Students' formative assessment perceptions: A multilevel analysis of gender and classroom effects in primary schools in Tanzania

Rajabu Abdalah Shafii[a,b,*], Katherine Fulgence[b], Florence Kyaruzi[b], Jean-Louis Berger[a]

[a] Department of Education Sciences, Faculty of Education, University of Fribourg, Rue Faucigny 2, CH - 1700 Fribourg, Switzerland, Switzerland.

[b] Dar es Salaam University College of Education, University of Dar es Salaam, P. O. Box 2329, Dar es Salaam, Tanzania.

(Paper submitted for publication)

Abstract

This study examined whether perceptions of formative assessment (FA) in Tanzanian primary schools arise from individual pupils' genders or the shared classroom environment. Using a cross-sectional survey of 1,558 students nested within 68 classrooms, the research applied multilevel modeling with Bayesian estimation. The results established scalar invariance, which validated latent mean comparisons across gender groups. The findings showed no statistically significant gender differences at the individual level. Approximately 28% of the total variance in FA perceptions was observed at the classroom level. The classrooms with stronger FA practices exhibited no significant gender gaps. Conversely, a substantial subject-specific effect was observed for Swahili, with its classrooms exhibiting a lower level of FA quality than other subjects. The findings suggest that policy should focus on classroom-level interventions to promote educational equity.

Keywords: formative assessment, gender, classroom assessment, classroom environment, multilevel modeling

1. Introduction

Learning assessments involve collecting, synthesizing, and interpreting information to inform classroom-based decision making, support student learning (formative assessments), and evaluate students' performance at a specific time (summative assessment; Razak & Lamola, 2019). Formative assessment (FA) can significantly enhance interactions between teachers and learners and promote equity in the classroom (Black & Wiliam, 1998, 2009). Because FA is increasingly emphasized to promote equity, it is important to determine whether similar patterns exist. However, poorly designed assessment practices could exacerbate gender inequalities during instruction (Rasooli & DeLuca, 2024; Razak & Lamola, 2019; Riddell & Salisbury, 2003). Such assessment practices often rely on teachers' implicit judgement (Bonesrønning, 2008; Copur-Gencturk et al., 2023; Lavy, 2008). Research has indicated that these judgments can favor specific behavior patterns associated with one gender (Cornwell et al., 2013; Voyer & Voyer, 2014). This can worsen inequalities in autonomy, feedback, and learning expectations during instruction (Jones & Myhill, 2004; Tiedemann, 2002). Grading and feedback have been linked to gender disparities in some educational contexts, although the evidence varies across systems (Bonesrønning, 2008; Cornwell et al., 2013; Lavy, 2008). Studies suggest that teachers tend to demonstrate bias between boys and girls in different subjects and educational contexts (Bonesrønning, 2008; Cornwell et al., 2013; Di Liberto et al., 2022; Protivínský & Münich, 2018).

For example, a study in the Czech Republic targeting primary and lower secondary school pupils (grades 4 and 8) reported a gender bias favoring girls in mathematics and language studies (Protivínský & Münich, 2018). Similarly, research in Chile on primary school pupils (Grade 4) reported that the teachers graded boys below girls across subjects (Contreras, 2024). Further studies, such as Angelo (2014) in Portugal, found grading bias against boys in Grade

12 in mathematics and Portuguese. Lavy (2008) in Israel also identified grading bias against boys in grades 10 to 12 across multiple subjects, including chemistry, computer science, mathematics, Bible studies, and biology. Global research consistently documents teacher-driven gender biases across diverse educational systems. However, there is limited evidence of this phenomenon in Sub-Saharan Africa, particularly Tanzania, where educational policy mandates FA and competence-based assessment (Education and Training Policy (ETP), 2014, revised 2023). The updated 2023 edition of the ETP places renewed emphasis on curriculum enhancement, skills certification, and assessment reforms that align with competence-based approaches (Ministry of Education Science and Technology (MoEST), 2023). The policy analyses indicate that while ETP encourages a move toward competence-based, interactive, and assessment-rich instruction, the persistence of established classroom practices has impeded its implementation (Munisi, 2025). This gap highlights the need to examine whether students' perceptions of FA arise from gender patterns or from the shared classroom environment.

Moreover, teachers may provide male pupils with greater autonomy and exhibit more confidence in their abilities, while offering less support to female pupils (Fennema et al., 1990; Tiedemann, 2002). Such biases may stem from non-cognitive skills, including behavior, adherence to rules, and the timely completion of homework, which are among the reasons for such biases (Cornwell et al., 2013; Voyer & Voyer, 2014). Gender differences in classroom assessment have been attributed to teachers' implicit beliefs or grading practices; however, these findings are inconsistent (Bonesrønning, 2008; Copur-Gencturk et al., 2023; Lavy, 2008). These inconsistencies likely arise from variations in assessment types.

Studies have focused on either high-stakes summative grading or daily formative interactions. This disparity reflects traditional gender biases in the classroom, where male pupils are often perceived as independent learners and female pupils as more reliant on social support (Copur-Gencturk et al., 2023). Grades typically serve as both a gatekeeping mechanism and an indicator of academic performance (Brookhart, 2011; The Organization for Economic Co-

operation and Development (OECD), 2013). Given that grades often influence admission decisions to prestigious institutions or programs, even subtle gender biases in assessment can lead to unequal educational opportunities (Brookhart, 2011; OECD, 2013). However, from the perspective of FA, grades are not intended to function as a selection criterion; instead, they are used to provide continuous feedback to support learning (Black & Wiliam, 1998; Nhan, 2024). Research has demonstrated that students who receive high-quality feedback and appropriate scaffolding (the core principles of FA) report favorable perceptions (Kyaruzi et al., 2019).

