achievement assessment* construct) with *ML-confirmatory factor analyses* (ML-CFA, Joreskog & Sorbom, 2022), (g) test measures with *factor analyses* and *graded response Rasch models* (Embretson & Reise, 2013), (h) indirect, ML mediation effects with *multi-level M-tests* (MacKinnon et al., 2004), (i) *cross-level interactions* (student gender x teacher gender) with *random parameters* in *random effects models* (Hox et al., 2017), (j) many hypotheses' *false positives* with the *two-stage linear step-up procedure* (Benjamini et al., 2006), (k) compare effect sizes with *Lagrange multiplier tests* (Bertsekas, 2014), (l) consistency of results across datasets (*robustness*) with analyses of data subsets and analyses of original (not estimated) data (Hansen, 2022).

2.4.1. Factor Analyses

We test each construct's (e.g., SES) survey items for type of factor structure (*single, multiple, hierarchical, nested/bi-factor*) and internal validity, while minimising their measurement errors with CFA (Joreskog & Sorbom, 2022). *Bartlett factor scores* yield unbiased estimates of factor score parameters (Joreskog & Sorbom, 2022). To assess CFA fit, we use the *comparative fit index* (CFI), *Tucker–Lewis index* (TLI), *root mean square error approximation* (RMSEA), and *standardised root mean square residual* (SRMR); they minimise Type I and Type II errors in many simulations (Hu & Bentler, 1999). We use two fit thresholds: good (CFI & TLI > 0.95;

RMSEA < 0.06; SRMR < 0.08) and moderate (0.90 < CFI & TLI < 0.95; 0.06 < RMSEA < 0.10; 0.08 < SRMR < 0.10).

Multi-group Rasch models for each item in each country showed similar parameters, and hence measurement equivalence across regions/countries (May, 2006). (Unlike CFA, multi-group Rasch models (a) model heterogeneous uses of ordinal rating scales and (b) require only one invariant anchor item across regions/countries; Rossi et al., 2001). All anchor items show acceptable discrimination ($\alpha > 0.50$) and threshold parameters ($-3 < \beta < 3$). Likewise, past studies showed similar survey responses and participant interpretations across regions/countries (G. Brown et al., 2005; Schulz, 2003).

2.4.2. Explanatory Model

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

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

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).

$$\begin{aligned}\textbf{Reading}_{ijk} = \beta_0 + e_{ijk} + f_{jk} + g_k + \beta_t\textbf{Country}_k + \beta_{ujk}\textbf{Student}_{ijk} + \beta_{vk}\textbf{School}_{ijk} \\ + \beta_{xjk}\textbf{Assessment}_{ijk} + \beta_{xjk}\textbf{Student\_views}_{ijk} + \beta_{zjk}\textbf{Interactions}_{ijk}\end{aligned} \quad (2)$$

Structural variables can affect process variables, so we enter **vectors** of the former explanatory variables before the latter. **Country** variables (real GDP per capita, GDP GINI, Confucian) can affect its people, so we enter them before **Student** and **School**. A *nested hypothesis test* (χ^2 log likelihood) indicates whether each set of explanatory variables is significant (Kennedy, 2008). Omitting *non-significant* variables does not cause *omitted variable bias*, so we safely remove them to increase precision and reduce *multicollinearity* (Kennedy, 2008). We apply this procedure to each vector.

As families and students often select their schools, we enter **Student** (girl, SES, native language speaker, grade level) before **School** (percentage of female students, school mean SES, class size, assessment for accountability, assessment for evaluative decisions, assessments for support learning, percentage of female teachers, mean years of school teaching experience, percentage of teachers receiving any formative assessment training, student–teacher relationship, classroom discipline, use of digital devices), followed by **Assessment** (assessment practice in teaching education or training programmes, received training emphasising reading comprehension assessment, teacher formative assessment processes: clarify goals and monitor progress, provide feedback, instructional adjustments), and student view of assessment practices (student view of clarifying goals and monitoring progress, student view of providing feedback, student view of instructional adjustments school mean of student view of clarifying goals and monitoring progress, school mean of student view of providing feedback, and school mean of student view of instructional adjustments, SD clarify goals and monitor progress, SD provide feedback, and SD instructional adjustments, clarify goals and monitor progress difference, provide feedback difference, instructional adjustments difference). We also analyse residuals for influential outliers.

3. Results

3.1. Factor Analysis

The ML-CFA showed acceptable fits for all factors, but the reliability coefficient of achievement accountability was only 0.572 (see Table 2; see factor loadings in Appendix A Table A2). So, we entered its components as separate explanatory variables into the regression: public posting of achievement data, administrative tracking of it, and directly giving it to parents. Among responses to assessment for evaluation questions, those for whether assessments determined a student's retention or promotion did not align with the others, so that question was removed from the factor.

3.2. Summary Statistics

See Table 3 for summary statistics.

Table 3. Summary statistics (N = 151,696).

Variable	Mean	SD	Min	Max
Reading test score	453.461	107.482	84.05	868.87
Country				
Real GDP per capita	26,112.169	15,264.377	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
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
Achievement data are posted publicly	0.265		0	1
Achievement data are tracked over time by an administrative authority	0.807		0	1
Achievement data are provided directly to parents	0.888		0	1
Teacher				
Female teacher	0.747	0.270	0.000	1.000
Years teaching experience	9.716	5.961	0.000	50.000
My teacher education/training . . .				
. . .had student assessment practices	0.703	0.267	0	1
. . .emphasised ways to assess reading comprehension	2.318	0.449	0	1
Formative assessment practice				
Learning goals and progress (teacher-reported)	1.499	0.351	1.000	4.000
Feedback (teacher-reported)	3.193	0.383	1.274	4.000
Instructional adjustments (teacher-reported)	3.233	0.364	1.000	4.000
Learning goals and progress (student-reported)	1.846	0.730	1.000	4.000
Feedback (student-reported)	2.422	0.856	1.000	4.000
Instructional adjustments (student-reported)	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

Note: For dichotomous variables, the mean indicates the proportion of participants with the attribute, and SD is not meaningful.

It can be seen that, in general, teachers are reported to frequently engage in formative assessment practices, as indicated by the scale means. These means tend to be higher than the mid-point of the scale (2.5) for positively worded items or lower than the mid-point for

negatively worded items. Moreover, teacher-reported formative assessment practices are relatively higher than student-reported formative assessment practices.

