Towards Sustainable Learning: Analyzing the Fairness of Student, Peer, and Teacher Judgments of Self-Assessment Practices using the Many-Facet Rasch Model (MFRM) in Indonesian Public Islamic University

A Thesis

Submitted to the Master's Study Program of Education at the Faculty of Education in partial fulfillment of the requirements for the degree of

Master of Arts (M.A.)

by:

Fitri Amalia

04212310005

UNIVERSITAS ISLAM INTERNASIONAL INDONESIA

DEPOK

2025

Towards Sustainable Learning: Analyzing the Fairness of Student, Peer, and Teacher Judgments of Self-Assessment Practices using the Many-Facet Rasch Model (MFRM) in Indonesian Public Islamic University

A Thesis

Submitted to the Master's Study Program of Education at the Faculty of Education in partial fulfillment of the requirements for the degree of

Master of Arts (M.A.)

by:

Fitri Amalia

04212310005

UNIVERSITAS ISLAM INTERNASIONAL INDONESIA

DEPOK

2025

ABSTRACT

Fitri Amalia
04212310005
fitri.amalia@uiii.ac.id
MA in Education
Universitas Islam Internasional Indonesia

Self-assessment as a learning strategy offers many benefits, particularly for university students. However, the accuracy of student self-assessment is often debated, which in turn makes peer and teacher assessments often used as comparison tools and standards in determining the accuracy aspect of self-assessment. However, peer and teacher assessments are not immune to bias and inaccuracy. In addition, the forms of rating behavior that arise from students when evaluating self-assessment, as well as the tendency of the practices they adopt, are not well explored, especially in the Indonesian context. Therefore, employing non-experimental quantitative research with a cross-sectional design, this study aims to investigate the practice of self-assessment performed by undergraduate students, explore the forms of rating behavior that arise from self-, peer, and teacher assessments, and explore trends in self-assessment practices implemented by students. Through the Many-Facet Rasch Model (MFRM) approach, the study also aimed to evaluate the psychometric properties of the Self-Assessment Practice Scale (SaPS) instrument adapted to meet the needs of peer and teacher assessors. A total of 2057 responses were collected from the participants, consisting of 31 students from the English Education Program who enrolled in the English Literature course, and a lecturer who taught the course. The data were analyzed using the FACETS computer software program. The results of this study showed that the SaPS had good psychometric properties after being adapted for use by peers and teachers. Then, the analysis found that most students are biased towards underestimating their own abilities. Furthermore, lecturer ratings were found to be more lenient than self- and peer-ratings. In addition, the study showed that items from the SEFI (Seeking External Feedback through Inquiry) dimension were most difficult to agree with, indicating that students tend to be reluctant to seek external feedback. This research highlights the importance of improving students' feedback literacy to maximize the benefits of self-assessment practices, which in turn affect self-regulation and ultimately impact lifelong learning.

Keywords**:** *self-assessment, higher education, many-facet Rasch measurement, rater factors, self-assessment practice scale (SaPS), feedback literacy.*

ACKNOWLEDGEMENT

All praises to Allah SWT for the grace, guidance, and blessings that have been given to me so that I can complete this thesis. This thesis is evidence of a remarkable academic journey over the past two years. A journey full of challenges and endless support from many people. Without His permission and will, this journey would not have been possible.

First of all, I would like to express my deepest appreciation to my esteemed supervisors, Bambang Sumintono, Ph.D., and Associate Professor Charyna Ayu Rizkiyanti, Ph.D, who have provided very meaningful guidance in the completion of this thesis. Thank you for always being very patient with me and for being there whenever I need you. Also, my sincere gratitude to the esteemed examiners, Anindito Aditomo, Ph.D., and Soeharto, Ph.D., for their valuable insights, suggestions, and direction, upon the improvement of this thesis. Your knowledge and expertise are deeply appreciated.

To all lecturers who directly or indirectly contributed to this thesis: Prof. Nina Nurmila, PhD., Associate Professor Tati D. Wardi, Ph.D., Dr. Destina Wahyu Winarti, Dr. Lukman Nul Hakim, Prof. Bambang Suryadi, Prof. Muhammad Zuhdi, Ph.D., Prof. Suwarsih Madya, Ph.D., and many more lecturers whom I cannot mention one by one, I feel very honored to have had the opportunity to learn from such dedicated scholars. Your influence will resonate within me for a lifetime.

Additionally, I am also very grateful to my family, father, mother in *jannah*, sister, and brothers, for your endless prayers, encouragement, and patience. Your unconditional love and support have provided me with a solid foundation so I could pursue my dreams. I would also extend my appreciation to my classmates, the third batch family, who always accompany me on this journey. In UIII, we meet as strangers yet grow like family, we learn together, strengthen each other, and every each of you is precious to me.

Finally, I would like to express my sincere appreciation to every research participant who has contributed to this study. Your commitment to taking part in this process and your contributions have been invaluable to me. In addition to all those whose names cannot be mentioned one by one here, who have played a part in the journey of this thesis, I am eternally grateful. Your love, support, and trust in me have gotten me here, and I dedicate this thesis to each and every one of you.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

APPENDIX

ABBREVIATION DIRECTORY

MFRM	: *Many-Facet Rasch Model*
MnSq	: *Mean Square*
ZStd	: *Z-standardized*
FAM	: *Fair Average Measure*
AaL	: *Assessment as Learning*
AfL	: *Assessment for Learning*
AoL	: *Assessment of Learning*
SA	: *Self Assessment*
SSA	: *Students Self Assessment*
SaPS	: *Self-assessment Practice Scale*
SEFM	: *Seeking External Feedback through Monitoring*
SEFI	: *Seeking External Feedback through Inquiry*
SIF	: *Seeking Internal Feedback*
SR	:*Self Reflection*
CSA	: *Cyclical Self-Assessment*
APA	: *American Psychological Association*
TALIS	: *Teaching and Learning International Survey*

CHAPTER ONE
INTRODUCTION

"*Assessment can contribute to learning, or it can hinder the learning, depending on how the assessment is designed and implemented in a particular learning environment. […] Making assessment act as leverage to facilitate student learning is not only a desirable practice in classrooms but also an important goal of global assessment reforms."* (Yan & Yang, 2022, p.1)

1.1 Background

"Why bother with students' self-assessment when it is the teacher's role to evaluate them?" This is a common question that may arise upon first encountering the term "self-assessment". Such a reaction reflects the traditional paradigm of educational assessment, where assessment is associated with a summative task that serves a specific function, namely to test students' knowledge at the end of the learning (Schuwirth & Van Der Vleuten, 2011). However, studies have shown that assessment no longer serves as a mere testing purpose; it has become an internal part of the learning process itself (Alemdag & Narciss, 2025; Yan & Yang, 2022). Therefore, just as assessment plays an important role in examining the extent to which students learn, it is believed to also act as a learning strategy for students themselves.

Until recently, the most popular form of examination practice in higher education is *assessment of learning* (AoL) (Zulfikar, 2018). In this approach, students' knowledge is assessed after the learning process concludes, typically at the end of a course or module, using summative instead of formative feedback. Although AoL can, in principle, target either knowledge or competence, in practice, it often prioritizes knowledge recall over skill development. Consequently, it provides limited impact on students' motivation and few opportunities to improve their performance, rendering it inadequate for equipping students with future-relevant lifelong learning (Boud & Falchikov, 2007; Ehlers, 2013).

Further, the *assessment for learning* (AfL) approach began to gain recognition. The Assessment for Learning (AfL) approach offers a more advanced view by making assessment an integral part of the learning process itself (Ehlers, 2013). Here, teachers are at the center point of implementing interventions through assessments, where students are not only the subject of assessment, but also the direct beneficiaries of the assessment (Schuwirth & Van Der Vleuten, 2011; Wiliam, 2011). Rather than simply measuring the knowledge, the aim is to increase students' motivation to learn and

support the development of their overall competencies and performance (Stiggins, 2008). However, with the shift towards more self-directed learning (Ehlers, 2020; Ehlers & Eigbrecht, 2024), there is a need to go beyond AfL as the standard practice of assessment in higher education and start adopting the *Assessment as Learning* (AaL) approach.

In contrast to traditional assessment approaches, through direct engagement in assessment tasks, AaL encourages students to acquire new knowledge or further develop their competencies (Yan & Boud, 2022). Instead of being considered just an evaluation method, Yan and Yang (2022) define AaL as a learning strategy. In this approach, student learning emerges from direct interaction with assessment tasks and the accompanying activities. Tasks in AaL are designed to create learning opportunities that go beyond recalling old knowledge, but also encourage the development of metacognitive skills and self-regulation so that students can monitor their own performance and meet their ongoing learning needs (Yan & Yang, 2022). Besides, since active engagement is required in this approach, which in turn will equip students with self-regulation skills essential for lifelong learning (Lee et al., 2019).

In addition, the emergence of AaL aims to complement the implementation of AfL and AoL, which in practice have several issues to address—first, the issue of limitations experienced by teachers. There is a demand to closely monitor the development of each student while at the same time carrying out various administrative tasks and other 'teacher' responsibilities (Boud & Molloy, 2013). Moreover, the OECD TALIS 2018 report also found that teachers spend less than half of their time teaching, with the rest consumed by administrative matters, grading, and classroom management (OECD, 2019). In addition, a large class of students with different needs often makes personalized feedback challenging (Gunawardena et al., 2024; Yan & Brown, 2021). Therefore, integrating AaL as a learning process is increasingly viewed as a pedagogical necessity to complement teachers' limited capacity and to better accommodate diverse student learning needs.

Secondly, the education system has not optimally prepared students for life after completing formal education (Earl, 2013). Traditional assessment tends to put the entire evaluation process on the teacher's shoulders, resulting in shallow feedback and impacting students' opportunities for personal growth, and becomes more complicated when it has a bigger class size (Black & Wiliam, 2009). Similarly, in higher education, the primary goal is to equip students with the foundations for lifelong learning, so that they are able to continue to learn and evaluate themselves independently as they enter

the world of work after graduation (Boud & Falchikov, 2006). However, in previous research, Boud identified a fundamental flaw in existing assessment practices in higher education, which is that conventional assessments fail to prepare students for the challenges of lifelong learning (Boud & Soler, 2016). For example, grading-oriented assessment often ignores students' ability to be critical of their performance. This ability is very important to support independent learning in the future. Therefore, AaL can be seen as the foundation of AoL and AfL, as it inspires students to learn and strengthens the learning process itself (Yan & Boud, 2022).

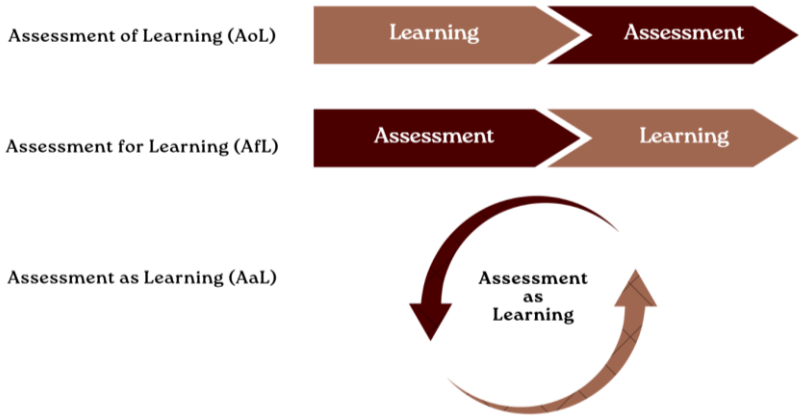


Figure 1.1 *Relationship between learning and assessment in the concepts AoL, AfL, and AsL* (Yan & Boud, 2022, p. 15)

Boud suggested that assessing the effectiveness of assessment should be based on its contribution to students' capacity for self-directed learning. The concept of sustainable assessment was then emerged as to present solution (Boud, 2000). It is identified as an assessment that meets current needs in terms of formative and summative assessment demands, but also prepares students to meet their future learning needs. This means that assessment activities should not only focus on certification or short-term feedback, but should also encourage reflection and independent learning into the future.

One concrete example of sustainable assessment is self-assessment. Self-assessment and self-grading are two distinct concepts and practices, although often misunderstood as the same thing, self-assessment encompasses a broader process (Andrade, 2019). Yan and Brown (2017) proposed self-assessment as a four-stage cyclical process: setting assessment criteria, where students determine their performance standards; self-directed feedback seeking, where students actively seek external and internal feedback; and self-reflection, where students evaluate

weaknesses and strengths based on decided criteria; and calibration, where students readjust the initial criteria based on the judgement from accumulation of feedback (Yan, 2022). The process is dynamic in that the assessment is continually recalibrated as feedback is obtained or criteria change over time.

Although it is called self-assessment, this practice not only involves oneself but also it can involve others too. According to Yan and Carless (2022), one of the most important stages in conducting self-assessment is seeking external feedback, whether from friends, teachers, or parents. This interaction can help students identify blind spots that are not visible through introspection alone (Panadero et al., 2016, 2019). Furthermore, reflective self-assessment can increase students' sense of autonomy, commitment to learning, and self-efficacy (Brown & Harris, 2013). Thus, if implemented properly and consistently, self-assessment can reduce dependence on teachers and increase intrinsic motivation for lifelong learning.

Despite its recognition as an effective approach in promoting students' self-regulated learning (Andersson & Palm, 2017; Andrade & Brookhart, 2020; Panadero et al., 2016; Yan, 2022), many studies criticize its accuracy. Based on meta-analysis studies, the accuracy of self-assessment results is the most frequently scrutinized topic among self-assessment studies (Andrade, 2019). To address the issue of accuracy, the most common practice is to compare the self-assessment with an expert assessment (e.g., teachers or tutors) as the gold standard to assess the construct validity of the self-assessment (Aryadoust, 2015; Brown et al., 2015). Some studies even compare self-assessments with teacher and peer assessments to evaluate their validity (Alias et al., 2015; Chang et al., 2012; Salehi & Masoule, 2017). Based on these standards, self-assessment has been widely criticized for its inaccuracy (Lew et al., 2010) and lack of correlation with competency (Yates et al., 2022).

However, looking back, this never-ending quest for accurate self-assessment is the result of a misunderstanding of the original conception that self-assessment can replace teacher assessment. As a result, accuracy continues to be a concern in this topic. Yan (2016, 2022) argued that the greatest benefit of self-assessment comes from students' active and reflective engagement in self-assessment to develop metacognition, not from accuracy per se, where the focus is on whether students can assess correctly and in accordance with reality. Therefore, the way the accuracy of self-assessment is perceived needs to be refined (Yan, 2022). Instead of a summative assessment method, self-assessment should be seen as a learning process or a learning strategy.

Nevertheless, accuracy remains an essential goal, as students will be more focused on determining areas of improvement when they know exactly what they have and have not mastered (Boud et al., 2013). It will also help them determine the optimal strategies to improve their learning performance (Yan, 2022). Therefore, the self-assessment approach should not only focus on aligning assessment results with teachers' or experts' judgments. Instead, utilizing external judgment should aim to help students exercise their thoroughness and inform them of areas that can only be spotted from the external. By shifting the focus from judging the accuracy to providing insight, it is acknowledged that external assessors themselves are not entirely free from subjectivity (Alemdag & Narciss, 2025; Aryadoust, 2015; Tandiono & Limijaya, 2025; YeşiLçinar & Şata, 2021), and thus their role is better understood as offering complementary perspectives rather than definitive judgements.

In this regard, peer assessment and teacher assessment are frequently used as benchmarks for validating self-assessment (Alemdag & Narciss, 2025). Peer assessment is considered to be contributing in providing relevant alternative viewpoints because students are in similar learning situations, while teacher assessment is considered to offer a professional point of view based on pedagogical experiences that tend to be more extensive. The integration of the three perspectives (self, peer, and teacher) is expected to yield a more comprehensive understanding of self-assessment. Nevertheless, the innate characteristics of the assessors, such as subjective perception factors, differences in judgment standards, personal relationships, or tendencies for leniency and severity in judgment, may affect the objectivity of their assessments. Therefore, it is vital to address these issues with appropriate analytical tools that are able to account for biases.