The Tanzanian education system recently implemented a gender-responsive pedagogy (GRP) policy to mitigate gender bias within both teaching practices and learning materials. The GRP places significant emphasis on the creation of classroom environments that ensure equitable learning opportunities for all students (Thabiti et al., 2025). Despite ongoing efforts, empirical evidence suggests that the effectiveness of GRP is limited because many teachers have insufficient knowledge of its principles (Mhewa et al., 2020; Thabiti et al., 2025). This lack of teacher knowledge is largely due to two key issues: inadequate attention to GRP during teacher training and insufficient opportunities for ongoing professional development (Mhewa et al., 2020). These gaps corroborate evidence from Sub-Saharan Africa, which suggests that teachers have limited assessment literacy (Kanjee & Mthembu, 2015). This is particularly observed in designing, interpreting, and using evidence of learning to improve student learning. Besides more general pedagogical practices, it is important to examine the classroom assessment environment in which gender bias may be evident. Considering Tanzania's recent focus on competence-based assessment (Abdala & Vuzo, 2024; Kahembe & Jackson, 2020), its classrooms offer a significant context for exploring pupils' perceptions of FA. This national shift was formally established in the Tanzania ETP 2014, which identifies competence-based education and assessment as central tenets of curriculum reform. These principles encompass changes to pedagogy and assessment practices (MoEST, 2023).

In addition, the learning environment fostered by teachers has a significant impact on pupils' academic performance, irrespective of gender (Razak & Lamola, 2019). In this context, a classroom's learning environment is defined as a socially constructed setting within which all pedagogical and assessment activities take place (Black & Wiliam, 2009; Creemers & Kyriakides, 2008; Heritage, 2010). This environment is viewed as a dynamic setting in which practices are embedded in social and instructional interactions rather than individual learner attributes (Creemers & Kyriakides, 2008). FA practices include both formal evaluations of academic work and informal observations of learning progress (Black & Wiliam, 2009; Bond et al., 2020; Heritage, 2010). Teachers can use FA principles differently with younger pupils than with older pupils (Browne, 2016; Maskos et al., 2025; Veugen et al., 2024). These differences pose a critical question: How do these variations in teachers' FA practices intersect with pupils' gender? It is of paramount importance to examine how pupils experience FA at the individual and classroom levels, focusing on gender. While pupils must be assessed against defined learning outcomes, adaptable methods are needed (Biggs & Tang, 2011; Black & Wiliam, 1998). A key distinction in assessment is between measuring learning outcomes within a specific curriculum and evaluating a pupil's ability to apply knowledge and skills in diverse real-world contexts (OECD, 2018). Classroom assessments require considerable flexibility to ensure the engagement of both genders (Black & Wiliam, 1998; Lavy, 2008). Despite their importance, most gender-related educational policies focus on teachers, learning materials, and classroom equipment (Razak & Lamola, 2019).

1.1 Formative assessment as a pedagogical promise and its challenges

FA is best understood not only as a set of assessment techniques but also as a complex pedagogical philosophy whose effectiveness is contingent upon its equitable implementation (Black & Wiliam, 2009; Maskos et al., 2025). When enacted effectively, its core principles can transform the classroom learning environment by using constructive feedback to develop self-regulated learning and enhance pupils' learning (Black & Wiliam, 2009; Marchisio et al., 2018;

Maskos et al., 2025). The core principles of FA help pupils take ownership of their learning (Andrade & Heritage, 2018; Black & Wiliam, 2009; Bond et al., 2020). Therefore, it is essential to examine how FA practices are perceived across diverse instructional settings. Tanzanian classrooms are a key setting for understanding how FA affects classroom dynamics and determining whether it functions as intended across genders and subjects. However, in developing countries, the implementation of FA faces challenges. These challenges include limited assessment literacy, large class sizes, and an overreliance on summative assessments (Halai et al., 2018; Kyaruzi et al., 2019; Yongqi Gu & Lam, 2023). The described challenges further limit opportunities for individualized feedback, compounded by the linguistic dynamics inherent in Tanzanian primary education (Ali & Mjenda, 2024; Halai et al., 2023; Kahembe & Jackson, 2020). In this context, Swahili serves dual roles: as its own subject and as the medium of instruction, shaping feedback dialogue and clarifications in daily lessons (Ali & Mjenda, 2024; Kahembe & Jackson, 2020). Unconscious biases among teachers can lead teachers to demonstrate greater confidence in the abilities of male pupils, while providing less instructional support to female pupils (Bonesrønning, 2008; Copur-Gencturk et al., 2023; Lavy, 2008). This imbalance affects iterative feedback and may reinforce traditional gender stereotypes about independent learning (Copur-Gencturk et al., 2023).

Research confirms that the assessment procedures and classroom environment created by teachers greatly shape a pupil's academic performance (Hattie, 2008; Razak & Lamola, 2019). Classrooms vary in how much they foster positive FA experiences (Wiliam, 2011). This is due to a combination of factors, including a teacher's assessment literacy, pedagogical choices such as the quality of feedback and opportunities for goal setting, and implicit beliefs (Shute, 2008). In particular, the teacher's overall approach to feedback and grading helps shape the classroom's learning environment. These elements combine to create distinct assessment environments that can differ significantly from one classroom to another (Alkharusi, 2011; Copur-Gencturk et al., 2023; Maskos et al., 2025). Wiliam (2011) highlights the classroom as

the primary unit for educational change. Therefore, the assessment culture within the classroom significantly impacts the success of classroom instruction. Brookhart (2007) also argues that students' perceptions are influenced by their immediate social environment. Given that FA is primarily implemented through interactions between teachers and pupils during instruction, the concept of the classroom learning environment emerges as the central theoretical lens for understanding its impact (Black & Wiliam, 2009; Heritage, 2010). Since all pedagogical and assessment activities unfold within the classroom environment, they are the most determinant of how FA is perceived and experienced. Pupils cluster in classrooms, and FA is enacted through shared classroom practices. Given these dynamics, FA should be conceptualized as a consolidated construct at the classroom level rather than individual interactions. A methodological shift from an individual pupil to a classroom learning environment is required to understand differences in pupils' FA experiences, particularly those suspected to be associated with gender (Alkharusi, 2011; Bonesrønning, 2008). Hence, assessment environments at the classroom level may be more influential than individual gender identity, which necessitates a multilevel approach.

1.2 The current study

Existing research on gender bias in grading and assessment has largely targeted summative grading and individual-level characteristics (Bonefeld & Dickhäuser, 2018; Contreras, 2024; Niemi, 2010), paying little attention to how the shared instructional environment influences FA. This study aimed to examine whether gender differences, mostly documented in summative grading, also emerged in pupils' lived perceptions of FA. Guided by FA theories, which posit that FA is a socially constructed classroom process (Black & Wiliam, 2009; Heritage, 2010; Bond et al., 2020), this study examined the perceptions of Tanzanian primary school pupils of their teachers' FA practices. Previous research in Tanzania has not examined whether gender differences in pupils' perceptions of their teachers' FA practices arise from individual pupils' genders, classroom environments, or subject-specific factors. This study

separated the sources of variation in these perceptions from within-class (individual) to between-class (shared environment) to answer the following research questions (RQs).