3.3. Model Testing

Half of the differences in reading achievement were at the student level (50%), with 28% at the country level and 23% at the school level (see Table 4). All results discussed below described the first entry into the regression, controlling for all previously included variables. Ancillary regressions and statistical tests are available upon request.

Table 4. Summary of unstandardised regression coefficients (with standard errors in parentheses) of 3-level analyses of students' reading test scores.

	Reading Test Score							
Explanatory Variable	**Model 1 Country**		**Model 2 + Student + School**		**Model 3 + Assessment**		**Model 4 + Student Views**	
Real GDP per capita ($1000s)	2.285	***	1.314	**	1.179	*	1.029	*
	(0.444)		(0.477)		(0.456)		(0.422)	
Confucian country	45.940	*	62.980	**	63.940	**	56.380	**
	(19.160)		(20.610)		(19.670)		(18.170)	
Girl			16.340	***	16.350	***	16.310	***
			(0.419)		(0.419)		(0.418)	
SES			5.256	***	5.256	***	5.267	***
			(0.236)		(0.236)		(0.236)	
Native language speaker			17.450	***	17.540	***	17.850	***
			(0.701)		(0.700)		(0.699)	
Grade level			38.340	***	38.330	***	38.020	***
			(0.341)		(0.341)		(0.340)	
School mean SES			49.060	***	49.120	***	42.570	***
			(1.198)		(1.193)		(1.203)	
Student–teacher relationship			6.570	***	6.568	***	6.317	***
			(0.262)		(0.262)		(0.263)	
Classroom discipline			8.724	***	8.726	***	8.702	***
			(0.290)		(0.290)		(0.290)	
My teacher education/training . . .								
. . .had student assessment practices					11.030	***	9.055	***
					(2.144)		(2.052)	
. . .emphasised ways to assess reading comprehension					−8.119	***	−6.450	***
					(1.664)		(1.596)	
Assessment for accountability					4.149	**	3.927	**
					(1.404)		(1.340)	
Assessment for evaluative purposes					−7.585	***	−6.699	***
					(1.936)		(1.849)	
Student views of . . .								
. . .Teacher clarify goals and monitor progress—School mean							58.860	***
							(3.372)	
. . .Instructional adjustments —School mean							36.120	***
							(2.912)	
. . .Provide feedback —School standard deviation (SD)							−57.700	***
							(3.845)	
Variance at each level	Explained variance at each level							
Country (28%)	0.720		0.672		0.702		0.746	
School (23%)	0.000		0.495		0.503		0.554	
Student (50%)	0.000		0.111		0.111		0.111	
Total variance explained	0.198		0.353		0.363		0.387	

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Country (real GDP per capita, Confucian), student, school, teacher, and assessment variables accounted for differences in students' reading achievement (see Table 4). Students

in a richer region/country with $1000 more real GDP per capita than the mean scored 2 points higher in reading on average (see Table 4, model 1, top left). In regions/countries with strong Confucian cultural influences, students scored 46 points higher in reading, compared to other students. These country-level variables accounted for 28% of the differences in reading scores (see Table 4, model 1, bottom left).

Student attributes (girl, SES, language, grade level), school attributes (schoolmate SES), and teaching practices (relationships with students, classroom discipline) were also linked to reading achievement. Girls outperformed boys in reading by 16 points (see Table 4, model 2, top middle). Students in a richer family with one level higher SES than the mean scored 5 points higher in reading than other students. Native language speakers outperformed other students by 17 points in reading. Students in one grade higher than the mean scored 38 points higher in reading. Also, students whose schoolmates' families had one level higher SES than the mean scored 49 points higher in reading than other students. Students who reported better relationships with their teachers or better classroom discipline outperformed other students in reading (+7% and +9%, respectively). These variables accounted for an additional 15% of the variance in reading scores (see Table 4, model 2, bottom middle; 15% ≈ 0.353–0.198).

School assessment policies (assessment for accountability and assessment for evaluative purposes) were linked to reading achievement. In schools that publicly posted achievement data for accountability, students scored 4 points higher, compared to students in other schools. However, students in schools that more often use assessment for evaluative purposes scored 8 points lower in reading than students in other schools did. These assessment variables accounted for an additional 1% of the variance in reading scores (see Table 4, model 3, bottom, second column from right; 1% = 0.363–0.353).

Student perceptions of their teacher's formative assessment practices were also linked to their reading test scores. When students in a school perceive their teachers to clarify their goals and monitor student progress more often, they show higher reading achievement (+59 points; see Table 4, model 4, bottom, rightmost column). Likewise, students in a school that report their teacher making instructional adjustments more often show higher reading achievement (+36 points). By contrast, students in schools with one standard deviation greater difference in frequency of teacher feedback across students than the mean show lower reading achievement (−57 points). These student views of teacher assessment practices accounted for an additional 2% of the variance in reading scores (see Table 4, model 3, bottom, second column from right; 2% ≈ 0.387–0.363).

Teacher assessment training was also linked to student reading achievement. Teachers' experience in assessment practice in teaching education or training programmes leads to higher student reading scores (+9 points). However, when teachers receive training emphasising reading comprehension assessment, the students show lower reading scores (−6 points).

All other explanatory variables and interactions were not significant. These explanatory variables accounted for nearly 39% of the differences in the students' reading scores. Analysis of residuals showed no substantial outliers. Robustness tests on data subsets and on the original data showed similar results.

4. Discussion

This study examined the associations between school assessment policies, teacher formative assessment practices and training and students' reading performance. The findings are discussed in the following sections.

4.1. The Relationship Between School Assessment Policies Regarding the Uses of Assessment Data and Reading Performance

We examined the relationship between three types of school assessment policies regarding the use of assessment (i.e., for learning, for evaluative purposes, and for accountability) and reading performance. The results show that school assessment policies for learning are not significantly associated with student reading scores, showing no support for H-1a. This finding is inconsistent with previous empirical studies that revealed a positive relationship in Germany (Förster & Souvignier, 2014) and in the US (Konstantopoulos et al., 2013), probably because the current study included data from multiple countries. Another possible explanation is that teachers, as a bridge between policy and students, receive insufficient training or support in implementing those assessment policies in classrooms (Nortvedt et al., 2016). Hence, assessment policies appear to have little correlation with student academic performance.

Students' lower reading scores were observed in contexts where schools have policies that use student assessment data for evaluative purposes, supporting H-1b. This negative association may be because using assessment data for evaluative purposes (e.g., teacher evaluation, and students' retention or promotion) could cause pressure for both teachers and students, exerting a negative effect on student academic performance and motivation (Moss, 2013).