Thus, in this context, the Many-Facet Rasch Model (MFRM) approach offers an appropriate analytical framework for the needs. MFRM allows the breakdown of influences from various facets in the assessment system, such as differences in severity levels between raters (Aryadoust, 2015), variation in item difficulty levels (Azizah et al., 2021), and individual student characteristics (Aryadoust, 2016), among others. Therefore, MFRM not only increases the validity of the assessment results but also provides diagnostic information on which aspects of each facet that potentially contribute to the accuracy of self-assessment (Engelhard & Wang, 2024). This approach becomes particularly relevant in research involving multiple raters situation in self-assessment involving peer and teacher assessment.

Additionally, the use of the Many-Facet Rasch Model (MFRM) in this study helps overcome methodological limitations that commonly occur in research on multi-rater assessment, where comparisons between raters are usually presented in descriptive or correlational terms without the judging behavior being represented (Riznanda, 2024; Susanti et al., 2023; Wulandari et al., 2020). Although descriptive statistics can deliver surface-level evidence regarding the variances between self-, peer, and teacher ratings, they do not recognize how these variances arise, especially when judging behavior, item difficulty, and student ability interact in complex ways. MFRM facilitates greater insights into these aspects because it tests them concurrently, enabling the researcher to identify systematic trends in scoring behavior and adjust to distortion due to the subjectivity of the rater (Engelhard & Wang, 2024; Linacre & Wright, 2002). The use of MFRM provides a methodological improvement in areas where multi-rater assessments are applied with limited levels of interpretation.

Furthermore, to attain theoretical coherence and level of analysis of self-assessment practices, this paper uses the Cyclical Self-Assessment Model (Yan & Brown, 2017) to view the self-assessment as an integrated, dynamic process of learning. The cyclical model focuses on four stages, i.e., setting criteria, soliciting feedback, self-reflection, and calibration, which are interrelated with each other (Pastore et al., 2025). The stages are linked to each other, and require one stage to add onto the next to increase the level of metacognitive awareness and self-regulation. More specifically, the feedback-seeking stage has been taken to the center of the updated Assessment-as-Learning (AaL) framework (Yan & Boud, 2022), which has put the issue of feedback at the pedagogical center of learning. The applicability of this theoretical underpinning is especially due to learners at the higher education level, who must acquire the skill of learning in autonomous ways as well as critically assess feedback provided by a variety of sources. The cyclical model, therefore, provides a thorough framework in which the practice and its implications can be examined.

In line with this theoretical orientation, this study utilized the Self-Assessment Practices Scale (SaPS), which is directly based on the Cyclical Self-Assessment Model (Yan, 2018). The SaPS is one of the few validated instruments specifically designed to capture student engagement across all four stages of the self-assessment process. SaPS focuses on actual practices as well as beliefs or attitudes, making it particularly suitable for investigating the behavioral aspects of self-assessment. Its structure allows researchers to analyze the extent to which students apply each phase of the self-assessment cycle, while allowing adaptation for use by peer and teacher assessors.

This alignment between the theoretical model and the measurement instrument increased construct validity and interpretability in this study.

Therefore, this study intends to fill the gap of previous research related to the practice of students' self-assessment. By positioning self-assessment as a learning strategy, the use of teacher and peer assessment aims to identify raters' rating behavior that potentially threatens the fairness of self-assessment. Finally, this study introduces MFRM as an alternative of methodological tool that can disentangle complexities as well as identify biases in multi-rater assessments. By implementing this model, not only are the biases of each group of raters revealed, but the tendencies of raters and the implications from the items used in assessment are also analyzed and discussed in the findings. Furthermore, the implications of linking self-assessment as AaL with lifelong learning and sustainable learning are discussed.

1.2 Research Objectives

Drawing on the background of the study, this research explores self-assessment in Indonesian higher education, where assessment remains largely summative and teacher-guided. Despite the increasing popularity of more formative and student-centred forms of learning, the use of self-assessment as a learning strategy remains underdeveloped and little understood. Therefore, the objectives of the study are written as follows:

1) To examine the Self-assessment Practice Scale (SaPS) psychometric quality when used to assess university students' self-assessment practice.
2) To explore how students, peers, and teachers differ in the ways they assess students' self-assessment practice.
3) To identify patterns of rating behavior, such as severity, leniency, or bias, across different groups of assessors.
4) To examine the role of self-assessment as a learning strategy.

1.3 Research Questions

In line with the research objective, this study developed research questions to determine the area of exploration and to guide the analysis systematically. Thus, the questions of the study are formulated as follows:

1) How does the Self-assessment Practice Scale (SaPS) perform in terms of its psychometric quality and unidimensionality when used to assess university students' self-assessment practice?

2) How do self, peer, and teacher assessments differ in the way they reflect students' self-assessment practices?
3) What kind of judging behaviors or rating patterns exist among raters?
4) In what ways do students' self-assessments reveal opportunities or challenges for using self-assessment as a learning strategy?

1.4 The Significance of the Study

Theoretically, this study contributes to the development of self-assessment studies in higher education within the Indonesian context. Specifically, empirical evidence regarding the accuracy of Indonesian students' self-assessment and the potential biases that affect it are discussed in this study. By applying the Many-Facet Rasch Model (MFRM), this study also expands the understanding of how judging behavior can be identified and accommodated quantitatively, thus contributing to the literature on alternative measurement and validity of assessment in education.

Practically, the results of this study can provide insights for educational practitioners in designing and implementing more valid and fair self-assessment practices. By understanding the potential biases in self-assessment and comparing them with peer and teacher assessment results, educators can develop more effective self-assessment training strategies. In addition, the findings of this study can also be utilized to improve the formative assessment system in higher education to better support the development of students' learning independence and self-regulated learning.

CHAPTER TWO

THEORETICAL FOUNDATIONS

This chapter serves as a theoretical foundation for researchers conducting this study. The main focus of this section is divided into two parts: literature review and theoretical framework. The former part discusses how self-assessment has evolved until recently and contains previous research related to self-assessment, both in and outside the Indonesian context. This aims to identify gaps and novelty within the self-assessment topic in the body of knowledge. Moreover, the theoretical framework section contains an explanation of the theories employed as a reference in this research. This section outlines the definitions used in the research and the conceptual framework that guides this research.

2.1. Literature Review

In education, assessment is an important factor since it determines what students consider important, influences their motivation levels to learning activities, and future learning orientation (Boud & Falchikov, 2006; Gibbs & Simpson, 2004; Jimaa, 2011). Conventional thinking behind assessment is on examining the student's progress every so often, which is referred to as assessment of learning (AoL). In turn, the modern strategy lays stress on the regular and constant evaluation, which is built into the everyday classroom activities and teacher-led, known as assessment for learning (AfL) (Murchan, 2017; Wiliam, 2011).

Taking this further, assessment as learning (AaL) places students in control of the assessment process, whereby students continue to monitor, reflect, and control their own learning (Earl, 2013; Yan & Yang, 2022). The shift to bring Assessment of Learning (AoL) to Assessment as Learning (AaL) is an attempt to encourage students to become more active participants in the learning process and to control their self-designed lifelong learning strategies. Nonetheless, this transformation requires a core shift in higher education assessment (and the Indonesian education in particular), where the purpose of assessing immediate knowledge would be replaced with the positive support of a lifelong development process of the students.

2.1.1 Self-Assessment Definition and Core Concepts

Yan and Brown (2017) define self-assessment as a process where students collect information about their performance actively, evaluate it against explicit criteria, and reflect on their strengths and weaknesses to enhance learning. In addition, it is viewed as a reflective and active process that requires students to take initiative,

engage with the purpose of their study, and apply evaluative skills to identify aspects of their work that fall short of expected standards (Andrade & Brown, 2016; Brown & Harris, 2013). In other words, self-assessment is a learning strategy where students use feedback to reflect on their performance in learning against specific criteria.

Practically, there are a lot of ways and strategies by which self-assessment could be applied. In the early stages, self-assessment was more likely to be provided in the form of a static activity (e.g., guessing, scoring one's own work) (Boud & Falchikov, 1989). But this method was not found to have a major effect on learning. The recent trends have changed the orientation of self-assessment into a more dynamic and integrated process in learning through the active participation of students in reflective thinking and decision-making based on their own comprehension about the quality of their own learning (Panadero et al., 2019; Yan, 2022; Yan & Carless, 2022). Within this context, self-assessment no longer functions as a method of assessment but, rather, as an operating method of learning that promotes the formation of metacognition and learning accountability.

2.1.2 Self-Assessment as Assessment as Learning

In the developing world of education, self-assessment is increasingly recognized as a crucial component of an approach known as Assessment as Learning (AaL). This concept sees assessment not merely as an evaluation tool at the end of learning, but as an integral part of the learning process itself (Earl, 2013). Fundamentally, this approach emphasizes the active role of students in constructing understanding through continuous critical reflection. In line with Mendoza and Yan (2021), the essence of self-assessment as an AaL strategy lies in the quality of its implementation, which should extend beyond mere scoring.

Furthermore, Mendoza and Yan (2021) emphatically state that the depth of the reflective process and the active involvement of students in determining assessment criteria are aspects that greatly affect the effectiveness of self-assessment as AaL. Meaning that superficial self-assessment, which only stops at giving grades, without being followed by an in-depth evaluation process of learning outcomes, is no longer considered relevant. On the contrary, meaningful self-assessment should include a series of systematic actions starting from setting evaluation standards by learners, seeking feedback from various sources, to critical reflection on learning progress.

In line with Yan and Brown's (2017) views and findings regarding the cyclical nature of effective self-assessment, ideally, learner should evaluate their positions in the learning journey on an ongoing basis. They assess how far and to what extent goals have been achieved, evaluate the strategies used, and adjust the results for the next step. This dynamic process distinguishes AaL self-assessment from traditional summative approaches.

In higher education settings, the Self-assessment Practice Scale (SaPS) by Yan (2018) is one such instrument that attempts to make self-assessment more of a systematic and theory-based activity rather than an incidental and unplanned one. Through a series of specific behavioral indicators that develop in SaPS, the students are guided not only to analyze the outcomes but also to conduct reflective practices at various levels of the learning process (Yan, 2022).

However, it is important to note that the SaPS is not the only instrument relevant to self-assessment practices. Previous research has leveraged related instruments such as the Self-Report Self-Regulatory Strategies Inventory (Cleary, 2006), the Feedback Seeking Scale (Williams & Johnson, 2000), the Motivated Strategies for Learning Questionnaire (Pintrich, 1991), the Self-Directed Learning Scale (Mok et al., 2006), and the Inventory of Learning Approaches and Skills for Students (Entwistle et al., 2013). These instruments have subscales for seeking feedback, self-monitoring, or self-reflection components that overlap with aspects of self-assessment. However, their limitations lie in their development outside of a consistent self-assessment framework, which often involves general statements such as “I evaluate my performance in learning” without elaborating on the detailed actions of self-assessment.

In response to this gap, the SaPS was developed in accordance with the Cyclical Self-assessment Model (Yan & Brown, 2017), making it one of the few theory-based instruments specifically tailored to understanding self-assessment as a holistic and dynamic process. Unlike previous approaches that bring together unrelated subscales, the SaPS provides a unified framework by focusing on three interrelated actions: seeking external feedback (through monitoring and inquiry), seeking internal feedback, and self-reflecting. This theoretical consistency makes SaPS particularly suitable for research that aims to analyze self-assessment as a learning strategy and not just as a performance evaluation (Yan, 2022).

2.1.3 Self-assessment Accuracy

The issue of accuracy is one of the problems that remain prominent in self-assessment discussions. The problem of having difficulties in establishing valid comparison criteria is one of the primary issues concerning this discussion. Typically, the most frequently used method to get quantitative evidence of assessment alignment is associated with comparing the outcomes of the self-assessment process carried out by students with assessments formed by external actors, e.g., teachers or additional experts (Brown et al., 2015). In this way, the self-assessment concept is condemned to be inaccurate and regarded as of low correlation to the actual performance (Baxter & Norman, 2011; Lew et al., 2010; Yates et al., 2022).

These criticisms, however, are underpinned by a misconception that self-assessment was to be used to replace an expert assessment. Pedagogically, the core advantage of assessment is associated with the emergence of metacognition and the active participation of the student in the educational course, not necessarily with the correspondence of the assessment outcomes to external norms (Andrade & Cizek, 2010; Yan, 2016). Therefore, self-assessment cannot be judged purely on accuracy. What matters more is that it is necessary to consider how far self-evaluation can help develop adaptive and reactive learning skills among students (Tan, 2012; Yan & Brown, 2017).

However, accuracy is still also an essential attribute when it comes to learning, since it enables more specific learning choices by giving students perceptions of self-performance (Boud et al., 2013). The students will be able to pay more attention to areas that they should improve when they can also realistically assess their accomplishments. Thus, accuracy as the goal to be reached completely in each self-assessment activity is not a necessary requirement. Still, the development of accuracy can be regarded as the relevant goal that is to be pursued, as well as the development of students' metacognitive skills and self-directed learning.

Multiple factors have been singled out as determinants of self-assessment precision. Accuracy is affected by individual characteristic factors, including age, level of skill, and experience in the area of performing (Brown & Harris, 2013; Topping, 2003; Yan et al., 2023). More accurate self-assessment is produced by students whose level of proficiency is higher. There is also the bias caused by the design and implementation of self-assessment. Self-assessment accuracy has been

seen to be enhanced by the provision of support through appropriate training (Li & Zhang, 2021), defined evaluation parameters (Kostons et al., 2012), and attention to specific process components (Panadero et al., 2016).

Several studies consistently demonstrate that self-assessment can generate bias. For instance, Tucker (2017) found that in group work, students tend to overrate their contribution to the task by scoring their self-assessment higher than their peers. Similarly, Johnston and Miles (2004) reported that students tend to rate higher on self-assessment compared to their peers, especially when the score influences the final grade, and this is reinforced by Sridharan et al. (2019) study, that peer assessment shows significant bias, particularly when it is used as the summative purpose. On the other hand, a study by Tandiono and Limijaya (2025) reveals that culture influences the bias in students' assessment, displaying a tendency to downplay their real academic achievements and be lenient towards friends due to social expectations and to maintain relationships.

Some other studies have also explored the link between the presence of bias in self-assessment and the level of students' academic performance. Sridharan et al. (2019) reported that students with lower academic achievement were likely to overrate themselves, while the high-performing students showed a more humble attitude in their self-assessment. This pattern is believed to be associated with students' position on the course grade, which influences how they perceive themselves in the context of assessment (Leach, 2012). High-achieving students can also experience greater levels of anxiety, which arises from a fear of failure and encourages them to underestimate their own abilities as a form of unconscious internal motivation to keep improving (DeLong, 2011). In addition, their humble attitude and concern that their self-assessment might not be consistent with that of their friends or lecturers also contribute to their lower scores (Aryadoust, 2015).

2.1.4 Self-assessment Calibration through Peer and Teacher Assessment

Effective learning requires an accurate assessment, and often, students struggle to apply objective assessment to themselves. Panadero et al. (2016) point out the importance of calibration, which can be defined as a match between self-assessment and external reference, such as peer or instructor feedback, which is crucial in improving the accuracy of self-assessment. Without these reference points, students can miss their own mistakes in those self-evaluations, which can lead to misleading cognitive beliefs. As an example, lower-ability students tend to overestimate their

work, whereas higher-ability students may underestimate their abilities (Boud & Falchikov, 1989; Brown & Harris, 2013). Students are thus able to recognize these biases and enhance their evaluations through the use of objective standards given by others through the calibration process.