RQ1. Do FA items operate equivalently across gender groups in Tanzanian primary school pupils?

RQ2. Do boys and girls in Tanzanian primary schools differ in their perceptions of their teachers' FA practices?

RQ3. To what extent do classroom-level differences in aggregated FA perceptions correspond with gender gap variations across classrooms?

2. Methods

2.1 Research design

This study used a cross-sectional survey research design. The method facilitates data collection from a large and diverse sample at a single time point (Creswell & Clark, 2018). It provides a comprehensive overview of the phenomenon under study (Fraenkel et al., 2019). In addition, this design facilitates direct statistical comparisons between groups (in this case, boys and girls) and identifies relationships among the variables of interest (Fraenkel et al., 2019). The cross-sectional design accounted for the nested structure of schools and classrooms. This approach enabled the distinction between individual perceptions and the shared classroom environment, which is central to understanding how FA is shared and perceived.

2.2 Sample and sampling methods

Purposeful sampling was used to select schools and classes to capture diverse educational contexts. This study included 37 public schools: 14 in Mwanza, 10 in the Coastal Region, and 13 in Morogoro. These three regions were chosen to ensure that the sample was geographically and socioeconomically diverse. The selection criteria were government ownership, a grade of five, six, or seven, and consent. From these schools, 68 classrooms were selected: 22 from Mwanza, 21 from the Coastal Region, and 25 from Morogoro. Within each school, one class per subject was selected. A total of 1,558 primary school pupils from 68

classes participated, drawn from three geographic regions. The study's population comprised 53.5% (n = 834) girls and 46.5% (n = 724) boys. Since the study focused on gender differences in FA practices, the regional variable was not included in the statistical analysis.

2.3 Instruments and psychometric properties

A set of seven FA items was used to assess students' perceptions of their teachers' FA practices (Luo & Lim, 2024). The seven items were selected due to their established validity in linking perceived FA with students' motivational beliefs and self-regulation (Black & Wiliam, 2009; Zimmerman, 2000). Their primary focus is clarifying learning goals and success criteria, providing feedback and support, and encouraging students to set their own goals. The items emphasize students' responsibility and goal setting; see Table 1 (Luo & Lim, 2024). These items were used to examine how pupils perceived and interpreted their teachers' FA practices in five subjects: Science, Mathematics, Moral Education, Swahili, and English. All items were rated on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Psychometric reliability was evaluated using McDonald's omega (ω). The original instrument reported a reliability of ω = 0.89 for perceived FA (Luo & Lim, 2024). However, the internal consistency observed in this study was ω = 0.75. This likely arises from the linguistic and cultural nuances introduced during the forward and backward translation of instruments into Swahili for the Tanzanian primary school context. Some English terms regarding FA lack a direct conceptual equivalent in Swahili. For instance, the English phrase "clear feedback on how to improve our work" was translated as "*mrejesho wa wazi juu ya jinsi ya kuboresha kazi yetu*." Although technically accurate, pupils may interpret *mrejesho* as any comment from the teacher, including praise or correction, rather than the actionable, learning-oriented feedback implied by FA theory. Similarly, "*kwa wazi*" may be interpreted as "public feedback" rather than "clear feedback," and *kuboresha kazi* may refer to improving future tasks rather than revising current work. Therefore, before multilevel modeling, it was necessary to rigorously test measurement invariance across gender groups.

2.4 Procedure

Ethical clearance was obtained before data collection, and all procedures were conducted according to institutional and national research ethics guidelines in Tanzania. The Dar es Salaam University College of Education approved the present study as part of the Strengthening Teacher Professional Development and Mentorship (STPDM) project. In accordance with the school schedule and the teacher's agreement, the researcher distributed questionnaires to pupils during regular class hours. Parental consent was obtained before the participants began completing the questionnaires. The pupils required between 15 and 20 minutes to complete a paper-and-pencil questionnaire. The researcher explained the meaning of FA and how it appears in lessons and exemplified FA practices, such as feedback that improves learning, clarifying goals and success criteria, and asking pupils to monitor and take responsibility for their learning. This was necessary because, despite items' translation, some English terms regarding FA lack a direct conceptual equivalent in Swahili. The researcher then provided examples and a conceptual overview of FA practices. Subsequently, the researcher guided the pupils in completing the survey by reading each item aloud, sequentially, until the end. These practices ensured that all participants understood the instructions promptly.

2.5 Data analysis

Descriptive statistics were computed to determine the means, standard deviations, and distributional properties for items separated by gender (Table 1). To test whether the items measured the same underlying construct for both boys and girls in the Tanzanian context, a multigroup confirmatory factor analysis was conducted. This test of measurement invariance (RQ1) was a necessary preliminary analysis to determine whether latent mean comparisons for gender (RQ2) could be statistically interpreted. Failure to consider measurement invariance can inflate the effect size of gender differences (Millsap, 2012). Therefore, any observed mean differences might be attributable to measurement bias rather than to authentic psychological

distinctions (Millsap, 2012; Steinmetz, 2013). Data were inspected using Jamovi Version 2.6.44, and Mplus Version 8.3 was used for subsequent analyses.

Table 1

Descriptive analyses (N = 1,558; boys n = 724, girls n = 834)

Item	Gender	M (SD)
My teacher encourages us to take responsibility for our own learning in this subject.	Boys	4.67 (0.79)
	Girls	4.52 (0.83)
My teacher provides us with hints or tips when necessary to facilitate our thinking in this subject.	Boys	4.37 (0.91)
	Girls	4.35 (0.85)
My teacher gives us clear feedback on how to improve our work in this subject.	Boys	4.28 (1.02)
	Girls	4.42 (0.88)
My teacher explains to us the criteria for good work in this subject.	Boys	4.48 (0.92)
	Girls	4.49 (0.88)
My teacher clearly tells us the learning goals to be achieved in this subject.	Boys	4.39 (1.01)
	Girls	4.39 (1.02)
My teacher asks us to check and correct errors in our own work in this subject	Boys	4.50 (0.93)
	Girls	4.33 (1.01)
My teacher encourages us to set our own goals for learning in this subject	Boys	4.39 (0.94)
	Girls	4.55 (0.85)