Schools' publicly posting achievement data for accountability positively predicted student reading performance, supporting H-1c. This finding suggests that public or administrative authorities' attention can promote students' learning. As Hellrung and Hartig's (2013) review revealed, public accountability can motivate teachers and school staff to self-evaluate and adjust their teaching or administrative practice. It is also possible that disclosing assessment data to parents could encourage them to pay more attention to their children's studies, thus fostering students' learning progress (Seitsinger et al., 2008).

4.2. The Relationship Between Teacher Formative Assessment Strategies and Reading Performance

Students' perceived teacher practice in clarifying their goals and monitoring their progress is a positive predictor of student reading scores, supporting H-2b. This result is in line with the findings from Shih and Reynolds' (2018) study, showing that such formative assessment strategies bear the potential for student learning: clarifying goals set benchmarks for students so that students can allocate their resources and choose the most suitable learning strategies to achieve the target (Rust et al., 2003; Torrance & Pryor, 2001); and monitoring can raise teachers' and students' awareness of the gap between current performance and expected outcomes and encourage them to close the gap timely by adjusting teaching or learning strategies (Harris et al., 2022; Yan et al., 2021;).

Students who reported higher teacher feedback frequency had lower reading scores, showing no support for H-2c. This finding is inconsistent with the previous empirical study with Swedish students (Mežek et al., 2022) and the secondary research focusing on Chinese students (Khine et al., 2023), both revealing a positive relationship between teacher feedback and student reading scores. We propose several possible reasons for the negative association found in this study: (a) The feedback information delivered by different teachers or in different countries may lead to different results; for instance, a negative relationship was also observed among students in the East but not in the West (Yan et al., 2021). Thus, the effect of teacher feedback highly depends on the feedback context (Panadero & Lipnevich, 2022). (b) The PISA measure of teacher feedback captured only feedback frequency, not quality or effectiveness. Given that the PISA measure of teacher feedback captured only feedback frequency, not quality or effectiveness, we could not comprehensively examine the link between teacher feedback and student reading performance; future research is

needed to untangle this association in detail. (c) An OECD (2016, p. 62) report about PISA 2015 pointed out that it was possible that "the relationship runs not from teaching strategy to student success on the items, but in the opposite direction"; hence, the negative relationship could result from teachers providing more feedback to lower-level students. (d) Teacher expertise and authority may be reinforced by overwhelming teacher feedback; as a result, students may over-rely on teacher feedback and fail to actively engage in the feedback process by making judgments and generating internal feedback, which is essential for their life-long learning (Han & Xu, 2020).

Higher reading scores were observed when students found that their teachers often made instructional adjustments, supporting H-2a. This result is consistent with previous studies that students showed higher reading scores if their teachers adjusted their reading instruction (Slavin, 2013; Yan et al., 2021). The positive relation reiterates that teachers' adjustment of instructions can facilitate students' progress by fitting their needs within their zone of proximal development (Vygotsky, 1978). Teachers making instructional adjustments from time to time can also help students address their learning gaps timely. Moreover, given that students' decreased academic motivation may be related to decreased teacher instructional behaviour (Maulana et al., 2016), students may be more motivated to learn if they perceive their teachers actively adjusting their teaching methods.

4.3. The Relationship Between Teacher Assessment Training and Reading Performance

Students had higher reading scores when their teachers had received teaching education or training programmes that included student assessment practice, supporting H-3a. The positive association supports the importance of providing teachers with professional development to refine their assessment practices; otherwise, they may not use assessment data effectively despite their strong willingness to do so (Andersson & Palm, 2017a, 2017b).

Surprisingly, when teachers received assessment training with an emphasis on reading comprehension, students were observed with lower reading scores, not supporting H-3b. The possible explanations for this negative association could be: (a) The training is not effective either because of the implementation quality (e.g., too short) or teachers' failure in comprehending assessment knowledge thoroughly or transferring knowledge into practice correctly (e.g., because of personal competency or contextual constraints). For instance, Liu et al. (2016) found that in China, teacher training content is not fully compliant with policy, and some training courses provided by college professors appear to be too theoretical for rural teachers. (b) Only one PISA item measured the practice of emphasising reading comprehension assessment in teacher training, but the effect may vary depending on the length and depth of the training and training methods. More empirical studies are needed in this area.

4.4. The Relationship Between Control Variables and Reading Performance

Context factors (including instructional-level, interpersonal-level, and socio-cultural-level factors) and student background factors were related to student reading performance as well. In terms of the instructional-level context, students' reading scores are positively correlated to better classroom discipline, consistent with Ma et al.'s (2013) findings based on student data from Hong Kong, Taipei, and Japan in PISA 2009. This result suggests that good classroom discipline is likely to facilitate student learning performance, since Rahimi and Karkami (2015) found that discipline strategies could relate to teaching effectiveness, motivation and student learning achievement. However, no significant relationship was found between the use of digital devices and student reading performances.

As for the interpersonal-level context, a positive relationship was observed between student–teacher relationships and student reading scores, which aligns with J. Lee's (2012)

PISA 2000 research using US data and reporting that teacher–student relationships significantly predicted reading performance. This positive link may be attributed to students' greater confidence and engagement with academic content when they have a good relationship with their teachers, achieving higher scores (Zee et al., 2021).

For socio-cultural-level context, national incomes positively predicted students' performance in the reading test, concurring with past studies (e.g., Chiu et al., 2017; Chiu & Chow, 2010). This result may be attributed to the availability of greater learning resources and the increased emphasis placed on children's education by parents in countries with higher national incomes. However, no significant relationship was found between income equality and student reading scores. Students in countries with stronger Confucian culture were observed with higher reading scores, probably because students have stronger motivation to succeed academically in Confucian culture (C. Y. Tan & Liu, 2018). Regarding school-level factors, only school mean SES was positively related to student reading scores, consistent with the study of Perry and McConney (2010). In schools with higher SES students, parents are more likely to obtain higher educational levels and may discuss and plan their children's study together more thoroughly. Other school-level factors (i.e., percentage of female students, percentage of female teachers, class size, and teaching experience), however, showed no significant relationship with student reading scores.