An important influencing factor of calibration determination is the degree of student mastery. Novices also tend to think their skills are higher because they have more cognitive loads and less knowledge at their disposal (Khonamri et al., 2021). Conversely, better-equipped students who have a better understanding of what is to be done tend to make better self-examinations. To help novices, they should adopt scaffolding in a structured manner, e.g., simplified rubrics or division of tasks into smaller components, which will reduce cognitive burden on an individual and improve their calibration (Kostons et al., 2012). Moreover, when educators use self-assessment tools, it is possible to enhance the recognition of strengths and weaknesses in the students by explicitly training them in the use of these tools (Yan et al., 2020).

Also, calibration depends on contextual and cultural differences because it determines its effectiveness. As an example, students in Indonesia claimed to be more uncomfortable and stressed when evaluating themselves rather than their classmates (Nawas, 2020). Meanwhile, in Malaysia, students emphasize the importance of face-saving actions, the suppression of negative evaluations, heaping scores as a strategy to minimize ambiguity, and relationship-enhancing scales (Cheah et al., 2018). This reveals the necessity of both a contextually and culturally distinct self-assessment practice.

In addition, calibration itself is not an activity that appears only once; it is a skill that is refined over time (Panadero et al., 2016; Yan, 2022). Prolonged studies like those done on medical students show that a cyclic process of self-determination and appraisal by peers or educators in conjunction would eventually increase calibration significantly (Khoiriyah & Roberts, 2025). As argued by Panadero et al. (2013), the accuracy of self-evaluation is a type of construct validation. What this means is that more accurate or realistic student self-ratings result in higher construct validity. This kind of precision shows that the students are able to criticize themselves adequately.

2.1.5 Many-Facet Rasch Model in Rasch Measurement

To analyze assessment data and evaluate the quality of instruments used to measure assessment, the Rasch model is an approach that has been widely used in educational measurement. Developed by Georg Rasch (1960), this model is based on fundamentally on the concept of comparison (Engelhard & Wang, 2021). This approach assumes that there are two things that affect a person's performance in answering an item, namely the difficulty level of the item itself and the ability of the individual concerned (Boone et al., 2014). In other words, it is based on a simple idea about what happens when a single person encounter a single item (Rasch, 1960/1980). In its most basic form, the Rasch model works by calculating the probability of a person answering an item correctly or incorrectly, based on a certain level of ability (Andrich & Marais, 2019; Sumintono, 2015). The main advantage of this model is that, on the same linear scale, it is able to place the individual ability of each participant and the item difficulty of each item, making it easier to interpret the results.

However, in the world of education itself, the form or model of assessment has several variations, not merely about how smart someone is or how hard someone has to work in order to be able to answer questions and get a good score. In many situations, there is a need for objective assessment, such as self-assessment, peer assessment, and lecturer assessment. In these assessment situations, there will be some additional factors that can affect the assessment score. For example, if there are multiple raters, then each individual rater may have different tendencies in giving scores (Eckes, 2015; Linacre & Wright, 2002). Some raters tend to be stricter in giving scores, while others are more generous (i.e. lenient rater). In such cases, the basic Rasch model is not able to capture these additional variations as it only considers individual ability and item difficulty.

Therefore, Linacre (1999) extended the basic Rasch model by developing the Many-Facet Rasch Model (MFRM), which has more facets than the traditional assessment situation that only involves two facets, namely individuals and items. The word 'facet' in MFRM refers to any element that can systematically affect assessment results, such as the rater, rating scale, task conditions, or other relevant factors. By including these facets, MFRM can estimate the influence of each facet on the assessment score simultaneously (Boone et al., 2014; Eckes, 2015; Engelhard & Wang, 2024). As in the context of this study, self-assessment involving peer assessment and lecturer assessment can be estimated by MFRM. Not only student

ability and item difficulty, but also the level of strictness or looseness of assessment from each group of assessors can also be estimated.

In MFRM analysis, the output or result obtained will be a logit measure (log odds unit) for each input facet, such as item difficulty measurement (first facet), individual participant ability (second facet), and rater severity or leniency (third facet). These three measures are placed on the same linear scale to allow for meaningful comparisons, called as Logit (logarithm odd probability unit) (Engelhard & Wind, 2018). For example, a rater who produces a high logit severity score can be interpreted as having a tendency to be more strict in giving scores; conversely, a rater who produces a lower logit score means that they tend to be more lenient in giving scores. In addition, to see the extent to which each element matches the expectations of the model, MFRM also produces fit statistics. If there is a misfit to the value of this statistic, it means that there are inconsistencies in the scoring behavior or there are unexpected response patterns detected (Bond et al., 2021). Through this analysis, MFRM provides a more detailed picture of the interaction between the three facets (individual, item, and rater), where it is compared to the ideal model (which is, in this case, MFRM of the Rasch model).

Therefore, the Many-Facet Rasch Model is very relevant to be used in this study, because the data being analyzed involves multiple raters (self, peers, and lecturers), each of whom may have different tendencies in making judgments. In addition, MFRM allows researchers to control for possible bias and inconsistency between raters, resulting in a more accurate estimate of student ability (Engelhard & Wang, 2024). Also, the model is able to identify patterns such as the degree of leniency, severity, inconsistency, and bias of each of these raters. This is crucial for answering the research questions related to self-assessment accuracy, inter-rater differences, and factors that may affect assessment accuracy in this study.

The mathematical model equation of the Many-Facet Rasch Model for the rating scale version in this study is presented as follows:

$$\text{Log}\,(P_{nijk} / P_{nij(k-1)}) = B_n - D_i - C_j - F_k$$

where

P_{nijk}	is the probability of ratee n being rated, on item i by rater j, a rating of category k
$P_{nij(k-1)}$	is the probability of ratee n being rated, on item i by rater j, a rating of category k-1
B_n	is the Ability of ratee n, where n = 1,N

D_i is the Difficulty of item i, where i = 1,L

C_j is the Severity of rater j, where j = 1,J

F_k is the Height of step k, where k = 1,K

(Linacre & Wright, 2002)

In this Facets model, every item test is characterized by a difficulty (D_i), every ratee is characterized by ability (B_n), and every rater is characterized by a level of severity (C_j). The log odds formulation of (1) puts these parameters on a common scale of log odds units or commonly referred to as "logits" (Linacre & Wright, 2002). Raters use a rating scale to assess the ratee's performance on each item. Each level on the scale shows a clear performance improvement. The symbol F_k has just one subscript, which means that the rating scale works the same way for all tasks and judges.

2.2 Theoretical Framework

This research is based on two main theoretical frameworks that serve as a conceptual foundation in understanding the self-assessment process conducted by students. The two theories are the Self-Assessment Process Model by Yan and Brown (2017) and the Assessment-as-Learning (AaL) concept, which is an advanced framework developed by Yan and Boud (2022). In the context of accuracy and rater bias, which is the focus of this study, these two theoretical frameworks provide a complementary conceptual foundation for understanding the student self-assessment process.

The Self-Assessment Process Model developed by Yan and Brown (2017) sees self-assessment as a series of cyclical processes involving a number of cognitive and metacognitive stages. The establishment of assessment criteria by students is the initial stage of this process, where students must have clear standards to use in assessing their performance. Next, students engage in feedback seeking, both from internal and external sources. Internal feedback involves reflecting on the student's own emotions, motivation, and condition, while external feedback can be input from teachers, classmates, or other learning resources. This process then continues into the self-reflection stage, where students reflect on the information gathered to assess their learning progress. Referring to the results of reflection and feedback that has been obtained, students finally make a self-assessment judgment. Thus, this model emphasizes students' active involvement in the whole process and underlines the role of feedback as a calibration mechanism that can improve the accuracy of their self-assessment.

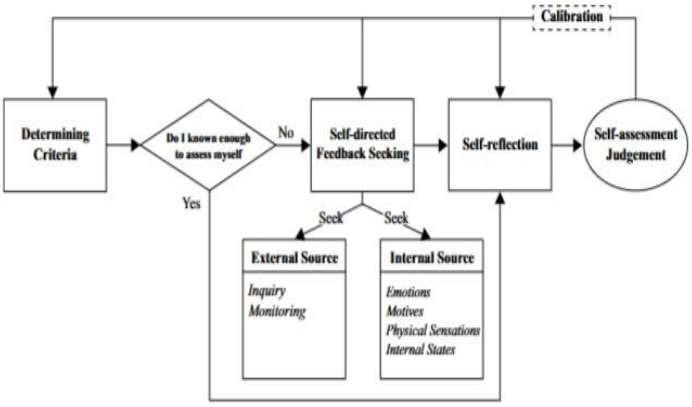


Figure 2.1. *Self-Assessment Process Model (Yan & Brown, 2017)*

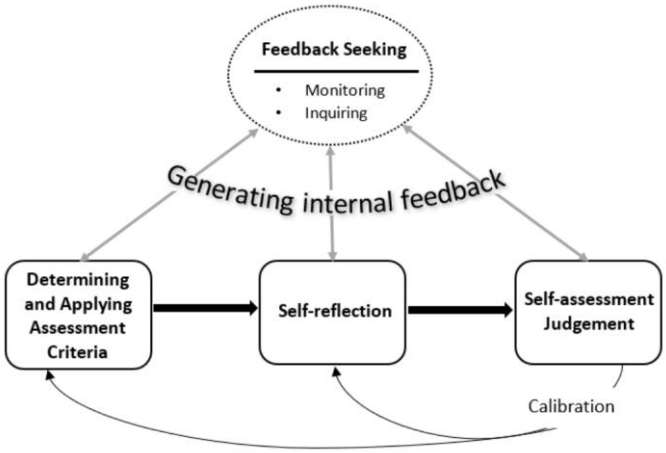


Figure 2.2. *Assessment-as-Learning Model (Yan & Boud, 2022)*

Meanwhile, the Assessment-as-Learning (AaL) framework, further developed by Yan and Boud (2022), integrates the concept of feedback literacy in the self-assessment process. In this framework, self-assessment is not only about the activity of assessing learning outcomes, but also as a core part of the learning process itself. Students are required to be able to utilize the feedback effectively, to regulate learning, build self-awareness of weaknesses and strengths, and develop sustainable reflective skills. The concept of feedback literacy emphasized in this model demands the ability of students to understand, evaluate, and apply feedback in the context of their learning (Yan & Carless, 2022).

Therefore, both theoretical frameworks were used to understand how students engaged in the self-assessment process in this study, which took place naturally in their learning environment. Although this study used a design that did not involve any intervention during the data collection period, the instrument used covered the various dimensions of the self-assessment process mentioned by Yan and brown (2017). As such, this study allows for the measurement of the extent to which

university students have independently engaged in the self-assessment process, as well as how such engagement relates to the accuracy and bias in their self-assessments.

In addition, to explore the fairness and rating behavior in student self-assessment, a comparison of assessment results from three sources of self, peers, and teachers was conducted. By using the Many-Facet Rasch Model analysis approach, it is possible to detect the accuracy and bias of the assessment. By integrating these two theoretical frameworks, it is hoped that this research can provide a comprehensive understanding of the dynamics of the self-assessment process in the context of learning in higher education.

2.3 Conceptual Framework

This study employed the theories of the Self-Assessment Process Model and the Assessment-as-Learning Model to systematically investigate and understand how students conduct self-assessment and how they utilize feedback in the context of learning. This framework allows the researcher to analyze how university students engage in self-assessment, as well as what steps they take in evaluating their learning. In addition, this framework became the foundation for the Self-assessment Practice Scale (SaPS), the self-assessment instrument developed by Yan (2018). In this instrument, there are four dimensions that measure the process of self-assessment, which focus on feedback seeking and self-reflection done by students: *seeking external feedback through monitoring (SEFM), seeking external feedback through inquiry (SEFI), seeking internal feedback (SIF),* and *self-reflection (SR).* With this instrument, the researcher tested the accuracy of the self-assessment conducted by university students by calibrating it with external assessors, including peers and lecturers.

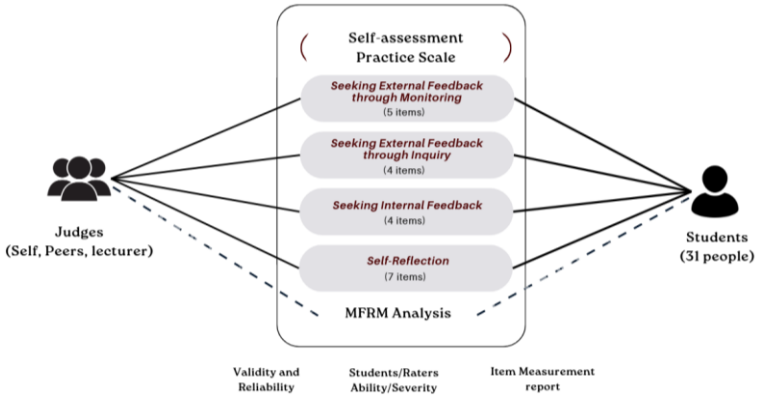


Figure 2.3. *Self-Produce Conceptual Framework*

The conceptual framework illustrated in Figure 2.3 explains the flow of this research. The two theories served as a foundation, then became the basis for examining the self-assessment practices of university students. Adhering to these theories, SaPS was used to obtain self-assessment information from related students, peers, and the course lecturer. Then, the assessment data from these three groups of assessors were analyzed using the Many-Facet Rasch Model (MFRM), the analytical framework used in this study to analyze the findings. Thus, this conceptual framework not only guides the process of operationalizing theory into research design but also serves as a conceptual map that connects the theories, instruments, and analytical methods used in this study to answer the research questions in a systematic and directed manner.

CHAPTER THREE

METHODOLOGY

This chapter describes the methodology used in collecting data and selecting samples to answer the research questions. The explanation in this section includes the research paradigm that forms the basis for how this research was designed and implemented, including data collection procedures, instruments, and data analysis methods, which use Many-Facet Rasch Measurement (MFRM). Efforts have been made to ensure that the methods employed are appropriate, robust, and easy to follow.

3.1. Research Paradigm

The positivist paradigm serves as the foundation of this research. This paradigm believes that human behavior and its traits can be observed, measured, and analyzed systematically. Positivism is characterized by its way of utilizing empirical observation, logical deduction, and objective pursuit of knowledge (Ali, 2024). Therefore, the researcher's position in this study is as an observer of a phenomenon.

Under this paradigm, the study employs Many-Facet Rasch Measurement (MFRM) approach in analyzing the data. Rasch measurement is grounded on a philosophical idea that latent traits, such as ability, judgement accuracy, as well as attitudes, could be meaningfully and precisely evaluated (Engelhard & Wang, 2024). Furthermore, it is possible to construct fair and consistent measures that focus on individuals. In this case, measuring how individuals interact with different elements of assessment, such as tasks or raters.

This is in line with the research objectives, which are to examine students' self-assessment practices seen from different raters (peer and teacher) evaluations, and whether any of the rater groups (self, peers, teacher) are more lenient or severe in terms of assessment compared to the others. This approach allows the researcher to be concerned with underlying patterns or tendencies in the data, not just the scores seen on the surface.

3.2. Research Design

This study applies a quantitative, non-experimental, cross-sectional design. Each component of this design was chosen to fit the purpose of the study and the nature of the data. First and foremost, quantitative research is an approach that tests theories by examining the relationship between variables using instruments to measure

information numerically by utilizing statistical procedures (Creswell & Creswell, 2018). In this study, a statistical tool is being used in analyzing the facets, where the information should be transformed into numbers to be processed by the specific software program, also called FACETS.