A hierarchical sequence of models was tested to establish configural, metric, and scalar invariance (Table 2). After establishing measurement invariance, the intraclass correlation coefficient (ICC) was calculated to quantify the proportion of total variance attributable to the classroom level. A multilevel modeling analysis was conducted to answer RQs 2 and 3 regarding gender differences while accounting for the nested structure of students within classrooms. By acknowledging this hierarchical structure, the analysis avoids the atomistic fallacy, which arises when group-level phenomena such as the shared classroom assessment environment are interpreted as purely individual-level traits (Creemers & Kyriakides, 2008). Moreover, the analysis partitioned the variance in the perception of FA into two levels: within-level for individual experiences (F_W) and between level for the shared classroom environment (F_B). The within-level model included subject indicators (Mathematics, Science, Moral Education, and Swahili, using English as the reference subject) as covariates predicting the

individual-level FA factor. Although the study did not include items that measured FA practices at the teacher or classroom level, the multilevel model partitioned student-reported FA perceptions into within-level (individual) and between-level latent factors. These latent factors represent aggregated classroom differences. The between-level factor does not directly reflect a measured "classroom environment"; instead, it captures the variance in student perceptions within each classroom. A random slope for gender was incorporated into the analysis, enabling its effects to vary across classrooms. To manage the complexity of a two-level random slope model, a Bayesian estimator with 10,000 iterations was chosen (Muthén & Asparouhov, 2012). Bayesian estimation is particularly appropriate for complex multilevel modeling involving random slopes and small cluster sizes (Gelman et al., 2013; Kaplan & Depaoli, 2012). The model achieved a potential scale reduction factor of 1.044, indicating that the Markov chain Monte Carlo chains successfully converged on a stable posterior distribution (Jang & Cohen, 2020). This ensured the reliability of the estimates.

3. Results

3.1 Descriptive statistics

Table 1 summarizes the univariate descriptive statistics for all items, separated by gender. The table displays the mean (M) and standard deviation (SD) values for male and female pupils' ratings across the seven items measuring their perceptions of their teacher's FA practices. The results indicate that pupils generally reported positive perceptions of their teachers' FA practices across all FA items, with mean scores ranging from 4.36 to 4.64.

3.2 Measurement Invariance

Before investigating gender differences (RQ2), measurement invariance was tested (RQ1). The stepwise testing approach was utilized to verify both the model-level fit and the equality of parameters across groups (Millsap, 2012). The purpose of this hierarchical testing

procedure was to evaluate whether the construct was measured equivalently across groups at the configural, metric, and scalar levels (see Table 2).

Table 2

Model Fit Statistics for Measurement Invariance (Configural, Metric, and Scalar)

Invariance Level	χ^2	df	p-value	Scaling Factor	RMSEA (90% CI)	CFI	TLI	SRMR	ΔCFI
Configural	75.472	42	0.0012	2.0678	0.032 (0.020, 0.043)	0.960	0.960	0.065	-
Metric	78.068	41	0.0004	1.9979	0.034 (0.022, 0.045)	0.956	0.955	0.065	-0.004
Scalar	78.357	40	0.0003	1.9903	0.035 (0.023, 0.047)	0.955	0.952	0.065	-0.001

The configural model demonstrated a strong fit (CFI = .960, RMSEA = .032), indicating that both boys and girls perceived FA similarly. The metric model, in which the latent factor loadings were constrained to equality, demonstrated a strong fit (CFI = 0.956, RMSEA = 0.034). In the metric model, unstandardized factor loadings were constrained to be equal across gender groups (e.g., Item 2 loading = 1.490 for both boys and girls, Item 5 loading = 1.870 for both). The scalar model, with further constraints applied to item intercepts, also demonstrated a good fit (CFI = .965, RMSEA = .035), meeting the recommended measurement invariance threshold of CFI < .01 (Chen, 2007; Cheung & Rensvold, 2002; Millsap, 2012). Mplus Version 8.3 estimated that the intercepts were equal for boys and girls (e.g., Item 1 intercept = 4.606), suggesting that the baseline levels of the FA items were equivalent for both genders. The findings provide substantial evidence for scalar invariance, confirming that the latent mean comparison between boys and girls is statistically justified.

3.3 Intraclass correlation coefficients

In educational settings, pupils are usually nested within classrooms where a shared learning environment can influence individual responses. These classroom effects are likely to reflect the pupils' shared experiences related to teaching practices and the general learning environment within the classroom. To account for this, the variance of FA items must be partitioned into "within-level" (individual students' variance) and "between-level" (classroom-

level variance). The ICC was calculated to quantify the proportion of total variance attributable to the classroom level. The item-level ICCs ranged from 0.026 to 0.073, while the latent factor ICCs indicated that 22% to 34% (on average, 28%) of the total variance in the perception of FA was attributable to classrooms. Despite their relative modesty, the cluster size was n = 23 of these ICCs, which yields design effects that necessitate multilevel modeling to avoid Type I error inflation.

3.4 Two-level random slope model

A two-level random slope model separated pupil-level perceptions (within-level) from a classroom-level shared environment to provide a broader understanding of gender differences. This approach allows researchers to examine how classroom-level covariates, such as academic subjects, influence the overall assessment environment and the magnitude of gender differences.

3.4.1 Within-level: Individual perceptions

The within-level model examined the relationship among FA items while accounting for classroom-level effects using Bayesian methods. At the individual level, the effect of gender on pupils' FA perception was not statistically significant (F_W factor regressed ON gender: $\beta = .035$, $p = .119$), indicating no significant difference in FA perceptions between boys and girls. Besides the gender slope, the within-level included subject indicators (Mathematics, Science, Moral Education, and Swahili) to examine whether pupils' perceptions of FA varied according to the subject in which they rated their teacher. The results show that none of the subject predictors were statistically significant (Mathematics: $\beta = 0.040$, $p = .490$; Science: $\beta = -0.014$, $p = .496$; Moral Education: $\beta = 0.000$, $p = .500$; Swahili: $\beta = -0.006$, $p = .498$).