Regarding the student background factors, girls outperformed boys in the reading test, probably because boys tend to place less value on reading and have lower motivation to read compared to girls with the same level of reading proficiency (Marinak & Gambrell, 2010). SES families are a positive predictor of student reading scores, probably because higher SES equips students with more education resources and learning opportunities (Chiu & McBride-Chang, 2010). Unsurprisingly, students who are native language speakers were observed with better reading performance than other students, because they are highly likely not to have language barriers in comprehending test items. Additionally, there is a positive association between student grades and their reading scores, concurring with the study by Konstantopoulos et al. (2013). This result could be explained by students' higher cognitive levels in higher grades and perhaps greater knowledge of test-taking techniques through more test training.

4.5. Policy and Practical Implications

This study contributes to school policymaking, teacher practices, and teacher education in relation to assessment practice. The study reveals that schools setting policies that use assessment data for accountability is associated with higher student reading scores, but using it for evaluation has a negative association with students' learning. These findings suggest that school policymakers reduce the use of assessment data for evaluative purposes, for instance, evaluating teachers' performances and determining students' retention based on students' scores. Instead, they are advised to disclose student assessment data for parents to support students' learning or for the public to motivate school self-evaluation. Schools are still suggested to use student assessment data for teaching and learning despite its non-significant relationship with student reading scores. The results reiterate the importance of awareness of cultural values and social norms when making and implementing assessment policies (G. T. Brown et al., 2009; Yan & Brown, 2021). Schools also need to systematically support teachers in transforming school assessment policies into classroom practices through various activities (e.g., workshops and seminars), since knowledge and capacity cannot be built overnight (Oo et al., 2024; Yan, 2021).

Regarding teachers' formative assessment strategies, two strategies (i.e., clarify learning goals and monitor progress, and instructional adjustment) were positively coupled with student reading performance. Despite the revealed adverse association with provid-

ing teacher feedback, we recommend that teachers adopt all three formative assessment practices in classrooms. As Sortwell et al.'s (2024) umbrella review highlighted, across 13 meta-analyses, formative assessment generally positively influenced student learning. To reconcile our divergent findings with it, we argue that feedback provided without concurrent pedagogical action may be insufficient or even counterproductive. Instead, the potential of feedback is likely to be shown and amplified when it is systematically integrated with other strategies, such as instructional adjustment catered to students' learning needs. In addition, teachers need to enhance feedback quality and activate students' agency in the feedback process. This is because the content and quality of feedback, the method of delivery, and student characteristics and agency all influence how students process and utilise teacher feedback, thereby affecting its impact on their learning performance (Panadero & Lipnevich, 2022). Feedback-literate students are more likely to actively engage with teacher feedback and make progress (Carless & Boud, 2018).

The positive relationship only appears between student reading scores and teacher training in assessment practice, but not in reading comprehension assessment, showing the necessity to provide teachers with comprehensive professional development in their career. While previous research proposed to provide professional development for teachers about both general and domain-specific assessment practice (Breiter & Light, 2006), our finding reveals the necessity to carefully design training and education programmes, for instance, what assessment knowledge or capacities teachers are supposed to acquire, how to teach, and how to examine teachers' gains after training. Poorly designed training or programmes may confuse teachers and worsen the expected outcomes.

4.6. Limitations and Future Directions

One of the limitations of the current study is that no data from textual-level and temporal-level contexts is included in the PISA 2018 dataset. Future studies can investigate the assessment-related factors in these two contexts to provide a more comprehensive picture. Additionally, teacher assessment training is an umbrella term in this study; however, teachers may improve their classroom practices in different assessment aspects to different degrees. Future studies can explore the link between teaching training in specific assessment practices (e.g., providing feedback, instructional adjustment) and student academic performance. As a cross-sectional large-scale study using the PISA dataset, we also acknowledged the possibilities of reciprocal relationships between assessment-related practices and student achievement (Forestier & Adamson, 2017). For instance, as we found a negative association between the frequency of teacher feedback practice and student reading performance, lower-achieving students likely receive more feedback, but not more teacher feedback, leading to lower performance. In addition, the methodological limitation involves modest reliability estimates, especially for some school-level policy constructs, which may influence the significant associations identified between these assessment policies and reading achievement, requiring the audience to interpret the results cautiously. We also did not examine country-level assessment policies as a predictor, as investigated in Klieme's (2020) study.

5. Conclusions

Using PISA 2018 data, this ecological study examines how different aspects of assessment, in particular school assessment policies, teacher assessment practices, and training, interact to be connected to students' reading performance. The findings revealed that when schools publicly posted achievement data for accountability purposes, teachers frequently clarified learning goals and monitored student progress, made instructional adjustments, and received training that included assessment practice, students tended to achieve higher

reading scores. In contrast, when schools relied more on assessment data for evaluation, teachers provided more feedback, and teacher training emphasised reading comprehension assessment, students were more likely to have lower reading scores. Nevertheless, it is important to interpret these results while considering cultural values and social norms. Overall, these findings contribute to the development of a comprehensive model that explores the relationship between assessment and learning.

Author Contributions: Conceptualization, Z.Y.; methodology, Z.Y. and M.M.C.; formal analysis, M.M.C.; writing—original draft preparation, Z.Y. M.M.C. and J.G.; writing—review and editing, Z.Y., M.M.C., J.G., L.Y. and Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

Table A1. Questions on formative assessment strategies.

	Questions
Assessment for Accountability In your school, are achievement data used in any of the following <accountability procedures>? (Yes/No) . . . (Achievement data include aggregated school or grade-level test scores or grades, or graduation rates.)	
1	. . .Achievement data are posted publicly (e.g., in the media)
2	. . .Achievement data are tracked over time by an administrative authority
3	. . .Achievement data are provided directly to parents
Assessment for Learning In your school, are assessments of students in <national modal grade for 15-year-olds> used for any of the following purposes? (Yes/No) . . .	
1	. . .To guide students' learning
2	. . .To group students for instructional purposes
3	. . .To monitor the school's progress from year to year
4	. . .To identify aspects of instruction or the curriculum that could be improved
5	. . .To adapt teaching to the students' needs
Assessment for Evaluative Purposes In your school, are assessments of students in <national modal grade for 15-year-olds> used for any of the following purposes? (Yes/No) . . .	
1	. . .To inform parents about their child's progress
2	. . .To make decisions about students' retention or promotion
3	. . .To compare the school to <district or national> performance
4	. . .To make judgements about teachers' effectiveness
5	. . .To compare the school with other schools
6	. . .To award certificates to students

Table A1. *Cont.*

	Questions
Clarifying goals and monitoring progress	
1	I set clear goals for the students' learning.
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.
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.
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.

Table A2. Factor loadings.