Then, a non-experimental design is carried out in the absence of manipulation from the researcher and lack of random and the lack of random group assignment (Johnson & Christensen, 2025). Participants were asked to provide self, peer, and teacher ratings using the SaPS instrument in a natural classroom setting without any interference before or after the data collection process.

Finally, the study followed a cross-sectional design, with data collected at a single point in time during the final weeks of the mid-semester in a higher education institution in Indonesia. A cross-sectional design is the opposite of a longitudinal study, which does data collection over an extended period of time (Mills et al., 2010). A total of one class of students participated by rating themselves and their peers, along with one lecturer who rated the students using the adapted SaPS items. All data were collected simultaneously over a one-week period, thus providing a snapshot of participants' self-assessment practices and rater behaviors. This design allowed the researcher to explore whether students' self-assessments with those of their peers and teachers were fair in terms of their judging behavior, and whether rater characteristics (e.g., severity or bias) influenced the assessments.

3.3. Research Participant and Data Collection Methods

The study is located at a public university in Banten Province, Indonesia. The participants are drawn from a single class of undergraduate-level English education department students enrolled in the English Literature course in their second year of study. The course was chosen according to its appropriateness and suitability for assessment activities that involve multiple assignments and presentations. The more assignments students have, the more chances they have to reflect on their performance. This situation is ideal to examine their self-assessment behavior.

As for the rater participants, there will be three different groups of raters, each addressed as self-group, peer-group, and teacher-group. The first group of raters, the self-group, consists of students who performed the self-assessment. This group comprises 31 students, or the total of the class members. The second group of raters includes students who rate their peers' performance, the self-group students. This peer-group student comprises the same 31 students from the exact class who, after rating

themselves, then proceed to rate two different friends each. That made up two peers assigned to rate each student. Lastly, the teacher involved is a lecturer who teaches the students in the literature course. This teacher rater assessed 10 students in the self-group. The detailed demographic information of the student participants is displayed in Table 1 below.

Table 3.1. *Demographics of students*

Demographic	Frequency	Percentage
Gender		
Male	7	22.6%
Female	24	77.4%
Parents job		
Government sector	7	22.6%
Private sector	10	32.3%
Informal sector	11	35.5%
Domestic work	3	9.7%
Parents Education Level		
Secondary High	10	32.3%
Diploma	2	6.5%
Undergraduate	18	58.1%
Masters	1	3.2%

In addition, one female lecturer who has been teaching the student participants for two semesters was involved in the study. With a background in a Master's degree in English Education, and 2 years of teaching experience, the lecturer gave insights from the teacher rater perspective. Hence, all assessments done by the raters used the same core instrument, the self-assessment practice scale instrument (SaPS), which was slightly modified in language to facilitate clarity for each of the rater groups (self, peer, and teacher). In addition, ratings were collected via paper forms to avoid confusion, particularly for the teacher.

3.4. Research Instrument

This study utilized the Self-assessment Practice Scale (SaPS) developed by Yan (2018). The original version of SaPS is rooted in the Cyclical Self-assessment Framework by Yan and Brown (2017). The modification was done to the instrument

only at the wording adjustment so that it is appropriate for peer- and teacher-raters. In addition, using the same instrument with slight wording modification facilitates comparison and supports the data analysis purpose. Since the study uses Many-Facet Rasch Measurement (MFRM) to analyze information, there is a need to use the same measurement tools to fairly examine the raters' performance.

The instrument consists of 20 items that measure specific self-assessment-related practice across four dimensions: *seeking external feedback through monitoring (SEFM), seeking external feedback through inquiry (SEFI), seeking internal feedback (SIF), and self-reflection (SR).* With 5 items, the SEFM dimension includes items that measure students' practices and strategies in monitoring their learning using external evidence, such as past tests, reference books, and the internet against legitimate standards. Furthermore, with 4 items, the SEFI dimension measures students' practices and strategies in monitoring their learning by looking at subjective evidence from others, such as teachers, friends, and parents. The SIF dimension, with 4 items, focuses on evaluating students' internal feedback. The four items in this dimension emphasize aspects such as feelings, emotions, physical conditions, and subjective internal conditions that can affect students' assessment and learning experience. Finally, the SR dimension, with seven items, evaluates students' actions in reviewing and reflecting on their performance and learning outcomes.

In the pilot study, all items were rated using the original format, a 6-point Likert scale, where 1 = strongly disagree and 6 = strongly agree. The use of this scale is intended to test how respondents respond to items in their original form without any initial modification. The results of this initial pilot became the basis for consideration to maintain or readjust the scale for the next stage based on empirical findings from the pilot data. Although there are three different instruments for each rater group, each version maintained identical item structure (e.g., "Saya" became "*Mahasiswa ini*" or "*Teman ini*"). Examples of the adapted items across rater groups are displayed in Table 2 below:

Table 3.2. *Examples of Instrument Items.*

Construct	Original version	Self-assessment	Peer-assessment	Teacher-assessment
SEFM 1	I check whether I have mastered the course	Saya mengerjakan latihan	Teman ini mengerjakan latihan	Mahasiswa ini mengerjakan latihan

	content by doing extra exercises.	tambahan untuk mengecek apakah saya telah menguasai materi perkuliahan	tambahan untuk mengecek apakah dia telah menguasai materi perkuliahan	tambahan untuk mengecek apakah dia telah menguasai materi perkuliahan
SEFFI 1	I ask my teachers to give me feedback about my performance.	Saya meminta feedback dari dosen mengenai performa belajar saya.	Teman ini meminta feedback dari dosen mengenai performa belajarnya.	Mahasiswa ini meminta feedback dari dosen mengenai performa belajarnya .
SIF 1	My gut feelings tell me whether my work is good or bad.	Saya mengandalkan firasat untuk memberitahu apakah saya telah mengerjakan tugas dengan baik atau tidak.	Teman ini mengandalkan firasat untuk memberitahu apakah ia telah mengerjakan tugas dengan baik atau tidak.	Mahasiswa ini mengandalkan firasat untuk memberitahu apakah ia telah mengerjakan tugas dengan baik atau tidak.
SR1	I seek out the reasons for mistakes I made after getting back marked work.	Saya mencari tahu sebab dari kesalahan yang saya buat setelah menerima hasil koreksi.	Teman ini mencari tahu sebab dari kesalahan yang ia buat setelah menerima hasil koreksi.	Mahasiswa ini mencari tahu sebab dari kesalahan yang ia buat setelah menerima hasil koreksi.

To ensure the validity of the adapted instrument, the study follows a cross-cultural validation approach (Hambleton & Lee, 2013). The instrument was originally adopted in English and then translated into Bahasa Indonesia following a forward-translation design, following the updated procedure by Hambleton and Patsula (1099). Furthermore, terminology was refined to keep the items suitable for participants' backgrounds. The instruments were piloted before the full data collection process with a small sample of students (n = 25) to check clarity, reliability, and item fit, as well as the adaptation results.

3.5. Pilot Testing

To test the validity and reliability of the instrument, a pilot study was conducted. Pilot testing is a test conducted before the main data collection on a small number of samples, which aims to assess whether there are deficiencies or problems that need to be corrected in the test instrument (Mat Roni et al., 2020). At the initial stage, the translated instrument's question items were validated by two experts, then readability tests were conducted on samples with similar characteristics to the target participants. After going through this process, the instrument was then checked again by the academic advisor, and finally, the pilot test was conducted.

First, the index of the item separation results is checked. The value of this index tells information about the distribution of items, or the grouping of item difficulty levels, that can measure participants' abilities (Boone, 2020). The results of the item separation value of this test showed a value of 1.55 with a strata value of 2.40. This result indicates that the test items on the instrument have two levels of difficulty. In other words, this test can measure students with both high and low abilities.

Next, the fit statistics of each item were checked to see if the data fit the model. Table 1 illustrates the summary statistics of each item. The acceptable value of infit and outfit MnSq is close to 1.00 (Engelhard, Jr. & Wang, 2024). Therefore, the values ranging from 0.5 to 1.5 are considered fit to the model. Meanwhile, for the ZStd, the acceptable range is between -2 to 2. Thus, according to the table below, five items are considered misfit (SEFI1, SIF2, SIF3, SR5, and SR7). Each of these items are written as *"Saya meminta feedback dari dosen mengenai performa belajar saya"* (I ask my teachers to give me feedback about my performance), *"Emosi saya memengaruhi penilaian saya terhadap kinerja belajar"* (My emotions influence my evaluation on my learning performance), "*Kondisi fisik saya mencerminkan seberapa baik saya belajar"* (How my body feels tells me how well I am doing), *"Ketika melakukan*

latihan, saya fokus kepada jawaban saya yang salah untuk membantu menentukan apa yang harus dipelajari selanjutnya" (When I do exercise, I look at what I got wrong or did poorly on to guide me as to what I should learn next), and *"Saya merenungkan kembali kelemahan saya (dalam belajar) ketika berdiskusi mengenai studi dengan teman sekelas"* (I reflect on my weaknesses when I discuss study related issues with my classmates) respectively.

Table 3.3. *Item Fit Statistics*

Item	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd
SEFM1	0.87	-0.3	0.85	-0.4
SEFM2	0.74	-0.8	0.72	-0.9
SEFM3	0.60	-1.6	0.63	-1.3
SEFM4	0.57	-1.5	0.63	-1.3
SEFM5	0.58	-1.5	0.60	-1.5
SEFI1	2.80	4.3	3.25	5.1
SEFI2	0.78	-0.7	0.78	-0.7
SEFI3	1.32	1.0	1.36	1.2
SEFI4	0.96	0.0	0.90	-0.2
SIF1	1.00	0.1	0.94	-0.1
SIF2	1.92	2.5	1.92	2.5
SIF3	2.43	3.5	2.33	3.3
SIF4	0.63	-1.3	0.63	-1.4
SR1	1.06	0.2	0.95	0.0
SR2	0.63	-1.3	0.65	-1.2
SR3	0.91	-0.2	0.76	-0.7
SR4	0.87	-0.3	0.86	-0.4
SR5	0.40	-2.5	0.42	-2.4
SR6	0.75	-0.8	0.70	-1.0
SR7	0.33	-3.1	0.35	-2.9

Furthermore, to see whether participants can understand the rating scale categories used, a rating scale analysis was conducted. First, the value of the average measure should show a monotonous increase as the category increases (Bond & Fox, 2012). Based on the information provided in Table 3.3, the data fulfill this criterion except for rating scale number 1 (Strongly disagree), indicating an underused

category. Next, the fit of the rating scale to the Rasch model was tested. Based on the recommendation by Linacre (1999), as an indication of acceptable model-with-data fit for rating scales, the value of the Outfit mean square statistic should be less than 2.00. As shown in Table 3.4, category 1 (Strongly disagree) indicates a poor fit.

Table 3.4. *Rating scale analysis*

Rating Scale	Usage	Average Measure	Expected Measure	Outfit MNSQ	Andrich Thresholds	Labels
1	2%	-0.12	-1.30	3.0		Strongly disagree
2	5%	-0.67	-0.69	1.1	-2.27	Disagree
3	15%	-0.14	0.06	0.9	-1.37	Slightly disagree
4	29%	0.84	0.86	0.6	-0.18	Slightly agree
5	32%	1.63	1.61	1.0	1.15	Agree
6	17%	2.52	2.45	1.0	2.66	Strongly agree

Finally, the values of the Andrich thresholds should increase sequentially (Engelhard & Wind, 2018), which indicates that the categories can be clearly distinguished by the participants. Based on the analysis, all values of the thresholds increased progressively from -2.27 to 2.66, indicating that the categories functioned as expected.

Overall, the results of the pilot test showed good item discrimination, characterized by item separation that can distinguish between 2 levels of difficulty on the items. On the other hand, the analysis of item fit showed several items that required revision. Finally, the results of the rating scale analysis indicated categories whose use was not optimal. Therefore, revisions were made based on the results of this pilot, including improvements to items SEFI1, SIF2, SIF3, SR5, and SR7, as well as a reduction in the rating scale categories to five categories.

3.6. Ethical Considerations

This study received ethical approval from the Research Ethics Committee, Faculty of Education with reference number 211/Dek.FIP/UIII/UM.02/6/2025, and official permission to conduct the research was received from the English Education Study Program at the target university (see Appendix). Moreover, the researcher received written consent through an email message from the designer of the original instrument to allow the use and adaptation of the Self-assessment Practice Scale (SaPS) in order to facilitate the purposes of the current investigation.

All the participants gave their informed consent, which was obtained through both verbal and in writing formats. The participants were thoroughly informed regarding the intention of the study, the essence of their participation, and their rights, including the right to decline to participate and even withdraw their participation at any time without any consequences.

Names and other identifying information of the participants were not stored and did not appear in any section of the reporting of the data, with the aim of ensuring that their confidentiality is maintained. Instead, each of the participants was assigned different codes, and the reporting in the manuscript was anonymous. All digital and non-digital records were saved safely, and access to them is available only to the researcher. There were no known risks to the participants in this study, and all procedures were planned to meet the ethical requirements of conducting research on human participants.

3.7. Data Analysis Methods

Through the distribution of printed questionnaire, the data were collected from the participants. After collection, the responses were entered into Microsoft Excel, where participants' identities were coded to ensure anonymity. The data were then organized and saved in a .txt format, as it is the requirement of the FACETS computer program for it to read the data. Subsequently, a data script was formulated and processed using the FACETS computer program for analysis.

The data is analyzed using MFRM to examine rater judging behavior, such as the severity, consistency, and bias interaction that occurred in the rating process. The good thing about this model is that it can show the participant's ability, the item difficulty, and rater severity simultaneously in one single continuum (Boone, 2020; Linacre & Wright, 2002). There are several types of the Facets model, divided by the utilization of the rating scale (Linacre & Wright, 2002). The one used in this study is the model that utilizes the same rating scale across raters and items.

In the analysis, first, the researcher checked the reliability index of the data. It was done to ensure that the instrument is measuring what it intends to measure and is reliable. The acceptable reliability value is above 0.65 (Bond & Fox, 2012). The result from the unidimensionality test was also conducted to ensure that the index value falls within the minimum requirement of 20% of raw variance explained by the measures, which indicates a close-enough unidimensionality to support the interpretation of the results from the Rasch model (Wind & Hua, 2022).

In addition, the separation index value was calculated for all three facets: students, raters, and items. The value from this specific analysis tells the estimation of how many groups of respondents emerged based on the ability level of the variable measured (Mohamat et al., 2022). Good separation indicated by more than 2 groups emerged (Linacre, 2002). Meaning that, for the person separation, if there are 3 groups of ability level participants, it can be categorized into low, medium, and high-level ability groups, and the same goes for items and raters.

Furthermore, the fit statistics of raters and items were measured next. This fit statistic is essential to evaluate because it tells the extent to which the data fits the Rasch model specification (Mohamat et al., 2022). Infit statistics are sensitive to unexpected patterns that are close to one's ability level; meanwhile, Outfit statistics are sensitive to outliers or extreme unexpected responses (Engelhard, Jr. & Wang, 2024). In other words, Infit statistics look for unusual responses on items that should match one's ability level, and Outfit statistics look for unusual responses at the two extreme ends (very easy or very difficult items). The value of MnSq = 1 shows that the data is in the most ideal condition for the Rasch model specification. Therefore, the value should fall within the acceptable range of the model, which ranges between 0.5 to 1.5 (Bond & Fox, 2012). Finally, the analysis of bias interaction between rater and ratee, as well as the unexpected responses, was examined to further confirm the results of the overall performance of each facet.

3.8. Validity and Reliability

The summary statistics from the analysis, as shown in Table 4.1, provide a comprehensive overview of the performance and interactions among the three key facets involved in the assessment process: raters, students, and items. Each facet plays a distinct role in shaping the outcomes, and these statistics assist in interpreting the quality and characteristics of the measurement system. Moreover, all three facets are combined into a single construct within a single continuum, making the comparison meaningful and helpful (Wilson, 2005, 2023).