3.4.2 Between-level classroom effects

The between-level analysis estimated using a Bayesian method examined the variance of FA perceptions across 68 classroom clusters and how subjects influence these perceptions. The between-level latent factor (F_B) represents the shared classroom variance of pupils' FA

perceptions, with all items loading significantly onto this factor. The covariance between the classroom FA environment (F_B) and the random gender slope (S) was -0.049, corresponding to a strong standardized correlation ρ = -0.826. The negative correlation reveals that, while the overall average is equivalent, high-quality FA practices mitigate local gender gaps, while lower-quality FA environments appear to exacerbate these disparities. Importantly, the null-gender effect indicates no statistically significant difference in boys' and girls' perceptions of FA when considering overall sample mean scores.

3.4.3 Subject effects

The between-level analysis also examined the effects of the subject's specific classroom context on the classroom's latent factor F_B and the gender slope. To include these categorical subjects in the multilevel model, this study employed dummy coding, a standard method in regression-based analyses. In this analysis, Math, Science, Moral Education, Swahili, and English were included as categorical predictors to examine whether each subject was associated with pupils' FA perceptions. The subject of English was selected as the reference, so the dummy variables for Math, Science, Moral Education, and Swahili each represented the difference between that subject and English at the classroom level. Consequently, the predictors of F_B included average classroom performance in four subjects: Math, Science, Moral Education and Swahili. Swahili had a statistically significant negative effect on the classroom FA factor (F_B ON Swahili β = -0.375, 95% CI [−0.733, -0.004]). This demonstrates that Swahili classrooms are perceived to have a lower-quality FA environment relative to other subjects. Furthermore, the results demonstrate that a shared instructional environment shapes how learners experience FA practices. These findings suggest that classroom characteristics, not individual gender differences, are the primary sources of variability in pupils' FA perceptions.

4. Discussion

This study examined whether FA perceptions in Tanzanian primary schools are driven by individual gender identities or by the influence of shared classroom assessment

environments. A significant contribution of the study is its multilevel analysis, which acknowledges that FA is not a static trait but rather a set of dynamic assessment practices enacted through teacher–student interactions within a specific social context (i.e., the classroom). The results of the multilevel analysis revealed significant variability between classrooms. The ICC data for the individual items ranged from .027 to .073, suggesting that approximately 3% to 7.3% of the variance in students' FA perceptions was attributable to the classroom level. The finding that approximately 28% of the variation in perceptions of FA is attributable to classrooms provides empirical support for theories of FA that focus on the classroom environment (Heritage, 2010; Maskos et al., 2025). These models posit that FA practices are inherently embedded in social and instructional interactions rather than in individual learners' attributes. Classrooms with a higher quality FA environment tend to show smaller gender differences in pupils' perceptions. Conversely, in classrooms where FA practices are weaker, gender differences in FA perception become more pronounced. The significant influence of the classroom's assessment environment is underscored by a strong negative correlation ($\rho = -.826$) between the quality of classroom FA and gender differences. These findings demonstrate that, rather than gender differences, learning outcomes across genders are strongly associated with teacher-related practices, including assessment practices and pedagogical alignment. This aligns with the dynamic approach to promoting equity (Kyriakides et al., 2018), which posits that enhancing the quality of teaching factors, such as assessment, is the most effective method for reducing achievement gaps between students. The substantial variation between classrooms confirms that FA is best conceptualized as a classroom-level construct shaped by teachers' assessment literacy, feedback routines, and pedagogical norms.

This study's multilevel evidence clarifies FA as a socially constructed instructional process rather than merely a collection of techniques. A significant between-class variance (28%) indicates that pupils' perceptions of FA are derived from shared classroom routines,

feedback, and interactional norms, rather than individual characteristics. This supports models that view the classroom as the primary unit of educational change, where assessments gain meaning through collective social processes (Black & Wiliam, 2009; Heritage, 2010). Given that FA is co-produced through teacher–pupil dialogue, classroom norms, and shared expectations, it should be considered a classroom-level construct. Research that focuses solely on individual differences within classes may, therefore, obscure the structural mechanisms through which FA influences equity and learning. Consequently, we advocate for a theoretical and methodological shift toward classroom-level analyses to better understand the function of FA in diverse instructional contexts.

At the individual level, none of the subject-specific indicators predicted pupils' perceptions of the FA practices. All subject-specific effects were statistically insignificant and characterized by wide uncertainty intervals. This demonstrates that pupils' FA experiences are not simply a function of the subject they are being taught. Instead, FA perceptions differed significantly in classrooms, highlighting shared teaching methods over subject-specific content in shaping experiences.

This study examined academic subjects as predictors to identify the effects on the classroom environment and gender gap. Using English as a reference subject, the Bayesian posterior estimates unveiled substantial subject-specific variations in Swahili. Swahili classrooms had a significant negative effect on the latent factor, illustrating that they have a lower-quality FA environment than other subjects. This finding reveals a structural paradox within Tanzanian classrooms: Despite Swahili serving as the medium of instruction and theoretically facilitating richer feedback and clarification, Swahili lessons exhibit the weakest FA environments. This paradox suggests that linguistic accessibility alone does not generate dialogic assessment interactions. Instead, long-established pedagogical traditions in language instruction centered on recitation, repetition, dictation, and teacher-led corrective routines override the opportunities afforded by a shared language (Lisanza, 2014; Wawire et al., 2021).

These instructional patterns embed a transmission-oriented culture that leaves little space for collaborative meaning making, co-regulated learning, or iterative feedback cycles. Therefore, the barrier to FA in Swahili classrooms is not linguistic but pedagogical: The historical legacy or rote form-focused language teaching suppresses the dialogic processes upon which FA relies. This suggests that a subject-specific pedagogical culture exerts a stronger influence on FA quality than the medium of instruction, itself. Extending the observations of Pitt and Carless (2022) and Winstone et al. (2022), FA's effectiveness depends on how learning goals and practices are embedded within disciplinary traditions and a subject-specific pedagogical culture. Traditionally, Swahili instruction relies on recitation, repetition, and teacher-led correction methods rather than dialogic meaning making (Lisanza, 2014; Mhewa et al., 2020; Wawire et al., 2021). These practices contrast with FA principles, which prioritize student-centered approaches and provide feedback to improve student learning. Because Swahili is also the medium of instruction in Tanzania, teachers may be likelier to rely on traditional and rote-based assessment methods in Swahili because its practices are deeply embedded in historical linguistic instruction. This instructional orientation may explain why Swahili classrooms exhibited weaker FA environments than other subjects, where problem solving aligns with FA cycles.