	School Level			Student Level		
Variable	**Factor Loading**	**SE**	**Unique**	**Factor Loading**	**SE**	**Unique**
SES						
Father's years of schooling	0.559	0.003	0.031	0.670	0.004	0.551
Mother's years of schooling	0.562	0.003	0.009	0.722	0.004	0.479
Highest job status	0.563	0.004	0.073	0.487	0.004	0.763
School						
Assessment for Accountability						
1 Publicly post achievement data	0.608	0.055	0.630			
2 Track achievement data	0.610	0.053	0.628			
3 Provide achievement data to parents	0.369	0.039	0.864			
Assessment for Learning						
1 Guide students' learning	0.753	0.025	0.433			
2 Group students	0.472	0.022	0.778			
3 Monitor the school's progress	0.699	0.021	0.511			
4 Identify areas for improvement	0.852	0.016	0.274			
5 Adapt teaching	0.900	0.015	0.190			
Assessment for Evaluative Purposes						
1 Inform parents	0.753	0.025	0.433			
2 Compare with district/nation	0.472	0.022	0.778			
3 Judge teacher effectiveness	0.699	0.021	0.511			
4 Compare with other schools	0.852	0.016	0.274			
5 Award student certificates	0.900	0.015	0.190			
Teacher						
Clarify goals and monitor progress						
1 Set learning goals	0.358	0.015	0.003	0.708	0.006	0.499
2 Ask questions to check progresses	0.291	0.014	0.022	0.764	0.006	0.417
3 Summarise the previous lesson	0.198	0.030	0.053	0.615	0.006	0.623
4 Tell students what to learn	0.250	0.020	0.019	0.743	0.006	0.448

Table A2. *Cont.*

Variable	School Level			Student Level		
	Factor Loading	SE	Unique	Factor Loading	SE	Unique
Provide feedback						
1 Comment on students' strengths	0.354	0.012	0.015	0.704	0.005	0.505
2 Comment on areas for improvement	0.303	0.012	0.017	0.886	0.004	0.215
3 Tell students how to improve	0.356	0.011	0.011	0.788	0.005	0.380
Instructional adjustments						
1 Adjust teaching to meet students' needs	0.383	0.021	0.014	0.638	0.007	0.593
2 Provide individual help	0.333	0.021	0.051	0.665	0.007	0.558
3 Change the lesson structure	0.226	0.024	0.032	0.629	0.007	0.604
Student view of Teacher						
Clarify goals and monitor progress						
1 Set learning goals	0.327	0.005	0.016	0.697	0.003	0.514
2 Ask questions to check progresses	0.305	0.004	0.010	0.724	0.003	0.477
3 Summarise the previous lesson	0.397	0.007	0.078	0.635	0.003	0.596
4 Tell students what to learn	0.302	0.006	0.019	0.666	0.003	0.557
Provide feedback						
1 Comment on students' strengths	0.364	0.005	0.016	0.715	0.002	0.489
2 Comment on areas for improvement	0.297	0.003	0.009	0.878	0.002	0.230
3 Tell students how to improve	0.301	0.004	0.010	0.818	0.002	0.330
Instructional adjustments						
1 Adjust teaching to meet students' needs	0.232	0.010	0.008	0.739	0.003	0.454
2 Provide individual help	0.254	0.009	0.008	0.736	0.003	0.459
3 Change the lesson structure	0.237	0.010	0.007	0.690	0.003	0.524

Appendix B. Ancillary Analyses

Table A3. Correlations, variances, and covariances in the lower left, diagonal, and upper right matrices.

	Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
1	Reading test score	11,552	9.54	561	6.69	26.01	5.08	22.22	6.01	10.41	21.22	1.36	−4.81	1.84	−11.12	9.69	1.32	0.72
2	Confucian	0.24	0.13	1.35	0.00	0.00	0.01	−0.02	0.01	0.04	0.01	0.00	0.00	0.02	0.00	0.02	0.01	0.00
3	Real GDP per capita (1000s)	0.34	0.24	233	−0.05	2.83	−0.17	1.09	−0.51	0.11	2.81	0.14	−0.43	−0.02	−2.41	1.99	−0.50	0.20
4	Girl	0.12	−0.01	−0.01	0.25	−0.01	0.00	0.03	0.01	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
5	SES	0.24	0.01	0.19	−0.02	1.00	0.02	0.14	0.03	0.03	0.34	0.02	0.01	0.00	−0.01	0.02	0.03	0.01
6	Native language speaker	0.12	0.07	−0.03	0.02	0.04	0.15	0.01	0.01	0.01	0.01	0.00	−0.01	0.00	−0.01	−0.01	0.01	0.00
7	Grade level	0.27	−0.07	0.09	0.09	0.19	0.04	0.58	0.02	0.05	0.11	0.04	0.02	0.00	0.02	0.00	0.02	0.01
8	Student–teacher relationship	0.07	0.02	−0.04	0.02	0.03	0.04	0.04	0.63	0.15	0.03	0.01	0.02	0.00	0.02	−0.05	0.07	0.01
9	Classroom discipline	0.13	0.14	0.01	0.05	0.04	0.05	0.08	0.25	0.56	0.04	0.01	0.02	0.01	0.03	−0.04	0.05	0.00
10	School mean SES	0.34	0.03	0.32	0.00	0.58	0.03	0.26	0.06	0.09	0.34	0.02	0.01	0.00	−0.01	0.02	0.03	0.01
11	Students tests in public	0.03	−0.01	0.02	0.00	0.04	0.00	0.12	0.04	0.04	0.07	0.19	0.03	0.01	0.03	−0.02	0.01	0.00
12	Assess to evaluate	−0.13	0.02	−0.08	−0.01	0.03	−0.08	0.08	0.06	0.08	0.06	0.19	0.12	0.00	0.05	−0.03	0.01	0.00
13	Teacher ed: Assess students	0.06	0.18	−0.01	−0.01	0.01	−0.01	−0.02	0.01	0.03	0.01	0.05	−0.03	0.07	0.01	0.00	0.00	0.00
14	Teacher ed: Assess reading	−0.23	0.01	−0.35	0.00	−0.03	−0.07	0.06	0.06	0.08	−0.04	0.13	0.29	0.10	0.20	−0.06	0.02	0.00
	Student views of Teacher																	
15	Goals & monitor: School mean	0.30	0.18	0.43	−0.01	0.07	−0.07	0.01	−0.22	−0.18	0.12	−0.11	−0.27	0.03	−0.43	0.09	−0.05	0.00
16	Adjust instruction: School mean	0.05	0.06	−0.12	0.02	0.10	0.09	0.08	0.34	0.25	0.17	0.11	0.16	0.04	0.17	−0.62	0.07	0.01
17	Provide feedback: School SD	0.06	0.05	0.11	−0.03	0.12	0.05	0.09	0.06	0.04	0.21	0.04	0.03	−0.04	0.04	−0.08	0.18	0.01