Table 3.5. *Rasch summary statistics*

Statistics	Rater	Student	Item
N	32	31	20
Mean of logit	-0.87	0.00	0.00
Standard deviation (SD)	0.16	0.34	0.43

Standard error (SE)	0.44	0.15	0.12
Separation	2.85	2.24	3.55
Strata	4.13	3.32	5.07
Reliability	0.89	0.83	0.93
Significance (probability) (p)	0.00	0.00	0.00
Observed Exact Agreement (%)		36.9%	
Expected Agreement (%)		32.7%	
Variance explained by Rasch measures (%)		26.69%	

The number of raters involved in the analysis was 32 in total, consisting of 31 students and 1 lecturer. They provided evaluations for 31 students across 20 items. The logit scale, which informs the measures in log-odds units (logit), serves as the common metric for estimating the severity of raters, the ability of students, and the difficulty of items (Engelhard, 2013; Engelhard & Wang, 2024). According to the results, the mean logit of student and item measures is centered at zero logit (0.00) as the starting point of measurement. However, the rater facet had a mean logit below zero (-0.87). This indicates that, on average, the instruments are able to detect the trends in rater judgements, which is that raters tend to be more lenient than the model expected; overall, they tended to give higher ratings than predicted.

Furthermore, the standard deviation (SD) values provide insight into the distribution of the measures within each facet. Items exhibited the greatest variability, with a standard deviation of 0.43 logits. This indicates a relatively wide range of item difficulty levels. Conversely, students demonstrated a moderate standard deviation of 0.34, signifying some variation in their abilities. In contrast, raters displayed the least variability, with a standard deviation of 0.16. This value reflects a more consistent or similar severity level among raters. However, this consistency in ratings was accompanied by a relatively high standard error for the rater facet at 0.44, which signals greater uncertainty in estimating individual rater severity, though this is still acceptable because the value is below 0.5 logit. This uncertainty is likely due to the limited number of ratings from each rater or the uneven distribution of ratings. Students and items had lower standard errors of 0.15 and 0.12, respectively, suggesting more precise estimations of their measures.

The separation values help to determine how well the model can distinguish between different levels of performance or severity. All three facets demonstrate an acceptable to strong separation, with items having the highest separation index at 3.55.

This high value indicates that the model was capable of differentiating around multiple groups of item difficulty. In addition, students and raters also endorsed good separation value at 2.24 and 2.85, respectively, indicating that the model can distinguish between students' abilities and raters' severity meaningfully. These values from separation translate into distinct strata, or statistically distinguishable levels. In more specific, the model identified approximately five distinct levels of item difficulty (5.07), four levels of rater severity (4.13), and over three levels of student ability estimate (3.32).

The reliability index further supports the instruments' quality. For all three facets, the reliability was high, with the highest observed measure for items (0.93), followed by raters (0.89), and students (0.83). These values reflect the consistency with which the model can estimate the relative positions of elements within each facet. In addition, the statistical significance of the variance in measures was confirmed, as all facets had p-values of 0.00, meaning that the observed differences were highly statistically significant, not due to chance.

The inter-rater reliability was determined by the observed exact agreement between raters, and the model percentage, which showed 36.9%. Notably, this observed agreement was slightly higher than the model's expected agreement percentage of 32.7%. This suggests that raters have a good agreement with each other. Finally, the model accounted for 26.69% of the variance in the data through Rasch measures. According to Wind and Hua (2022), a value of about 20% of the variance is evidence of "close enough" unidimensionality. While this may seem modest, the proportion is reasonable in the context of complex human ratings, where factors such as individual interpretation, contextual variability, and subjective judgement often contribute to unexpected variance.

Overall, these summary statistics indicate the good reliability and validity of the instrument, as well as the effective functioning of the MFRM model in accounting for and adjusting the inherent variability among raters, students, and items. The model's estimates were statistically significant and showed high reliability and meaningful separation, suggesting that the data can be interpreted fairly and meaningfully.

CHAPTER IV
FINDINGS AND DISCUSSIONS

This chapter presents the findings derived from the Many-Facets Rasch analysis. The discussion begins with a comprehensive examination of the summary statistics, including measures such as means, reliability, standard deviation, and separation indices, among others. The initial reports of descriptive statistics are intended to establish the fundamental properties of the dataset, thereby ensuring its adequacy for subsequent interpretation. Following this, the analysis of variable maps, fit statistics, bias, and unexpected responses was discussed. In addition, the interpretation of the results is guided by the research questions in the discussion section. Each answer to the research question will be presented through the lens of the Cyclical Self-assessment and Assessment-as-Learning model, respectively. Finally, relevant academic literature enriches the discussion.

4.1. Findings

4.1.1 Summary Statistics

Table 4.1 presents the overview of the statistical analysis of the Many-Facet Rasch Model (MFRM), including three facets of raters, students, and items. The Item facet consists of four dimensions of the self-assessment practice scale (SaPS): SEFM (seeking external feedback through monitoring), SEFI (seeking external feedback through inquiry), SIF (seeking internal feedback), and SR (self-reflection).

Table 4.1. *The statistical overview of the MFRM analysis.*

Statistics	Raters	Student	SEFM	SEFI	SIF	SR
Measure						
Mean	-0.87	0.00	0.00	0.00	0.00	0.00
SD	0.47	0.38	0.45	0.69	0.05	0.47
FAM						
Mean	3.81	3.82	3.78	3.53	3.82	4.10
SD	0.31	0.26	0.25	0.48	0.03	0.17
SE	0.16	0.34	0.13	0.12	0.13	0.16
Outfit MNSQ						
Mean	0.99	1.00	1.03	1.00	1.00	1.03
SD	0.45	0.37	0.17	0.15	0.12	0.15
Infit MNSQ						
Mean	0.99	1.00	1.02	0.99	0.99	1.00

SD	0.45	0.39	0.24	0.19	0.14	0.13
Outfit ZSTD						
Mean	-0.4	-0.2	0.2	0.0	0.0	0.3
SD	2.5	2.2	1.1	1.1	0.9	0.8
Infit ZSTD						
Mean	-0.3	-0.2	0.0	-0.1	-0.1	0.0
SD	2.5	2.3	1.5	1.4	1.0	0.8
Strata	4.13	3.32	4.65	7.66	0.33	4.00
Separation	2.85	2.24	3.24	5.49	0.00	2.75
Reliability	0.89	0.83	0.91	0.97	0.00	0.88
Inter-rater reliability						
Observed Exact Agreement			36.9%			
Expected %			32.7%			

SD = Standard deviation, SE = Standard error, FAM = Fair average measure.

The total number of raters involved in the analysis was 32, comprising of 31 students and 1 lecturer. They provided evaluations for 31 students across 20 items. The measure values of raters -0.87 (0.47) and students 0.00 (0.38) both indicate that, on average, raters tended to be slightly more severe in their scoring compared to the expected mean of zero. The fit statistics of Infit and Outfit MNSQ remain close to 1.00 (0.99 – 1.03), showing that both raters and students conformed well to the Rasch model expectations. Furthermore, the mean value of ZSTD in Infit and Outfit measures is close to zero (ranging from -0.4 to 0.3), which further confirms the absence of substantial misfit. In addition, reliability indices are high for both raters and students, which are 0.89 and 0.83, respectively. With separation and strata indices for both raters and students being 2.85 (4.13) and 2.24 (3.32), respectively, it is suggested that the model was able to reliably distinguish between levels of severity among raters and performance among students.

Furthermore, across four dimensions of the items (SEFM, SEFI, SIF, SR), the mean of Fair Average Measure (FAM) values ranged from 3.53 to 4.10. This value indicates a relatively high endorsement of self-assessment practice among participants, with the SR dimension (M = 4.10, SD = 0.17) demonstrating the strongest presence. The fit statistics of Infit and Outfit MNSQ for all dimensions remain close to 1.00, with standard deviations below 0.25. This suggests that the items within each dimension fit well with the Rasch model expectation and function consistently (Eckes, 2015). Similarly, the Infit and Outfit ZSTD values fall close to 0, further confirming

the validity of the measurement across dimensions (Eckes, 2015; Engelhard & Wang, 2021).

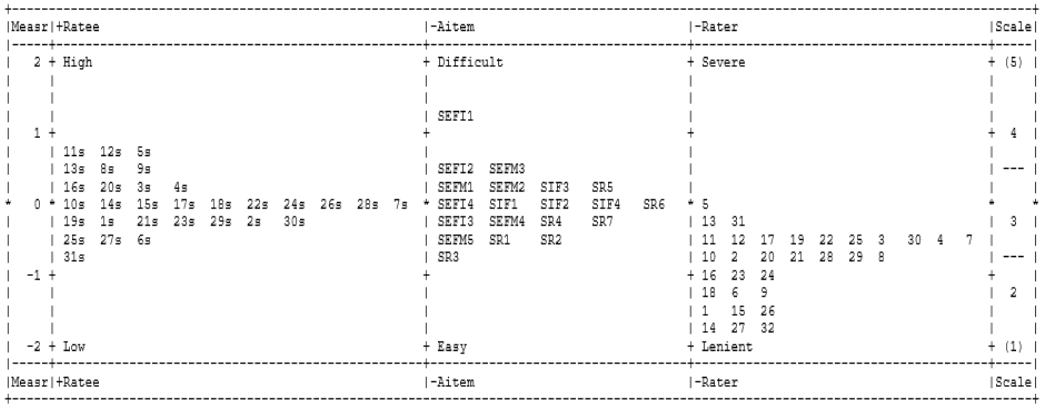
The Separation indices that ranged from 0.00 (SIF) to 5.49 (SEFI), together with their corresponding reliability values from 0.00 in SIF to 0.97 in SEFI, indicate variability in the instrument's ability to distinguish among different levels of students' performance across dimensions. Notably, SEFI demonstrates the highest reliability (0.97), suggesting strong discriminating power, while SIF, on the other hand, shows no effective separation, indicating limited variability among participants in this dimension. This is an understandable and reasonable result, as the items in the SIF dimension prompt participants to reflect on their self-assessment based on internal feedback, which is inherently personal. This helps explain the uniform answers that emerged from participants, why they may struggle to evaluate these items, particularly when the task involves judging others whose internal feedback processes cannot be directly accessed by peers and teachers (Yan, 2022). Moreover, the inter-rater agreement shows 36.9% observed exact agreement and 32.7% expected agreement. The close percentage ranges between these values showed enough level of agreement between the raters, where raters have a common perspective for assessing the quality of assessment.

4.1.2 Variable Map and Person Fit

In the FACETS software result, the variable map (i.e., Wright map) illustrates the logit value of students' ability level, the discrimination of item difficulty, and the severity or leniency tendency of raters, on the same measurement construct. Wright's map provides a comprehensive view of the data distribution, which is very helpful for seeing a snapshot of the overall data and immediate comparison among each rater's assessment tendencies, as well as direct comparison between the facets (Boone, 2020; Eckes, 2015).

In the FACETS program, to obtain a comprehensive picture of the characteristics of each rater group, the analysis was conducted in two stages: a combined analysis and a separate analysis of each facet. The combined analysis contains overall data that includes all raters in one model (self, peer, teacher). On the other hand, the separate analysis utilizes the Self dataset, which includes only data from self-ratings, and the Other-raters dataset, which comprises data from peer and lecturer ratings. This separation was made to observe the characteristics and tendencies of each group of raters in further depth. Figure 4.1 displays the variable map of the combined analysis.

Figure 4.1. *Variable Map (Combine)*

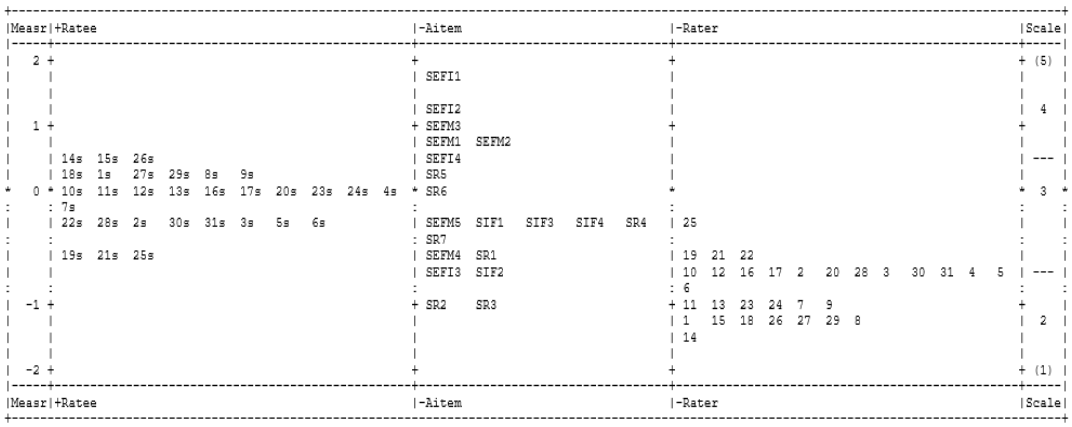


Three facets were calibrated, and the first column of Figure 4.1 illustrates the logit scale that emerged from the data, which ranges from -2 to 2 logits. The logit scale indicated the position of each facet (student, item, rater). The second column shows the location of the ratee (students) on the map, which informs the estimated proficiency level of their self-assessment performance. Each students were rated by 3-4 raters, including the teacher. Students' abilities ranged from those of high-achieving students at the top to those of low-achieving students at the bottom (Linacre, 2025). From the map, the most able students are students' identity of 11s, 12s, and 5s, respectively. Meanwhile, in the lowest position of the ratee distribution is student 31s, who stands as the least able student.

The item column shows the distribution of item difficulty. The higher the item is positioned on the map the more difficult it is, according to participants. Conversely, the lower the item is positioned, the less difficult it is. Item SEFI1 (*Seeking External Feedback through Inquiry* number 1) is the most difficult item to agree with, according to participants. Followed by item SR3 (*Self Reflection* number 3), which stands as the easiest one.

The next column displays the rater location, ranging from the most severe rater to the most lenient rater. At the top (rater 5) is the most severe rater compared to the raters at the lowest end of the map (raters 14, 22, and 32), being the most lenient. As seen from the map, the center position of the raters' severity distribution is around the -1 logit. When the data is in (-) logits, it represents leniency or overestimation in rating. On the other hand, when the data is positioned in (+) logits, it indicates the severity or underestimation in rating (Eckes, 2015; Engelhard & Wang, 2024). Thus, the majority of raters demonstrate generosity in terms of judging the students' self-assessment behavior.

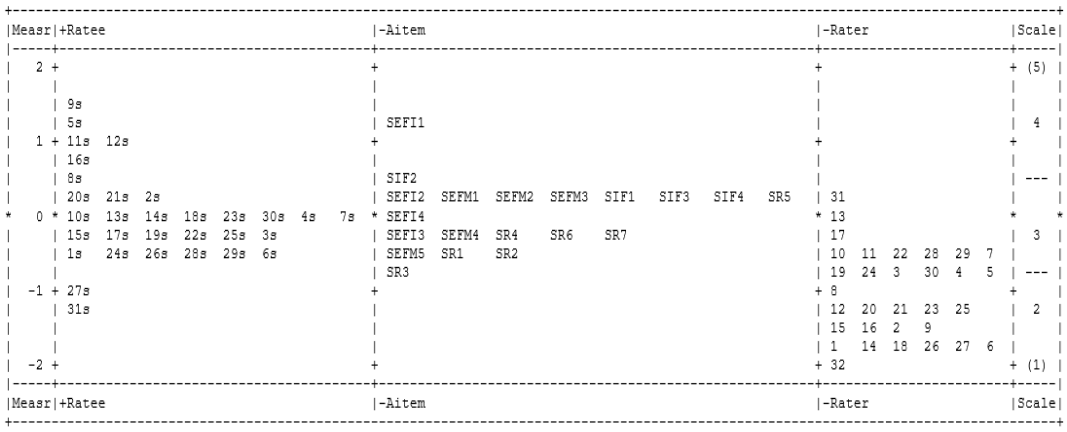
Figure 4.2. *Variable Map (Self)*



Moving on to Figure 4.2, a variable map derived from the Self-raters dataset is presented. The logit scale column reveals that the data ranges from -2 to 2 logits. Next, the ratee column reflects the students' ability levels, with three students ranked at the top (14, 15, 26) and three students at the bottom of the map (19, 21, 25). When compared to the combined data, there is a difference in the ranking position of student ability on self-rating. Meanwhile, in the rater column, rater 25 is identified as the most severe rater, while rater 14 is the most lenient one. This makes sense when rater 25 is harsh toward herself, her position in the ability estimate is the lowest, as she is the most severe rater among the others. The same applies to rater 14. She is very generous in rating herself, which makes her stay at the top in the ability estimate. Both extreme raters in this dataset are female, each from a different province, and have distinct socioeconomic backgrounds. Thus, it can be said that diverse demographic characteristics among raters suggest no specific identity.