The findings revealed a null-gender effect, indicating that boys and girls perceive their teachers' FA practices similarly. The lack of gender differences at the individual level likely reflects the inherent structure of FA as a practice that limits opportunities for implicit bias. While studies often report instances of gender bias in teacher-assigned grades that favor one gender and penalize another across subjects (Contreras, 2024; Protivínský & Münich, 2018), FA practices rely on continuous and interactive dialogue and iterative feedback. Such assessment practices are likely to reduce implicit biases that influence assessment practices (Black & Wiliam, 2009; Murillo & Hidalgo, 2020). As FA shifts the focus of teachers from evaluative scoring to iterative feedback cycles, the implicit stereotypes documented in studies

of high-stakes grading (e.g., Copur-Gencturk et al., 2023) are less likely to affect students' daily learning experiences. This mechanism helps to explain the negative correlation (ρ = -.826) observed between classroom FA quality and gender gaps, suggesting that a robust FA environment systematically suppresses behavioral pathways, though bias commonly affects classroom assessment. Therefore, the null-gender effect is theoretically consistent with models that position FA as a pedagogical system that promotes equity. However, the present data do not allow causal inferences regarding protective effects. Traditional summative grading depends on single evaluations in which teachers rely on holistic impressions, behavioral compliance, or adherence to rules, precisely the areas where gendered stereotypes are most influential (Bonesrønning, 2008; Cornwell et al., 2013; Voyer & Voyer, 2014). Conversely, high-quality FA operates through interactive feedback, the ongoing clarification of learning goals, and dialogic exchanges that require teachers to ground their judgments in observable evidence of learning rather than intuitive impressions (Black & Wiliam, 2009; Heritage, 2010; Maskos et al., 2025).

4.1 Implications

These findings highlight a persistent "atomistic fallacy" in Tanzanian educational policy, whereby interventions often assume that inequity originates within individual pupils' characteristics, such as gender, motivation, or behavior, rather than within the socially constructed classroom setting where learning occurs. The evidence from this study contradicts that assumption: With 28% of the variance in FA perceptions attributable to classroom-level factors, the shared assessment environment exerts considerably greater influence on learners' experiences of fairness than individual demographic traits. Treating fairness as a personal attribute of students obscures the structural role of teachers' assessment literacy, feedback routines, and classroom norms. Instead, fairness should be conceptualized as a property of the instructional setting, itself. To create a fair learning environment, policy should prioritize dynamic, teacher–student-led FA practices (Black & Wiliam, 2009). These practices would

ensure all students receive high-quality feedback and support (Kyaruzi et al., 2019; Maskos et al., 2025). This necessitates a policy shift away from learner-focused remediation and toward strengthening the professional capacity of teachers to enact high-quality, feedback-driven FA practices that create equitable conditions for all pupils, regardless of gender. This aligns with findings that focusing on the quality of classroom processes is a more effective approach to achieving equity than traditional resource-based interventions (Kyriakides et al., 2018; Razak & Lamola, 2019).

These findings provide a critical link to Tanzania's Gender-Responsive Pedagogy (GRP) policy, which aims to mitigate bias in instructional practice. However, the policy's implementation has been limited by teachers' insufficient knowledge of GRP principles (Mhewa et al., 2020; Thabiti et al., 2025). The present results suggest that FA may serve as a practical means through which GRP can be realized in everyday classroom instruction. High-quality FA environments characterized by explicit success criteria, iterative feedback cycles, and opportunities for pupil self-regulation naturally reduce the behavioral pathways through which implicit gender biases manifest. Given that FA structures teachers' evaluative decisions around observable evidence of learning rather than intuitive judgments, strengthening teachers' FA competence may provide a more actionable and sustainable pathway for achieving the aims of GRP than general gender-sensitivity training alone. Positioning FA as an operational mechanism for GRP reframes assessment literacy not as a technical skill set, but as a core equity intervention in the Tanzanian context.

While the findings offer considerable empirical evidence that classroom processes are significant for pupils' experiences with FA, it is important to interpret these results as providing a baseline map of assessment equity in Tanzanian primary schools. The study drew on three regions; however, these do not fully capture the county's cultural, linguistic, and pedagogical diversity. The observed null-gender effect and the subject-specific Swahili effect suggest

conditions specific to the sampled regions and, thus, should serve as an empirical starting point for national dialogue and broader system-level evaluation rather than a universal rule.

4.2 Methodological limitations

Notably, this study's results are not without limitations and must be interpreted with caution. The study used a cross-sectional design, which prevents making definitive causal relationships about the development of perceptions of FA over time and cannot account for the influence of professional development, policy changes, and classroom dynamics on the longitudinal development of classroom FA practices.

In addition, the study exclusively focused on pupils' self-reports, which limits the triangulation of findings with teachers' practices, task-level analysis, and classroom observations. However, self-reports offer a unique perspective of pupils' internal experiences, which is an FA dimension that cannot be inferred from teacher behavior alone. The discovery of the subject-specific Swahili effect originates directly from pupils' experiences. Including qualitative insights could help capture the rich lived experience of FA in the classroom. The geographic focus of this study was limited to the Morogoro, Mwanza, and Coastal regions, which limits generalizability. Correspondingly, it is imperative that the findings are interpreted as region-specific insights that provide a basis for future national evaluations. Nonetheless, the model successfully distinguishes classroom variances. The findings robustly justify FA as a classroom-level construct and provide a baseline map of the FA environment in Tanzanian schools.

4.3 Future directions

The current study proposes a thematic shift in the research focus from the individual to prioritizing classroom learning environments. Shifting the research lens to the classroom level will significantly transform the development of gender-responsive interventions in assessment practices and address the systemic factors that dictate whether a learning environment is perceived as equitable. Future research can investigate how teachers apply feedback and

scaffolding differently across genders and should employ longitudinal methods to track how FA influences high-stakes summative outcomes. Future research also should evaluate whether positive FA experiences translate into equitable access to higher education. Prospective investigations may include a qualitative method, such as ethnography, to uncover the mechanism behind a strong negative correlation between a high-quality FA environment and the magnitude of gender gaps.