References

Andersson, C., & Palm, T. (2017a). Characteristics of improved formative assessment practice. *Education Inquiry*, *8*(2), 104–122. [CrossRef]

Andersson, C., & Palm, T. (2017b). The impact of formative assessment on student achievement: A study of the effects of changes to classroom practice after a comprehensive professional development programme. *Learning and Instruction*, *49*, 92–102. [CrossRef]

Andrew, M. (2014). The scarring effects of primary-grade retention? A study of cumulative advantage in the educational career. *Social Forces*, *93*(2), 653–685. [CrossRef]

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

Berryhill, J., Linney, J. A., & Fromewick, J. (2009). The effects of education accountability on teachers: Are policies too-stress provoking for their own good? *International Journal of Education Policy and Leadership*, *4*(5), 1–14. [CrossRef]

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

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan*, *86*(1), 8–21. [CrossRef]

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. [CrossRef]

Breiter, A., & Light, D. (2006). Data for school improvement: Factors for designing effective information systems to support decision-making in schools. *Journal of Educational Technology & Society*, *9*(3), 206–217.

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. [CrossRef]

Brown, G. T. (2011). Teachers' conceptions of assessment: Comparing primary and secondary teachers in New Zealand. *Assessment Matters*, *3*, 45–70. [CrossRef]

Brown, G. T., & Harris, L. R. (2009). Unintended consequences of using tests to improve learning: How improvement-oriented resources heighten conceptions of assessment as school accountability. *Journal of MultiDisciplinary Evaluation*, *6*(12), 68–91. [CrossRef]

Brown, G. T., Kennedy, K. J., Fok, P. K., Chan, J. K. S., & Yu, W. M. (2009). Assessment for student improvement: Understanding Hong Kong teachers' conceptions and practices of assessment. *Assessment in Education: Principles, Policy & Practice*, *16*(3), 347–363. [CrossRef]

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. [CrossRef]

Chen, Q., Kettle, M., Klenowski, V., & May, L. (2013). Interpretations of formative assessment in the teaching of English at two Chinese universities: A sociocultural perspective. *Assessment & Evaluation in Higher Education*, *38*(7), 831–846. [CrossRef]

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. [CrossRef]

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. [CrossRef]

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. [CrossRef]

Chong, S. W., & Isaacs, T. (2023). An ecological perspective on classroom-based assessment. *TESOL Quarterly*, *57*(4), 1558–1570. [CrossRef]

Chudgar, A., & Sankar, V. (2008). The relationship between teacher gender and student achievement: Evidence from five Indian states. *Compare*, *38*(5), 627–642. [CrossRef]

Cumming, J. J., Maxwell, G. S., & Wyatt-Smith, C. M. (2016). School leadership in assessment in an environment of external accountability: Developing an assessment for learning culture. In G. Johnson, & N. Dempster (Eds.), *Leadership in diverse learning contexts*. Springer.

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., Hondrich, A. L., Rieser, S., Hertel, S., & Hardy, I. (2015). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal*, *52*(6), 1133–1159. [CrossRef]

El Helou, M., Nabhani, M., & Bahous, R. (2016). Teachers' views on causes leading to their burnout. *School Leadership & Management*, *36*(5), 551–567. [CrossRef]

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: A Journal of Comparative and International Education*, *47*(3), 359–373. [CrossRef]

Förster, N., & Souvignier, E. (2014). Learning progress assessment and goal setting: Effects on reading achievement, reading motivation and reading self-concept. *Learning and Instruction*, *32*, 91–100. [CrossRef]

Gerber, J. P., Wheeler, L., & Suls, J. (2018). A social comparison theory meta-analysis 60+ years on. *Psychological Bulletin*, *144*(2), 177–197. [CrossRef]

Glazer, N. (2014). Formative plus summative assessment in large undergraduate courses: Why both? *International Journal of Teaching and Learning in Higher Education*, *26*(2), 276–286.

Goldstein, H. (2011). *Multilevel statistical models*. John Wiley & Sons.