Furthermore, for the item column, the map indicates that item SEFI1 (*Seeking External Feedback through Inquiry* number 1) remains the most challenging item to agree with, followed by items SR2 and SR3 (*Self Reflection* number 2 and 3), which are the easiest. Thus, except for student 25, who showed underestimation in her rating, and students 19 and 21 (who showed the average rating on themselves), the rest of the students assessed themselves showed overestimation in their ratings, as indicated on the Wright map.

Figure 4.3. *Variable Map (Others)*



Lastly, Figure 4.3 displays a variable map from students that is being rated by others (peers and a teacher). In this figure, the logit scale is still ranging from -2 to 2 logits, just as in the two previous figures. However, there is a change in students' ability estimates, with student 9 being at the top and student 31 at the bottom. The same goes for the rater severity estimates. At the top, rater 31 is the most severe, while rater 32 is the most lenient one. Since this map displays assessments done by others, there is no self-assessment data in this section. Therefore, student 31, being the most severe rater, has nothing to do with his own ability in the ratee column. Student 31 is very severe in terms of rating his peers. At the same time, he is rated severely by his peers, which places him at the lowest in terms of ability estimates. Meanwhile, rater 32, the most lenient one, is the lecturer. This showed that the lecturer is very lenient in assessing students' self-assessment practice compared to their peers.

Meanwhile, for item difficulty level, SEFI1 (Seeking External Feedback through inquiry number 1) is the most challenging item, followed by SR3 (Self Reflection number 3) as the easiest item. Interestingly, SEFI 1 is consistently staying at the top for the most challenging item across all three datasets. Turns out every rater disagree towards this statement *"saya/ teman ini/ mahasiswa ini meminta feedback dari dosen mengenai performa belajar"* (I ask my teachers to give me feedback about my performance). Similarly, item SR3 also consistently stayed at the bottom as the easiest item to agree with. The item is from SR (*Self Reflection*) dimension that says *"jika ada bagian yang membuat saya/teman ini/mahasiswa ini ragu setelah menyelesaikan tugas, saya akan memeriksanya kembali"* (any areas I am unsure of after finishing my work, I go over again). This might indicate that students are more familiar with internal self-checking as a learning habit rather than asking external feedback, as indicated by the previous item.

Interestingly, among self-raters, item SR2 is also being the easiest item to agree with. The item is *"saya mempertimbangkan apakah komentar dari orang lain (dosen, keluarga, atau teman) tentang tugas yang saya kerjakan masuk akal dan bermanfaat bagi saya* (I think about how much sense the comments of other people (e.g., teachers, family members, friends) regarding my work make to me). This item sounds almost general, and there is a possibility that participants may not understand the item specifically within the self-assessment context. Moreover, considering the Indonesian collectivist culture, where people tend to act upon maintaining social harmony, it is natural that the item is easily agreed upon by participants.

From the analysis of these variable maps, it can be concluded that each group of assessors has varying assessment tendencies. Table 4.2 summarizes the severity/leniency tendencies of raters from the three variable maps. The distribution of the severity level of the Self group is 1.14 logits, ranging from 0.52 (rater 14) to -0.62 (rater 25), In comparison, the distribution of the severity level of the Others group is 2.13 logits, twice as much as the distribution of the Self group, ranging from 0.17 (rater 31) to -1.96 (rater 32). With a mean Logit measure of -1.04, the Others group (peers and teachers) is more generous than the Self group, with a mean Logit measure of 0.00. Meaning that, students are more generous when judging their friends compared to themselves. In addition, the lecturer is most lenient in judging students' performance.

Table 4.2 *Rater severity measure across groups.*

Rater	Measure (Combine)	Measure (Self)	Measure (Peers & teacher)
1	-1.62	0.34	-1.65
2	-0.87	-0.17	-1.38
3	-0.56	-0.14	-0.72
4	-0.56	-0.03	-0.85
5	-0.04	-0.17	-0.67
6	-1.26	-0.14	-1.85
7	-0.62	0.04	-0.41
8	-0.77	0.30	-0.91
9	-1.22	0.21	-1.60
10	-0.86	-0.07	-0.57
11	-0.47	0.08	-0.52
12	-0.49	-0.10	-1.20

13	-0.29	0.12	-0.08
14	-1.75	0.52	-1.80
15	-1.55	0.44	-1.45
16	-0.94	-0.03	-1.56
17	-0.60	-0.07	-0.15
18	-1.28	0.25	-1.69
19	-0.51	-0.40	-0.82
20	-0.75	-0.03	-1.13
21	-0.78	-0.47	-1.37
22	-0.51	-0.37	-0.49
23	-0.97	0.08	-1.31
24	-0.89	0.01	-0.77
25	-1.47	-0.62	-1.14
26	-1.56	0.39	-1.76
27	-1.84	0.30	-1.85
28	-0.66	-0.24	-0.59
29	-0.66	0.30	-0.51
30	-0.48	-0.21	-0.74
31	-0.32	-0.14	0.17
32	-1.69	-	-1.96
Mean	**-0.87**	**0.00**	**-1.04**

Moreover, fit analysis describes how well the data align with the expectations of the model. The Infit and Outfit statistics of MNSQ are usually used to inform about data that does not fit the expectation of the model, either from individuals or from items. However, for reporting purposes, only the Outfit statistic values need to be presented (Boone et al., 2014; J. M. Linacre, 2025). Therefore, using these statistics as a guide, a fit analysis of the model data was conducted.

Table 4.2 presents a summary of the rater misfit, which indicates Outfit MnSq values outside the ideal limits. The first column displays the information of the raters, who identified that seven out of 32 raters indicated that they were unfit. Of the seven identified, five indicated misfit (R31, R29, R3, R30, and R1) and two indicated overfit (R7 and R13). Misfit indicates an unexpected pattern of response resulting from inconsistency in answers. In addition, overfit indicates too predictable answers, commonly caused by less variety in scoring or overly consistent responses in a certain

category (Aryadoust, 2016). The demographics of the participants consisted of four males and three females, representing different cities and having diverse socioeconomic backgrounds. It can be concluded that demographics did not determine the participants' scoring tendencies in this matter.

Table 4.3. *Person Fit Statistics*

Rater	Infit MnSq	Outfit MnSq	PtMea
R7	0.23	0.23	0.40
R13	0.41	0.42	0.38
R31	1.52	1.52	0.09
R29	1.75	1.74	0.30
R3	1.91	1.84	0.44
R30	1.87	1.99	0.45
R1	2.01	1.97	0.43

The findings from misfit and overfit among some raters suggest that not all raters or students understand and respond to the items in a consistent manner (Engelhard & Wind, 2018). This reflects the variations in individual appraisal habits that may be influenced by understanding of the instrument, self-reflection keenness, or appraisal tendencies. By understanding these tendencies, students can determine their next steps more specifically, and institutions can design interventions that suit these needs.

All in all, the trend of the three variable maps analysis indicated that there was a difference in scoring patterns among rater groups, each in the estimates of the ability of the students, the difficulty of the items, and rater severity or leniency. Self-raters were excessive in rating themselves, whereas others (peers and teachers) gave slightly varied judgments, although conservative in their approach. Additionally, fit analysis shows that most of the raters fit the model reasonably well, even though some of the raters fall slightly above and below the acceptable range (Engelhard & Wang, 2024). Overall, the model makes an important contribution by helping to systematically identify individual assessment tendencies.

4.1.3 Item Measurement Report

Table 4.4 displays the difficulty measures, fair average measure (FAM), standard error (SE), and MNSQ infit and outfit indices of the scoring criteria. The difficulty level of the assessment criteria shows a distribution of 2.01 logits, which is 0.43 logits wider than the distribution of student ability. The separation statistic has a value of 3.55 with

a strata of 5.07, or equivalent to 5 groups of item difficulty levels. Then, the standard error value of the difficulty level estimation falls within the range of 0.10 to 0.14, indicating a relatively high level of precision (Aryadoust, 2016). Furthermore, the differences in values between the observed average and the fair average measure for each item are relatively small for the majority of items, indicating stability.

Table 4.4 *Item Measurement Report*

Item	Observed average	FAM	Measure (Logits)	SE	Infit MnSq	Outfit MnSq
SEFM1	3.57	3.57	0.37	0.11	0.86	0.85
SEFM2	3.61	1.61	0.32	0.11	0.73	0.71
SEFM3	3.51	3.51	0.44	0.11	1.16	1.17
SEFM4	4.05	4.05	-0.33	0.13	0.98	0.99
SEFM5	4.09	4.09	-0.39	0.13	1.39	1.31
SEFI1	2.86	2.81	1.19	0.10	0.72	0.76
SEFI2	3.41	3.39	0.57	0.11	1.46	1.47
SEFI3	4.07	4.07	-0.36	0.13	0.82	0.88
SEFI4	3.80	3.79	0.06	0.12	1.22	1.24
SIF1	3.76	3.75	0.12	0.12	1.11	1.16
SIF2	3.78	3.77	0.09	0.12	1.44	1.45
SIF3	3.73	3.73	0.16	0.12	1.56	1.61
SIF4	3.80	3.79	0.06	0.12	0.89	0.92
SR1	4.13	4.13	-0.46	0.13	0.79	0.77
SR2	4.19	4.20	-0.59	0.14	0.91	0.90
SR3	4.31	4.32	-0.82	0.14	0.68	0.71
SR4	3.97	3.97	-0.20	0.12	0.73	0.78
SR5	3.72	3.71	0.17	0.12	0.72	0.74
SR6	3.91	3.91	-0.11	0.12	1.04	0.99
SR7	4.03	4.03	-0.29	0.13	0.64	0.65

Note. FAM = Fair average measure

Meanwhile, for item fit statistics, only one item was detected to have a misfit, which is item SIF3 with an Infit MnSq of 1.56 and an Outfit MnSq of 1.61, exceeding the threshold of 1.5 (Engelhard & Wind, 2018). This item is part of *Seeking Internal Feedback* dimension which written as *"kondisi fisik saya/teman ini mencerminkan seberapa baik dia belajar."* This indicates that the pattern of responses to this item is

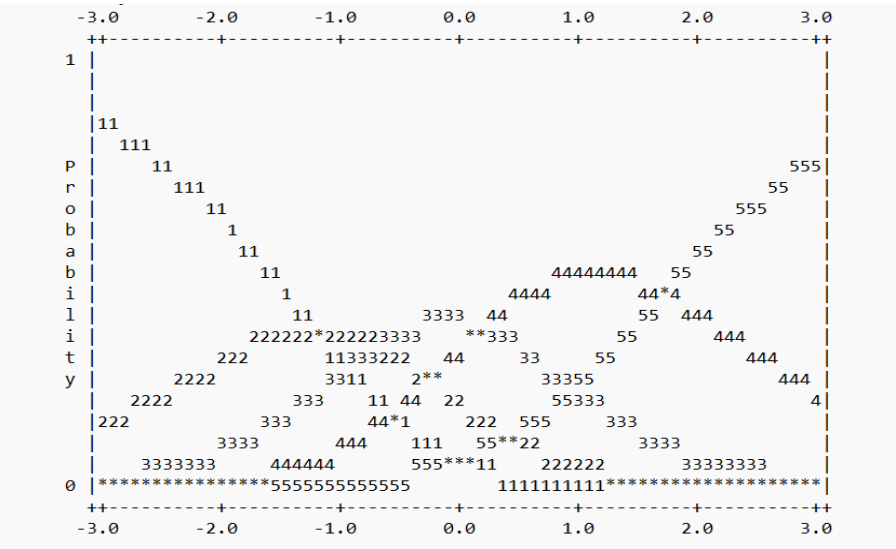
inconsistent and deviates from the model predictions. Given the fact that the SIF dimension is a very personal and internal aspect, when this reflective effort is assessed by others, it is very likely that there will be a discrepancy in perception (Yan, 2022). This item requires the rater to make a judgment regarding a person's internal cognitive process only through external observations. Therefore, the misfit of this item reflects the natural limitation of evaluating internal aspects through external observation.

Table 4.5 displays the statistics related to rating scale categories, which ranged from category 1 (strongly disagree) to category 5 (strongly agree). The average measure for each category is showing a gradual increase from -0.02 to 1.42, indicating a consistent improvement (Bond et al., 2021; B. D. Wright, 1994). For the Outfit MNSQ values, each category has a value close to 1, ranging from 0.9 to 1.1 value. This indicates that the five categories have a good fit validity (Boone et al., 2014; Linacre, 2025). Moreover, the Rasch-Andrich thresholds also displayed a consistent improvement from low to 1.72. This indicates that the raters had a clear understanding of the rating scale categories. In addition, the probability curve in Figure 4.4 showed a slight overlap between category 1 (strongly disagree) and category 2 (disagree). However, it most likely happened due to limited usage of the category.

Table 4.5. *Rating scale category statistics.*

Rating scale category	Total	Percent	Average measure	Expected measure	Outfit MNSQ	Rasch-Andrich Thresholds
1 (strongly disagree)	42	2%	-0.02	-0.12	1.1	low
2 (disagree)	153	7%	0.33	0.25	1.1	-1.22
3 (neutral)	468	23%	0.55	0.62	0.9	-0.68
4 (agree)	875	43%	1.01	1.00	1.0	0.18
5 (Strongly agree)	519	25%	1.42	1.41	1.0	1.72

Figure 4.4. *Rating scale probability curve*



4.1.4 Bias Interaction Analysis Report

To examine the interaction between students and raters under the Many-Facet Rasch Model, a bias analysis was performed. This analysis informs whether particular raters are biased against or toward particular students (Aryadoust, 2015; Eckes, 2015; Engelhard & Wang, 2021). This discrimination is partisan and does not necessarily imply that the overall estimate of the rater is an invalid judgment. Bias can be generally detected by referring to the Bias/Interaction report, where the interaction value should be marked as notable when the *t* value is greater than 2.00 and the probability index is less than 0.05 (Boone, 2020; Boone et al., 2014). Table 4.6 shows only the bias interactions that occur between raters and ratees, determined by the *t* value, probability index, bias size, and outfit MNSQ value (Myford & Wolfe, 2003, 2004).