References

Abdala, J., & Vuzo, M. (2024). Practices in assessment for learning in English language classrooms within government secondary schools in Tanga City, Tanzania. *Papers in Education and Development*, *42*(1), 189–214. https://doi.org/10.56279/ped.v42i1.10

Ali, H. D., & Mjenda, M. (2024). Teachers' understanding of classroom assessment: Insights from English language teachers in Dodoma municipality, Tanzania. *Cogent Education*, *11*(1), 2380627. https://doi.org/10.1080/2331186X.2024.2380627

Alkharusi, H. (2011). Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience, and in-service assessment training. *Journal of Turkish Science Education*, *8*(2), 39–48.

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge, Taylor & Francis Group.

Angelo, C. S. (2014). *Is there a bias toward girls in non-anonymous evaluation?* Unpublished work.

Biggs, J. B., & Tang, C. S. (with Society for Research into Higher Education). (2011). *Teaching for quality learning at university: What the student does* (4th ed.). McGraw-Hill/Society for Research into Higher Education/Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, *5*(1), 7–74. https://doi.org/10.1080/0969595980050102

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation, and Accountability*, *21*(1), 5–31. https://doi.org/10.1007/s11092-008-9068-5

Bond, E., Woolcott, G., Markopoulos, C., & Southern Cross University. (2020). Why aren't teachers using formative assessment? What can be done about it? *Assessment Matters*, *14*, 112–136. https://doi.org/10.18296/am.0046

Bonefeld, M., & Dickhäuser, O. (2018). (Biased) Grading of students' performance: Students' names, performance level, and implicit attitudes. *Frontiers in Psychology*, *9*, 481. https://doi.org/10.3389/fpsyg.2018.00481

Bonesrønning, H. (2008). The effect of grading practices on gender differences in academic performance. *Bulletin of Economic Research*, *60*(3), 245–264. https://doi.org/10.1111/j.1467-8586.2008.00278.x

Brookhart, S. M. (2007). Expanding views about formative classroom assessment: A review of the literature. In: McMillan, J. H., Ed., *Formative classroom assessment: Theory into practice*, Teachers College Press, New York, 43–62.

Brookhart, S. M. (2011). *Grading and learning: Practices that support student achievement*. Solution Tree Press.

Browne, E. (2016). Evidence on formative classroom assessment for learning. *Knowledge, evidence, and learning for development* (K4D Helpdesk Report). Institute of Development Studies.

Chen, F. F. (2007). Sensitivity of goodness-of-fit indexes to lack of measurement invariance. *Structural Equation Modelling: A Multidisciplinary Journal*, *14*(3), 464–504. https://doi.org/10.1080/10705510701301834

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modelling: A Multidisciplinary Journal*, *9*(2), 233–255. https://doi.org/10.1207/S15328007SEM0902_5

Contreras, D. (2024). Gender differences in grading: Teacher bias or student behaviour? *Education Economics*, *32*(6), 762–785. https://doi.org/10.1080/09645292.2023.2252620

Copur-Gencturk, Y., Thacker, I., & Cimpian, J. R. (2023). Teachers' race and gender biases and the moderating effects of their beliefs and dispositions. *International Journal of STEM Education*, *10*(1), 31. https://doi.org/10.1186/s40594-023-00420-z

Cornwell, C., Mustard, D. B., & Van Parys, J. (2013). Noncognitive skills and the gender disparities in test scores and teacher assessments: Evidence from primary school. *Journal of Human Resources*, *48*(1), 236–264. https://doi.org/10.1353/jhr.2013.0002

Creemers, B., & Kyriakides, L. (2007). *The dynamics of educational effectiveness* (1st ed.). Routledge. https://doi.org/10.4324/9780203939185

Creswell, J. W., & Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). Sage.

Di Liberto, A., Casula, L., & Pau, S. (2022). Grading practices, gender bias, and educational outcomes: Evidence from Italy. *Education Economics*, *30*(5), 481–508. https://doi.org/10.1080/09645292.2021.2004999

Fennema, E., Peterson, P. L., Carpenter, T. P., & Lubinski, C. A. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics*, *21*(1), 55–69. https://doi.org/10.1007/BF00311015

Fraenkel, J. R., Wallen, N. E., & Hyun, H. H. (2019). *How to design and evaluate research in education* (10th ed., International Student Edition). McGraw-Hill Education.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC. https://doi.org/10.1201/b16018

Halai, A., Sarungi, V., & Hopfenbeck, T. N. (2018). Assessment for learning in Africa: Insights from classrooms in Tanzania. *Quality Mathematics Education for All*, 230–234.

Halai, A., Sarungi, V., & Hopfenbeck, T. N. (2023). Teachers' perspectives and practice of assessment for learning in classrooms in Tanzania. In *International Encyclopedia of Education* (4th ed.) (pp. 63–72). Elsevier. https://doi.org/10.1016/B978-0-12-818630-5.09039-4

Hattie, J. (2008). *Visible learning*. Routledge. https://doi.org/10.4324/9780203887332

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. Corwin.

Jang, Y., & Cohen, A. S. (2020). The impact of Markov chain convergence on estimation of mixture IRT model parameters. *Educational and Psychological Measurement*, *80*(5), 975–994. https://doi.org/10.1177/0013164419898228

Jones, S., & Myhill, D. (2004). "Troublesome boys" and "compliant girls": Gender identity and perceptions of achievement and underachievement. *British Journal of Sociology of Education*, *25*(5), 547–561. https://doi.org/10.1080/0142569042000252044

Kahembe, J., & Jackson, L. (2020). *Educational assessment in Tanzania: A sociocultural perspective*. Springer Singapore. https://doi.org/10.1007/978-981-15-9992-7

Kanjee, A., & Mthembu, J. (2015). Assessment literacy of foundation phase teachers: An exploratory study. *South African Journal of Childhood Education*, *5*(1), 26. https://doi.org/10.4102/sajce.v5i1.354

Kaplan, D., & Depaoli, S. (2012). Bayesian structural equation modelling. In R. H. Hoyle (Ed.), *Handbook of structural equation modelling* (pp. 650–673). The Guilford Press.

Kyaruzi, F., Strijbos, J.-W., Ufer, S., & Brown, G. T. L. (2019). Students' formative assessment perceptions, feedback use, and mathematics performance in secondary schools in Tanzania. *Assessment in Education: Principles, Policy, & Practice*, *26*(3), 278–302. https://doi.org/10.1080/0969594X.2019.1593103

Kyriakides, L., Creemers, B., & Charalambous, E. (2018). *Equity and quality dimensions in educational effectiveness.* Cham, Switzerland: Springer.