Han, Y., & Xu, Y. (2020). The development of student feedback literacy: The influences of teacher feedback on peer feedback. *Assessment & Evaluation in Higher Education*, *45*(5), 680–696. [CrossRef]
Hansen, B. (2022). *Econometrics*. Princeton University Press.
Harris, L. R., Adie, L., & Wyatt-Smith, C. (2022). Learning progression–based assessments: A systematic review of student and teacher uses. *Review of Educational Research*, *92*(6), 996–1040. [CrossRef]
Hattie, J. (2017). *Backup of Hattie's ranking list of 256 influences and effect sizes related to student achievement*. Visible Learning. Available online: https://visible-learning.org/backup-hattie-ranking-256-effects-2017/ (accessed on 1 November 2017).
Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. [CrossRef]
Hellrung, K., & Hartig, J. (2013). Understanding and using feedback—A review of empirical studies concerning feedback from external evaluations to teachers. *Educational Research Review*, *9*, 174–190. [CrossRef]
Hogan, E., & Payne, B. (2024). A mixed methods study of teachers' use of feedback within middle school social studies classrooms to promote reading comprehension. *Learning and Instruction*, *92*, 101938. [CrossRef]
Hopfenbeck, T. N., Flórez Petour, M. T., & Tolo, A. (2015). Balancing tensions in educational policy reforms: Large-scale implementation of assessment for learning in Norway. *Assessment in Education: Principles, Policy & Practice*, *22*(1), 44–60. [CrossRef]
Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis*. Routledge.
Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*(1), 1–55. [CrossRef]
Huang, F. L., & Moon, T. R. (2009). Is experience the best teacher? A multilevel analysis of teacher characteristics and student achievement in low performing schools. *Educational Assessment, Evaluation and Accountability*, *21*, 209–234. [CrossRef]
Joreskog, K., & Sorbom, D. (2022). *LISREL 12*. Scientific Software International.
Kennedy, P. (2008). *Guide to econometrics*. Wiley-Blackwell.
Khine, M. S., Fraser, B. J., Afari, E., & Liu, Y. (2023). Language learning environments and reading achievement among students in China: Evidence from PISA 2018 data. *Learning Environments Research*, *26*(1), 31–50. [CrossRef]
Kim, S. Y., Liu, L., & Cao, F. (2017). How does first language (L1) influence second language (L2) reading in the brain? Evidence from Korean-English and Chinese-English bilinguals. *Brain and Language*, *171*, 1–13. [CrossRef]
King, R., Chiu, M. M., & Du, H. (2022). Greater economic inequality, lower school belonging: Multilevel and cross-temporal analyses of 65 countries. *Journal of Educational Psychology*, *114*(5), 1101–1120. [CrossRef]
King, R. B., Cai, Y., & Elliot, A. J. (2024). Income inequality is associated with heightened test anxiety and lower academic achievement: A cross-national study in 51 countries. *Learning and Instruction*, *89*, 101825. [CrossRef]
Klieme, E. (2020). Policies and practices of assessment: A showcase for the use (and misuse) of international large scale assessments in educational effectiveness research. In J. Hall, A. Lindorff, & P. Sammons (Eds.), *International perspectives in educational effectiveness research* (pp. 147–181). Springer Nature Publishers. [CrossRef]
Konstantopoulos, S. (2008). The power of the test in three-level cluster randomized designs. *Journal of Research on Educational Effectiveness*, *1*, 66–88. [CrossRef]
Konstantopoulos, S., Miller, S. R., & van der Ploeg, A. (2013). The impact of Indiana's system of interim assessments on mathematics and reading achievement. *Educational Evaluation and Policy Analysis*, *35*(4), 481–499. [CrossRef]
Kuroda, Y. (2022). What does the disclosure of school quality information bring? The effect through the housing market. *Journal of Regional Science*, *62*(1), 125–149. [CrossRef]
Lee, J. (2012). The effects of the teacher–student relationship and academic press on student engagement and academic performance. *International Journal of Educational Research*, *53*, 330–340. [CrossRef]
Lee, S., Turner, L. J., Woo, S., & Kim, K. (2014). *All or nothing? The impact of school and classroom gender composition on effort and academic achievement* (No. w20722). National Bureau of Economic Research.
Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, *83*(404), 1198–1202. [CrossRef]
Liu, H., Liu, C., Chang, F., & Loyalka, P. (2016). Implementation of teacher training in China and its policy implications. *China & World Economy*, *24*(3), 86–104. [CrossRef]
Ma, X., Jong, C., & Yuan, J. (2013). Exploring reasons for the East Asian success in PISA. In H. Meyer, & A. Benavot (Eds.), *PISA, power, and policy: The emergence of global educational governance* (pp. 225–246). Symposium Books Ltd.
MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods. *Multivariate Behavioral Research*, *39*, 99–128. [CrossRef]
Maingot, M., & Zeghal, D. (2008). An analysis of voluntary disclosure of performance indicators by Canadian universities. *Tertiary Education and Management*, *14*, 269–283. [CrossRef]
Mao, Z., & Lee, I. (2022). Researching L2 student engagement with written feedback: Insights from sociocultural theory. *Tesol Quarterly*, *56*(2), 788–798. [CrossRef]

Marinak, B. A., & Gambrell, L. B. (2010). Reading motivation: Exploring the elementary gender gap. *Literacy Research and Instruction*, *49*(2), 129–141. [CrossRef]

Maulana, R., Opdenakker, M. C., & Bosker, R. (2016). Teachers' instructional behaviors as important predictors of academic motivation: Changes and links across the school year. *Learning and Individual Differences*, *50*, 147–156. [CrossRef]

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. [CrossRef]

Meraviglia, C., Ganzeboom, H. B., & De Luca, D. (2018). A new international measure of social stratification. In J. Jarman, & P. Lambert (Eds.), *Exploring social inequality in the 21st century* (pp. 23–51). Routledge.

Mežek, Š., McGrath, L., Negretti, R., & Berggren, J. (2022). Scaffolding L2 academic reading and self-regulation through task and feedback. *TESOL Quarterly*, *56*(1), 41–67. [CrossRef]

Monseur, C., & Adams, R. (2009). Plausible values. *Journal of Applied Measurement*, *10*, 320–334.

Moss, C. (2013). Research on classroom summative assessment. In *SAGE handbook of research on classroom assessment* (pp. 235–256). SAGE Publications, Inc.

Muthén, L. K., & Muthén, B. O. (2018). *Mplus 8.1*. Muthén & Muthén.

Nadelson, L. S., Throndsen, J., Campbell, J. E., Arp, M., Durfee, M., Dupree, K., Poll, T., & Schoepf, S. (2016). Are they using the data? Teacher perceptions of, practices with, and preparation to use assessment data. *International Journal of Education*, *8*(3), 50–70. [CrossRef]

Nguyen, C., & Griffin, P. (2010). Factors influencing student achievement in Vietnam. *Procedia-Social and Behavioral Sciences*, *2*(2), 1871–1877. [CrossRef]

Nortvedt, G. A., Santos, L., & Pinto, J. (2016). Assessment for learning in Norway and Portugal: The case of primary school mathematics teaching. *Assessment in Education: Principles, Policy & Practice*, *23*(3), 377–395. [CrossRef]

OECD. (2016). *How teachers teach and students learn: Successful strategies for school* (OECD Education Working Paper No. 130). OECD. [CrossRef]

OECD. (2018). *PISA 2018 technical report*. Available online: https://www.oecd.org/pisa/data/pisa2018technicalreport/ (accessed on 21 November 2020).

Oo, C. Z., Alonzo, D., Asih, R., Pelobillo, G., Lim, R., San, N. M. H., & O'Neill, S. (2024). Implementing school-based assessment reforms to enhance student learning: A systematic review. *Educational Assessment, Evaluation and Accountability*, *36*(1), 7–30. [CrossRef]

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*, 100416. [CrossRef]

Parsons, T. (2017). The school class as a social system: Some of its functions in American society. In *Exploring education* (pp. 151–164). Routledge.

Perry, L. B., & McConney, A. (2010). Does the SES of the school matter? An examination of socioeconomic status and student achievement using PISA 2003. *Teachers College Record*, *112*(4), 1137–1162. [CrossRef]

Peugh, J. L., & Enders, C. K. (2004). Missing data in educational research: A review of reporting practices and suggestions for improvement. *Review of Educational Research*, *74*(4), 525–556. [CrossRef]

Rahimi, M., & Karkami, F. H. (2015). The role of teachers' classroom discipline in their teaching effectiveness and students' language learning motivation and achievement: A path method. *Iranian Journal of Language teaching research*, *3*(1), 57–82.