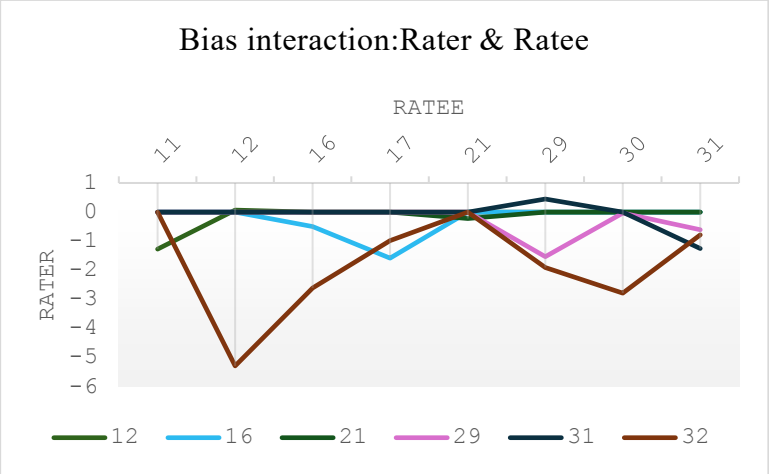
Table 4.6. *Bias Interaction between Rater and Ratee*

Rater	Ratee	t value	P value	Bias Measure	Outfit MnSq	Interpretation
32	31s	-3.86	0.00	-0.91	0.6	severe/underestimation
31	29s	-3.34	0.00	-0.77	1.8	severe/underestimation
29	30s	-2.70	0.01	-0.63	1.3	severe/underestimation
32	17s	-2.59	0.01	-0.71	0.8	severe/underestimation
21	21s	-2.36	0.03	-0.55	1.3	severe/underestimation
12	12s	-2.18	0.04	-0.56	0.6	severe/underestimation

16	17s	2.02	0.05	0.64	0.6	lenient/overestimation
32	12s	2.05	0.05	3.59	0.0	lenient/overestimation
32	16s	2.05	0.05	0.92	1.3	lenient/overestimation
12	11s	2.23	0.03	0.79	1.0	lenient/overestimation
32	30s	2.60	0.01	1.09	0.8	lenient/overestimation
29	29s	2.94	0.00	0.87	1.1	lenient/overestimation
31	31s	3.68	0.00	0.93	1.3	lenient/overestimation

Thirteen interactions demonstrate bias, both positive bias (lenient/overestimation) and negative bias (severe/underestimation). A t-value > 2.00 and p < 0.05 indicates that the interaction is statistically significant. Of the 13 bias interactions, surprisingly, five of them came from rater R32, who is a lecturer. Of the five interactions, it was detected that the lecturer rated severely twice to ratee 31s and 17s, and three times rated leniently to ratee 12s, 16s, and 30s. Particularly for ratee 12s (see Figure 4.5), the bias measure of 3.59 is a very extreme value, which indicates guessing or a fixed rating (always give a 5 rating) given by the rater. This shows that the lecturer might use students' general performance to determine their self-assessment behavior practice instead of direct observation.

Figure 4.5 *Bias interaction graph.*



Furthermore, three student raters exhibited bias twice each (R31, R29, and R12), and two raters showed interaction bias towards one ratee each (R16 and R21). Interestingly, raters R31, R21, R29, and R12 showed self-bias, which means they tend to evaluate themselves harshly or generously. Raters R31 and R29 are male raters, and they demonstrate an overestimation bias in their judgments. Meanwhile, raters R21 and R12, both female, demonstrate underestimation in their self-

judgments. This phenomenon shows the existence of variations in self-assessment patterns that may be influenced by individual or cultural factors (Tandiono & Limijaya, 2025).

The findings of bias indicate that neither lecturers nor students are completely free from subjective tendencies when conducting assessments. Moreover, the presence of self-bias, both in the form of underestimation and overestimation, suggests that self-assessment practices are not always carried out objectively, however, it can be developed over time (Boud et al., 2013). Both of these reflect that the process of self-assessment, when evaluated by others in a higher education environment, can be influenced by many factors, and the goal is to manage these factors so that they can bring good to the enhancement of future learning.

4.1.5 Unexpected Responses

In Rasch analysis, participant responses that deviate from model predictions are termed *unexpected responses*. When there are unexpected patterns of answers, such as when the same rater, ratee, or item appears multiple times on the list, this is an indication of local discrepancies that should be carefully investigated (Eckes, 2015; Engelhard & Wang, 2024). From the analysis, out of a total of 2057 responses, 85 responses (4.1%) were classified as having unexpected responses (Bond & Fox, 2012). Among them, 12 responses (0.5%) were associated with absolute standardized residuals ≥ 3, and 73 responses (3.5%) were associated with absolute standardized residuals ≥ 2 (refer to Appendix 4).

In addition, the majority of the responses (74 cases) fall under the under-value category, which means the rater assigned a lower score than the model expected. Meanwhile, the rest of the response (11 cases) falls under the category of over-value, meaning the rater assigned a higher score than the model expected. This explains the rater's tendency to give a lower score than expected is more dominant than the other way around. However, the detected percentage of the unexpected responses was too small, under 5%. Thus, it indicates that all raters had made a detailed and careful assessment (Mohamat et al., 2022).

Table 4.7 summarizes some of the most frequent responses of the unexpected for the three facets (rater, item, and ratee). This information also adds to the understanding of rater consistency and item quality. First, the rater column shows that there are six raters (R3, R29, R30, R1, R19, and R5) who showed the frequency of unexpected responses more than or equal to five times (≥ 5). On the other hand, raters R32 and R31 demonstrated the most striking unexpected responses, with a frequency of more

than or equal to ten times (≥ 10). This is consistent with the previous findings in the bias analysis, where the lecturer (R32) showed extreme judgment patterns.

Table 4.7. *Summary of Unexpected Response*

Rater		Item		Ratee	
Rater	**Frequency**	**Item**	**Frequency**	**Ratee**	**Frequency**
R3, R29, R30, R19, R1, R5	≥ 5	SIF3, SEFM5, SEFI2, SEFM3, SEFI4, SIF1, SIF2, SEFM4	≥ 5	3s, 30s, 8s, 29s, 5s, 31s, 20s	≥ 5
R32, R31	≥ 10				

Furthermore, eight items (SIF3, SEFM5, SEFI3, SEFI4, SIF1, SIF2, and SEFM4) showed a high frequency of unexpected responses. Five of these items are part of the *Seeking External Feedback* dimension, and the other three are part of the *Seeking Internal Feedback* dimension. This indicates that unexpected responses are more common in indicators that require the action of seeking feedback or reflective processes.

Finally, in the case of ratee, there were seven students who received a number of unexpected responses (3s, 30s, 8s, 29s, 5s, 31s, and 20s). The frequency of unexpected response patterns received by these seven ratees is quite high, about more than or equal to five times (≥ 5). This finding suggests that there are considerable perceptual differences between raters, which might be due to the profile or performance of the ratee.

The emergence of these findings reinforces previous interpretations that in the context of assessing self-assessment practices involving external assessors, there is a variation in responses that cannot be fully explained by the model. The fact that dimensions emphasizing both internal and external feedback seeking elicited more unexpected responses is a sign that it is likely that these dimensions have not yet become a part of study habits by the majority of the students involved. This can be an indication of passive action in learning habits demonstrated by university students within this context.

In general, this section of the results demonstrated that the SaPS instrument and the Many-Facet Rasch Model analysis manage to disclose the intricate nature of self-assessment practice in higher education. The technical results of reliability levels, ability distributions, model fits, and the existence of bias and unanticipated reactions have given a descriptive view of how students, colleagues, and lecturers respond to the assessment activities. Nevertheless, most importantly, these results create an opening for reflection that self-assessment cannot simply be a matter of using valid assessment instruments, but is also a question of personal willingness to reflect sincerely, to master the assessment indicators, and to receive and provide feedback constructively. The remainder of the next section will therefore be devoted to answering the four research questions based on these findings, thereby enhancing the understanding of the practice of self-assessment by different assessors.

4.2. Discussions

RQ1: How does SaPS perform psychometrically and unidimensionally when used by three distinct rater groups to assess students' SA practices?

To address this question, the analysis focused on three major psychometric indicators, which are reliability, separation indices, and unidimensionality. The analysis results based on the Many-Facet Rasch Model (MFRM) demonstrate that the Self-assessment Practice Scale (SaPS) instrument can be described as having good psychometric properties. It functions unidimensionally when used to assess self-assessment practice by the three rater groups. Several important measures are calculated in the foundation of this evaluation, which are reliability, separation for each facet, and unidimensionality tendency of the instrument.

First, regarding reliability, the index values that have been obtained on the three main facets of the assessment system, namely, raters, students, and items, are high. Evidenced by the value of Item reliability reached 0.93, rater reliability reached 0.89, and student reliability reached 0.83. These figures indicate that the instrument is relatively stable for measuring variance between individuals and between items (Eckes, 2015; B. Wright & Stone, 1999). To simplify, the model can distinguish well which students perform higher or lower in terms of self-assessment practice, which raters are more or less strict, and which items are more or less enjoyable.

Secondly, the separation ability is also relatively good in the instrument. The item differentiability values were 3.55, student 2.24, and rater 2.85. These values can be interpreted in terms of the number of strata (statistically distinguishable levels of

difference), which correspond to approximately five levels of item difficulty (5.07 strata), more than three levels of student ability (3.32 strata), and four groups of rater strictness (4.13 strata), respectively. This implies that the instrument not only has a low level of inconsistency but is also sensitive enough to detect true differences between the groups (B. Wright & Stone, 1999).

Third, the Rasch model successfully explained 26.69 percent of the total variance of the data, which is quite good in the context of complex human-based assessments. Wind and Hua (2022) noted that a threshold of 20 percent of the explained variance qualifies as initial evidence of unidimensionality - that is, most items in the instrument were related to the same construct, self-assessment practices. While this value is not perfect unidimensionality, this finding suggests that the SaPS construct is relatively precise and consistent in measuring one key dimension in total. For instance, the variable map shows that the item that presented the greatest challenge for participants to agree with was SEFI1 (Seeking External Feedback through Inquiry 1), while items SR3 and SR2 (Self-Reflection number 3 and 2) were the most agreeable among participants. This indicates a difference in difficulty between the items and is a characteristic of a psychometrically sound instrument. Moreover, rating scale statistics also shows that the categories are easily understood by raters.

Taken together, these results indicate that the instruments produce stable measures across raters, students, and items; also, they function unidimensionally when applied across self, peer, and teacher assessment. Even though the rater measurements are somewhat more uncertain (this is evidenced by the large standard error values), this can be attributed to the fact that there is a limited number of observations per rater (the rater does not assess all students in the class). However, the value 0.44 of SE is still within the acceptable range. Thus, the overall results suggest that the SaPS can be viewed as a good and valid instrument for assessing students' self-assessment practices.

RQ2: How do self, peer, and teacher assessments differ in the way they assess students' self-assessment practice?

Building on evidence that SaPS functions reliably across rater groups, the next step is to examine how assessment differs in their scoring patterns among self, peer, and teacher. The findings of this study reveal that there are diverse scoring patterns between the self-, peer-, and teacher-assessments, respectively. This is reflected in the logit distribution of the combined map variable which shows that although all three rater groups tend to give higher scores than the model predicted (lenient), the logit mean of the others (peer and teacher) rater group is lower (-1.04) than that of the self-

group (-0.89). This explains that within the construct of the logit, external raters are more generous than students themselves in evaluating their self-assessment practices. Furthermore, from the Wright map visualization, it is observed that teachers (R32) often give higher ratings to students more frequently than peers. This is evidenced by the lowest logit value owned by the teacher, which is -1.96 logit. This finding suggests that in the context of practical self-assessment, students actually show a relatively stricter action in examining their own performance.

Based on the Cyclical Self-assessment theory, there are four steps in the self-assessment cycle: determine the criteria of success, seek feedback (both internally and externally), self-reflection, judgment, and calibration (repeat the cycle after the judgment process) (Yan & Brown, 2017). Therefore, the fact that students are severe toward themselves in self-assessment might indicate the high standard of success criteria they set for themselves. This high standard might be caused by too specific criteria or too general criteria; thus, further investigation should be conducted to verify this. Furthermore, the process of actively seeking feedback and self-reflection is a core part of the self-assessment cycle. Therefore, the finding that students were stricter in their self-assessment could be interpreted that students are not yet fully accustomed to making adjustments between internal perceptions and external standards, or the calibration process. In this context, students may face difficulties in understanding their own performance indices. This resonates with the findings of Panadero et al. (2016), who emphasized the importance of metacognitive skills in this process.

In addition, this difference was also seen in the tendency of answers for certain items. For example, item SEFI1 (*Seeking External Feedback through Inquiry*) was consistently awarded as the most difficult item in the eyes of all rater groups, including self, peers, and teachers. However, despite this, the level of agreement from each rater group remained different. For example, teachers tended to agree more with the item than students. In contrast, students and peers tended to disagree more with this item. This means that students' and teachers' perceptions of feedback seeking are likely to differ. Teachers may feel that students have demonstrated this behavior, but students feel otherwise. Another interpretation is that teachers might use students' general performance to assess this item, which does not necessarily reflect a particular behavior. This means that it is important for both students and teachers to have good feedback literacy to achieve alignment of understanding (Dawson et al., 2024; Yang et al., 2025). Thus, this kind of information tells us which area to improve.

The differences in perceptions and interpretations of self-assessment behavior by the three groups of assessors reflect variations in expectations, understanding, and experiences of this practice (Yan et al., 2023). This makes a strong argument that by looking at the practice of self-assessment from multiple perspectives, factors that are thought to influence student motivation and perceptions of this learning strategy can be carefully addressed. If it is felt that students have no experience in this practice, interventions can be designed around their needs. At the same time, if there are indications of misalignment in understanding, then this may indicate that one group needs further intervention to maximize certain practices and understandings. From the findings, it can be concluded that all three types of raters, self, peer, and teacher raters, are equally prone to subjectivity. This reinforces the notion that accuracy in self-assessment should not be understood in terms of replacing teacher assessment (Boud & Falchikov, 2006; Yan, 2022), but rather as a learning strategy that requires structural and pedagogical support from others. Under the Assessment-as-Learning dynamic, productive self-assessment needs metacognitive awareness, an awareness of criteria, and the ability to seek and use feedback constructively (Yan & Carless, 2022).

RQ3: What kind of judging behaviors or rating patterns exist among raters?

To understand the rating behaviors underlying self, peer, and teacher assessments, this study examined the bias interaction, fit statistics, and unexpected responses. These analyses suggest that perceptual differences between rater groups arise not only from differences in judgment levels but also from different patterns of judgment behavior and biases. There is clear evidence that both self-raters and other-raters exhibit various types of inconsistencies and personal tendencies, which indicate the way they make judgments.

For instance, rater 32 (the teacher) showed five significant cases of interaction bias, both in the form of overestimation and underestimation towards some students. Moreover, the teacher shows an extreme leniency towards some students, which indicates a superficial evaluation based on the general performance. This shows that even teachers, who are often considered the gold standard in assessment (Panadero et al., 2016), are also prone to bias. Limited involvement, particularly when the teacher's role is to provide generative feedback is contributes to the misalignment of the criteria understanding. In addition, inaccuracies in teacher assessment can occur due to limited involvement in this process. This finding resonates with the study of Van Zundert et al. (2010), which shows that the quality of student and teacher assessments depends heavily on a shared understanding of the criteria used. Limited involvement,

particularly when the teacher's role is to provide generative feedback, contributes to the misalignment of the criteria understanding between the two.

On the other hand, some students also exhibited self-bias, both in the form of underestimation or overestimation, indicating notable differences in their evaluation. If referred to the self-assessment cycle, the bias that emerged from some students indicates the suboptimal implementation of the overall stages within the cycle. This could be due to the absence of clear assessment criteria from students, a lack of effort in seeking feedback, superficial self-reflection, or a lack of reflective judgment and calibration by the end of the process. Every stage plays a pivotal role in complementing the overall cycle. Thus, further investigation within these areas might be necessary.