Lavy, V. (2008). Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment. *Journal of Public Economics*, *92*(10–11), 2083–2105. https://doi.org/10.1016/j.jpubeco.2008.02.009

Lisanza, E. M. (2014). Dialogic instruction and learning: The case of one Kiswahili classroom in Kenya. *Language, Culture, and Curriculum*, *27*(2), 121–135. https://doi.org/10.1080/07908318.2014.912285

Luo, W., & Lim, S. Q. W. (2024). Perceived formative assessment and student motivational beliefs and self-regulation strategies: A multilevel analysis. *Educational Psychology*, *44*(3), 284–302. https://doi.org/10.1080/01443410.2024.2354686

Marchisio, M., Barana, A., Fioravera, M., Rabellino, S., & Conte, A. (2018). A model of formative automatic assessment and interactive feedback for STEM. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, 1016–1025. https://doi.org/10.1109/compsac.2018.00178

Maskos, K., Schulz, A., Oeksuez, S. S., & Rakoczy, K. (2025). Formative assessment in mathematics education: A systematic review. *ZDM—Mathematics Education*, *57*(4), 679–693. https://doi.org/10.1007/s11858-025-01696-x

Mhewa, M. M., Bhalalusesa, E. P., & Kafanabo, E. (2020). Secondary school teachers' understanding of gender-responsive pedagogy in bridging inequalities of students' learning in Tanzania. *Papers in Education and Development*, *38*(2), 252–279.

Millsap, R. E. (2012). *Statistical approaches to measurement invariance* (1st ed.). Routledge. https://doi.org/10.4324/9780203821961

Ministry of Education, Science, and Technology. (2014). *Education and training policy* (2023 revised edition).

Munisi, R. J. (2025). The elements of competence-based education and training in the Tanzania Education and Training Policy 2014 (2023) Edition. *Journal of Research Innovation and Implications in Education*, *9*(4), 785–798. https://doi.org/10.59765/tr7w9l

Murillo, F. J., & Hidalgo, N. (2020). Fair student assessment: A phenomenographic study on teachers' conceptions. *Studies in Educational Evaluation*, *65*, 100860. https://doi.org/10.1016/j.stueduc.2020.100860

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modelling: A more flexible representation of substantive theory. *Psychological Methods*, *17*(3), 313–335. https://doi.org/10.1037/a0026802

Nhan, L. K. (2024). Enhancing teaching and learning through formative assessment. *International Journal of Science and Management Studies (IJSMS)*, 356–365. https://doi.org/10.51386/25815946/ijsms-v7i3p128

Niemi, N. S. (2010). Still failing at fairness: How gender bias cheats girls and boys in school and what we can do about it, by David Sadker, Myra Sadker, and Karen Zittleman. *Gender and Education*, *22*(1), 142–143. https://doi.org/10.1080/09540250903464773

OECD. (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. OECD. https://doi.org/10.1787/9789264190658-en

OECD. (2018). *The future of eduction and skills: Education 2030*. OECD Publishing.

Pitt, E., & Carless, D. (2022). Signature feedback practices in the creative arts: Integrating feedback within the curriculum. *Assessment & Evaluation in Higher Education*, *47*(6), 817–829. https://doi.org/10.1080/02602938.2021.1980769

Protivínský, T., & Münich, D. (2018). Gender bias in teachers' grading: What is in the grade? *Studies in Educational Evaluation*, *59*, 141–149. https://doi.org/10.1016/j.stueduc.2018.07.006

Rasooli, A., & DeLuca, C. (2024). A critical review of fairness from multiple perspectives: Implications for classroom assessment theory. *Applied Measurement in Education*, *37*(2), 148–164. https://doi.org/10.1080/08957347.2024.2345594

Razak, N. A., & Lamola, K. (2019). *Gender equity, equality, and learning assessments*. UNESCO. www.unesco.org/open-access/terms-use-ccbyncsa-en

Riddell, S., & Salisbury, J. (Eds.). (2003). Equity, assessment, and gender. In *Gender, Policy, and Educational Change* (1st ed., pp. 156–174). Routledge. https://doi.org/10.4324/9780203200056-16

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research*, *78*(1), 153–189. https://doi.org/10.3102/0034654307313795

Steinmetz, H. (2013). Analysing observed composite differences across groups: Is partial measurement invariance enough? *Methodology 6*, *9*(1), 1–12.

Thabiti, T. H., Mwadilawa, B., & Basela, J. (2025). Implementation of gender-responsive pedagogy approaches in public secondary schools in Mafia District, Tanzania. *International Journal of Sub-Saharan African Research (IJSSAR)*, *3*(2), 50–62.

Tiedemann, J. (2002). Teachers' gender stereotypes as determinants of teacher perceptions in elementary school mathematics. *Educational Studies in Mathematics*, *50*, 49–62.

Veugen, M. J., Gulikers, J. T. M., & Den Brok, P. (2024). Secondary school teachers' use of formative assessment practice to create co-regulated learning. *Journal of Formative Design in Learning*, *8*(1), 15–32. https://doi.org/10.1007/s41686-024-00089-9

Voyer, D., & Voyer, S. D. (2014). Gender differences in scholastic achievement: A meta-analysis. *Psychological Bulletin*, *140*(4), 1174–1204. https://doi.org/10.1037/a0036620

Wawire, B. A., Piper, B., & Liang, X. (2021). Examining the simple view of reading in Kiswahili: Longitudinal evidence from Kenya. *Learning and Individual Differences*, *90*, 102044. https://doi.org/10.1016/j.lindif.2021.102044

Wiliam, D. (2011). *Embedded formative assessment* (2nd ed.). Solution Tree Press.

Winstone, N. E., Balloo, K., & Carless, D. (2022). Discipline-specific feedback literacies: A framework for curriculum design. *Higher Education*, *83*(1), 57–77. https://doi.org/10.1007/s10734-020-00632-0

Yongqi Gu, P., & Lam, R. (2023). Developing assessment literacy for classroom-based formative assessment. *Chinese Journal of Applied Linguistics*, *46*(2), 155–161. https://doi.org/10.1515/cjal-2023-0201

Zimmerman, B. J. (2000). Attaining self-regulation. In *Handbook of Self-Regulation* (pp. 13–39). Elsevier. https://doi.org/10.1016/B978-012109890-2/50031-7