Reed, D. (2002). Clearly communicating the learning objective matters! *Middle School Journal*, *43*(5), 16–24. [CrossRef]

Reilly, D., Neumann, D. L., & Andrews, G. (2019). Gender differences in reading and writing achievement: Evidence from the National Assessment of Educational Progress (NAEP). *American Psychologist*, *74*(4), 445–458. [CrossRef]

Rossi, P. E., Gilula, Z., & Allenby, G. M. (2001). Overcoming scale usage heterogeneity. *Journal of the American Statistical Association*, *96*, 20–31. [CrossRef]

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. [CrossRef]

Sarrico, C. S., Rosa, M. J., & Manatos, M. J. (2012). School performance management practices and school achievement. *International Journal of Productivity and Performance Management*, *61*(3), 272–289. [CrossRef]

Schulz, W. (2003). *Validating questionnaire constructs in international studies*. Australian Council for Educational Research.

Seitsinger, A. M., Felner, R. D., Brand, S., & Burns, A. (2008). A large-scale examination of the nature and efficacy of teachers' practices to engage parents: Assessment, parental contact, and student-level impact. *Journal of School Psychology*, *46*(4), 477–505. [CrossRef]

Shen, T., & Konstantopoulos, S. (2017). Class size effects on reading achievement in Europe: Evidence from PIRLS. *Studies in Educational Evaluation*, *53*, 98–114. [CrossRef]

Shih, Y. C., & Reynolds, B. L. (2018). The effects of integrating goal setting and reading strategy instruction on English reading proficiency and learning motivation: A quasi-experimental study. *Applied Linguistics Review*, *9*(1), 35–62. [CrossRef]

Slavin, R. E. (2013). Effective programmes in reading and mathematics: Lessons from the best evidence encyclopaedia. *School Effectiveness and School Improvement*, *24*(4), 383–391. [CrossRef]

Sortwell, A., Trimble, K., Ferraz, R., Geelan, D. R., Hine, G., Ramirez-Campillo, R., Carter-Thuiller, B., Gkintoni, E., & Xuan, Q. (2024). A systematic review of meta-analyses on the impact of formative assessment on K-12 students' learning: Toward sustainable quality education. *Sustainability*, *16*(17), 7826. [CrossRef]

Swart, E. K., Nielen, T. M., & Sikkema-de Jong, M. T. (2022). Does feedback targeting text comprehension trigger the use of reading strategies or changes in readers' attitudes? A meta-analysis. *Journal of Research in Reading*, *45*(2), 171–188. [CrossRef]

Tan, A. L., & Towndrow, P. A. (2009). Catalysing student–teacher interactions and teacher learning in science practical formative assessment with digital video technology. *Teaching and Teacher Education*, *25*(1), 61–67. [CrossRef]

Tan, C. Y., & Liu, D. (2018). What is the influence of cultural capital on student reading achievement in Confucian as compared to non-Confucian heritage societies? *Compare: A Journal of Comparative and International Education*, *48*(6), 896–914. [CrossRef]

Teltemann, J., & Schunck, R. (2020). Standardized testing, use of assessment data, and low reading performance of immigrant and non-immigrant students in OECD countries. *Frontiers in Sociology*, *5*, 544628. [CrossRef]

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. [CrossRef]

Torres, J. O. (2019). Positive impact of utilizing more formative assessment over summative assessment in the EFL/ESL Classroom. *Open Journal of Modern Linguistics*, *9*(1), 1–11. [CrossRef]

Torres, R. (2021). Does test-based school accountability have an impact on student achievement and equity in education?: A panel approach using PISA. In *OECD education working papers* (No. 250). OECD Publishing. [CrossRef]

van Kuijk, M. F., Deunk, M. I., Bosker, R. J., & Ritzema, E. S. (2016). Goals, data use, and instruction: The effect of a teacher professional development program on reading achievement. *School Effectiveness and School Improvement*, *27*(2), 135–156. [CrossRef]

Vygotsky, L. S. (1978). *Mind in society*. MIT Press.

Wiliam, D., & Thompson, M. (2008). Integrating assessment with instruction. In C. A. Dwyer (Ed.), *The future of assessment* (pp. 53–82). Erlbaum.

Winstone, N., & Carless, D. (2019). *Designing effective feedback processes in higher education: A learning-focused approach*. Routledge.

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. [CrossRef]

Wong, V. W., Ruble, L. A., Yu, Y., & McGrew, J. H. (2017). Too stressed to teach? Teaching quality, student engagement, and IEP outcomes. *Exceptional Children*, *83*(4), 412–427. [CrossRef]

World Bank. (2018). *The world development report 2018*. Oxford University Press.

Xuan, Q., Cheung, A., & Sun, D. (2022). The effectiveness of formative assessment for enhancing reading achievement in K-12 classrooms: A meta-analysis. *Frontiers in Psychology*, *13*, 990196. [CrossRef]

Yan, Z. (2021). Assessment-as-learning in classrooms: The challenges and professional development. *Journal of Education for Teaching*, *47*(2), 293–295. [CrossRef]

Yan, Z., & Brown, G. T. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. *Studies in Educational Evaluation 68*, 100985. [CrossRef]

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*(1), 186–208. [CrossRef]

Yan, Z., & King, R. B. (2023). Assessment is contagious: The social contagion of formative assessment practices and self-efficacy among teachers. *Assessment in Education: Principles, Policy & Practice*, *30*(2), 130–150. [CrossRef]

Yan, Z., King, R. B., & Haw, J. Y. (2021). Formative assessment, growth mindset, and achievement: Examining their relations in the East and the West. *Assessment in Education: Principles, Policy & Practice*, *28*(5–6), 676–702. [CrossRef]

Yeh, S. S. (2010). Understanding and addressing the achievement gap through individualised instruction and formative assessment. *Assessment in Education*, *17*(2), 169–182. [CrossRef]

Zee, M., Rudasill, K. M., & Bosman, R. J. (2021). A cross-lagged study of students' motivation, academic achievement, and relationships with teachers from kindergarten to 6th grade. *Journal of Educational Psychology*, *113*(6), 1208. [CrossRef]

Zhai, X., Li, M., & Guo, Y. (2018). Teachers' use of learning progression-based formative assessment to inform teachers' instructional adjustment: A case study of two physics teachers' instruction. *International Journal of Science Education*, *40*(15), 1832–1856. [CrossRef]