For instance, raters 31 and 29 appeared consistently in three categories of indicators: interaction bias, statistical fit, and unexpected responses. Yet, both appear more than five times in the unexpected responses, showing that they rate some ratees unexpectedly and rate some items unexpectedly (see Table 4.5). The fact that both raters perform uniquely towards some items in every dimension (SEFM, SEFI, SIF, SR) shows that there is irregular behavior in the way they seek feedback and use it to reflect (see Appendix 2). From the Assessment-as-Learning perspective, actively seeking feedback is the core of learning (Yan & Carless, 2022). Inconsistencies in this area may reflect gaps in students' feedback literacy, a skill Yan and Charless (2022) identify as pivotal for effective self-assessment.

In addition, student raters also show unexpected behavior when rating their peers as shown by the unexpected rating scores. This suggests that they might have understood feedback seeking action differently, or probably the same as the teacher, they might evaluate their peers primarily on their general performance rather than specific criteria (Myford & Wolfe, 2003). Interestingly, gender also appeared to influence rating patterns, this is evidenced by two male raters tended toward self-overestimation, while two female raters leaned toward underestimation. Although this is not the central finding, this suggests that judgment-making may be shaped by personal and cultural identities (Carless & Boud, 2018; Tandiono & Limijaya, 2025), an important consideration in a cross-cultural education context.

In summary, the findings reveal how the process of judgment is carried out by various parties (self, peer, teacher). Showing that not only do the results of the assessment differ, but the way of judging also contains certain tendencies and patterns. Recognizing and addressing these patterns is essential for students, as it can enhance

the effectiveness of the self-assessment as a learning strategy, while for lecturers, it can improve the fairness and pedagogical value of assessments.

RQ4: In what ways do students' self-assessments reveal opportunities or challenges for using self-assessment as a learning strategy?

An analysis of university students' response patterns provided important insights into the opportunities and challenges of using self-assessment as a learning strategy. Evidence from the variable map, fit statistics, and unexpected responses revealed that there were consistent trends in the way students assessed themselves through certain items. This analysis of students' responses is a reflection of students' understanding, comfort, or even resistance to the learning reflection process.

One of the most notable findings was the position of item SEFI1 *(Saya meminta feedback dari dosen mengenai performa belajar saya/ I ask feedback from the teachers),* which appeared consistently across all three variable maps as the most difficult item to agree with, each by students, peers, and lecturers. This pattern indicates that most students are not yet fully engaged in the practice of seeking feedback from teachers. Compared to the other item in the same dimension, for instance, item SEFI 2 *(Saya meminta saran dari anggota keluarga mengenai tugas kuliah saya/ I ask my family members to give me advice on my work),* which is the second hardest item to agree with, it can be seen that students are genuinely find asking for feedback from teachers is more challenging than having them from family members. This highlights a specific discomfort or hesitation associated with teacher-student feedback interactions, potentially linked to cultural or institutional dynamics.

As noted by Chong (2021), student engagement with feedback is determined by the interaction of contextual and individual factors from ecological standpoints. Contextual factors in the traditional feedback environment include limited access to the teacher, large class size, power imbalance between students and teachers, insufficient details in comments, the lack of chances to resubmit the work, and the heavy workload of the teacher (Carless, 2023; Winstone & Boud, 2022). While individual factors include feedback literacy, learning motivation, prior experiences with feedback, and maturity, these are what influence feedback engagement in students (Shen & Chong, 2023). These findings imply that without addressing these factors, student engagement with feedback remains limited. Thus, these factors need to be addressed in order to promote students' engagement with feedback, which in turn will result in more meaningful self-assessment practice.

Within the Cyclical Self-Assessment framework (Yan & Brown, 2017), seeking feedback is a crucial stage before self-reflection and judgment calibration (see Figure 2.1). The weak agreement with item SEFI1 reflects that the learners have not invested in the process of seeking feedback. Such obstacles may be psychological in nature (embarrassment, fear of disrupting the lecturer), structural (no time or mechanism exists), or epistemic (the student does not see the value of seeking feedback). According to the study conducted by Carless and Winstone (2023), the effectiveness of self-assessment practice hinges on the availability of a learning environment that supports openness, psychological safety, and feedback literacy.

Coversely, item SR3 (*jika ada bagian yang membuat saya/teman ini/mahasiswa ini ragu setelah menyelesaikan tugas, saya akan memeriksanya kembali/ any areas I am unsure of after finishing my work, I go over again)* and item SR2 (*saya mempertimbangkan apakah komentar dari orang lain (dosen, keluarga, atau teman) tentang tugas yang saya kerjakan masuk akal dan bermanfaat bagi saya/ I think about how much sense the comments of other people (e.g., teachers, family members, friends*)), consistently ranked as the easiest item to agree with. Even though these items are intended to address and assess self-assessment practice, there is a high chance that participants interpret the items as in general social judgments. While these responses indicate that students are comfortable with self-reflection, they may also reflect general tendencies of social judgment rather than metacognitive practice.

Considering Indonesian collectivist culture, where maintaining social harmony is heavily valued, agreement with those items may stem from conformity and modesty norms. This finding resonates with Tandiono and Limijaya (2025), who argued that students exhibit leniency towards friends and show modesty towards self-assessment, driven by the need to conform to social expectations and to maintain good relationships. This cultural dimension helps explain why students appear more willing to engage in reflective practices that do not directly challenge hierarchical structures. However, González-Betancor et al. (2019) argued that in the case where self-assessment affects final grade, students often overestimate their contribution to the groupwork by giving themselves higher scores than peers. This illustrates the complexity of aligning reflective habits with authentic evaluation.

Furthermore, the data presented in the results of the unexpected responses supported findings of the earlier item analysis. It was discovered that the majority of the unexpected responses were represented by those items in two primary dimensions: Seeking External Feedback (both SEFI and SEFM), and Seeking Internal Feedback

(SIF). What this implies is that the aspects of self-assessment, which demand high metacognition and reflective awareness, have not been done accordingly by the students. In contrast to dimensions like Self Reflection (SR), it is seen as easier to examine, probably because of how objective and measurable it is according to the students. This follows the findings of Yan et al. (2020) that the effectiveness of SA as a learning strategy largely depends on how much students are engaged in meaningful and repeated assessment processes. These dimensions, therefore, require specific consideration when implementing such learning programs that would like to integrate self-assessment as a significant element in learning strategies.

Taken together, the contrasting patterns between the most difficult and easiest items to agree with reveal significant cultural and pedagogical implications. The reluctance to seek feedback from teachers (SEFI1), despite its central role in the self-assessment cycle, underscores a deeper issue tied to hierarchical norms, limited access, and psychological safety, factors that are particularly relevant in the Indonesian collectivist context. In such a culture, respect for authority and avoidance of confrontation may discourage students from actively requesting feedback, as it can be perceived as overstepping traditional roles between students and teachers. On the other hand, strong agreement with self-reflection items (SR2, SR3) may reflect not just metacognitive engagement, but also conformity and modesty norms commonly observed in collectivist societies. These tendencies suggest that students are more comfortable engaging in reflective behaviors that do not disrupt social harmony or hierarchy. Consequently, to implement self-assessment meaningfully in this context, educational programs must address both cultural values and structural limitations by creating feedback environments that foster openness, reduce power distance, and develop students' feedback literacy and metacognitive awareness.

Finally, reinforcing the idea that self-assessment is not a substitute for lecturer assessment, but a complementary strategy within Assessment-as-Learning (Yan & Boud, 2022) is important. When fully implemented, self-assessment can equip students not only to evaluate themselves more effectively but also to become autonomous, reflective, and responsible lifelong learners (Boud & Soler, 2016; Panadero et al., 2019).

CHAPTER V

CONCLUSION

This chapter provides an overview of the research topic and the results from the research. The discussion in this section includes five parts, namely (1) summary, (2) conclusion, (3) study limitations, (4) suggestions, and (5) future directions.

5.1 Summary

The current study examines the application of self-assessment in the Indonesian higher education sector by utilizing both the lenses of Assessment-as-Learning (AaL) and Cyclical Self-Assessment Theory (Yan & Brown, 2017), using the Self-assessment Practice Scale (SaPS) and Many-Facet Rasch Model (MFRM) for analysis. Through self-rating, peer rating, and lecturer rating, the research presents an empirical report of the psychometric quality of SaPS, the judgment behaviour of various groups of raters, and possibilities of self as a viable and sustainable learning strategy at university settings. The results showed that, although the SaPS displayed adequate levels of reliability and unidimensionality, notable scoring tendencies emerged. It was particularly noted that the students were highly dependent on inner reflection and endorsed less external feedback in its pursuit. These patterns indicate that the aspects of culture and context, i.e., the values of social harmony in Indonesian classrooms, determine the self-assessment practices. As argued by Boud and Soler (2016), assessment design cannot be disjointed from its social context, as the interpretations and behaviors of learners are strongly rooted in the cultural norms.

Unlike the self-assessment research that has been consistently studied within Western contexts, which often emphasizes critical self-judgement and peer dialogue (Andraden & Brown, 2016; Dawson et al., 2024; González-Betancor et al., 2019; Panadero et al., 2019), Indonesian students seem more comfortable with self-introspection and less likely to engage in a more dialogical feedback process. Similarly, in Confucian heritage contexts, studies have shown both a strong acceptance of teacher authority and challenges in fostering student autonomy (Aryadoust, 2015; Yan et al., 2023; Yan & Yang, 2022). Nonetheless, unlike Confucian students who might regard teacher feedback as authoritative, Indonesian students in the study were also reluctant to seek teacher feedback (i.e., SEFI1 emerged as the most challenging question), which may signal problems related to

psychological safety, power distance, or a limited institutional structure for dialogic feedback (Carless, 2023; Winstone & Boud, 2022).

This discrepancy evokes the crucial contextual issue which is that, although the Cyclical Self-Assessment model states that seeking third-party evaluation is a matter of the utmost importance before proceeding with the self-reflection and calibration (Yan & Brown, 2017), the reported low practice of it in Indonesia indicates that this cycle has yet to be fully used in practice. Rather, on the contrary, reflection becomes the stage that the students mostly perform, and it fits into what Panadero et al. (2019) call the common misconceptions of self-assessment as retrospective self-reflection instead of the active feedback-seeking process. Thus, the study showed how the preparedness of Indonesian students in reflective self-evaluation needs to be transformed further to the communicative and feedback-seeking constructs that should underline the regulation of metacognition and feedback literacy (Carless & Winstone, 2023; Yan & Carless, 2022). These findings support the idea of pedagogical interventions that create a safe feedback-rich learning environment and directly scaffold the feedback literacy skills. Filling these gaps, self-assessment will be a long-term learning process that helps students to acquire independence and critical consciousness needed to engage in lifelong learning (Boud, 2000; Boud & Soler, 2016)

5.2 Conclusion

The study underscores the importance of conceptualizing self-assessment not only as a grading tool but as a learning strategy. Findings indicate that students in Indonesian higher education are not accustomed to either requesting feedback actively or proactively, while in addition, lecturer demonstrate a certain tendency to evaluate the students rather globally, without focusing on particular behaviors. This suggests a limited understanding of the basic principles of self-assessment, both from students and lecturers.

This condition shows the importance of feedback literacy by both students and lecturers in supporting the implementation of self-assessment. Students need to be equipped with the understanding and skills to explore, request, and utilize feedback reflectively. On the other hand, lecturers need to support this process by providing formative and descriptive feedback, not just summative grades. The epistemic understanding that self-assessment is a process of building knowledge through feedback and reflection, as well as pedagogical improvements on how to

apply self-assessment in learning practices by lecturers, needs to be strengthened so that lecturers can assist students in this practice consciously and purposefully.

Additionally, it is further indicated in the results that self-assessment and peer-assessment practice in the Indonesian higher education system cannot be dissociated from collectivistic culture. In a culture where relationships between people are important and where social harmony is reinforced, students would grade themselves poorly to show humility, and they would grade their friends generously, as a means of solidarity. Such an attitude is not just a personal bias, but it indicates the dominant social values. Interpersonal relationships, unlike objectivity of judgment, are commonly valued by Indonesian university students. In that regard, outspoken criticism is regarded to be disrespectful, and strong self-evaluation can be noted as arrogant. Therefore, educators should create an assessment approach that takes into account this social process, such as ensuring that self- and peer-assessment is not the only source of final evaluation, but rather a guiding reflection. Through awareness of this cultural dimension, the university can lower the social pressure created by assessment practices whilst enabling the cultivation of metacognitive skills among students.

Finally, the use of MFRM in this study allowed for a more in-depth analysis of the assessment behavior of each group of raters. Surprisingly, students tend to be harsher in their self-assessment compared to the judgments of their peers and lecturers. This finding is consistent with previous studies that explain that cultural factors in Indonesia influence these student assessment tendencies. Altogether, this research highlights the significance of the paradigm shift in assessment, which should be moving towards formative and dialogical.

5.3 Study Limitation

Although this study has been systematically designed and produced some considerable results, several limitations remain to be considered as part of the research interpretation of results and further implications. These limitations relate to the scope of the research context, the distribution of participants, the instrument design, and the analytical approach used. First, the study was restricted to only one class within one of the study programs in a public university in Banten, Indonesia. This affects the representativeness of the data, especially in representing variations in student characteristics in other higher education institutions that have different geographical, social, and cultural backgrounds. Thus, the findings of this study cannot be widely generalized to the entire context of higher education in Indonesia,

particularly in institutions with different learning dynamics from the context where this study was conducted.

Secondly, the ratio of representation amongst the raters group presents another concern. In this study, all student subjects conducted self-assessment and peer assessment, so the group of raters from among students was relatively complete and evenly distributed. However, the involvement of only one lecturer from the teachers' side limits the diversity of pedagogical perspectives. Thus, it may affect the interpretation of the lecturer's assessment data, as it does not reflect the variety of perspectives among lecturers in assessing students' self-assessment practices.

Third, rater training before data collection was not included in this study, which may also be found as one reason explaining the relatively low percentage of raw variance explained by Rasch. Although the Self-assessment Practice Scale (SaPS) was well translated to self, peer, and lecturer assessment, there were no formal calibration sessions, so raters began assessment using the instrument without a common frame of reference regarding the criteria. Such a degree of misalignment could have contributed to a lack of judging consistency between rater group which resulting in a weaker representation of the Rasch analysis. To reduce the risk of idiosyncratic behavior and maximize construct validity in Many-Facet Rasch Model study, a systematic training of raters is necessary. Thus, special attention should be paid in future research to give prior training to the raters.

Fourth, the interpretative component of cultural factor analysis is also limited. This paper shows that students are likely to use internal reflection and inhibit the use of external feedback, which is believed to be connected to the culture of academics in Indonesia. Yet, this finding has not been supported with overt qualitative evidence like interviews or reflective notes that would describe student motivations and perceptions in greater detail. Therefore, the cultural interpretation in this study remains tentative and is to be proved empirically through further exploration. Overall, the limitations do not invalidate the research contribution, but serve as an important basis for directing future studies to be more comprehensive in developing an understanding of the dynamics of self-assessment within the context of the Indonesian higher education environment.

5.4 Suggestions

The findings and limitations discussed in the previous section lead to several suggestions that can be taken into account to enhance the efficacy of the self-assessment implementation as a learning strategy in the environment of Indonesian higher education. First, it is important to explicitly build feedback literacy early on in the learning process. The tendency to avoid seeking external feedback, particularly from the lecturers, is evident in many students participating in this study, which is most likely to be affected by negative attitudes about the meaning of criticism. As such, lecturers and institutions must create learning opportunities that can systematically prepare students with the basic skills of providing, taking, seeking, and using feedback.

Moreover, it is also important to establish a classroom culture that encourages the act of actively seeking feedback. In many cases, the unwillingness of students to seek feedback might be based on their fear of being taken as ignorant, undermining the authority of the lecturer, or being humiliated in the presence of peers. Hence, lecturers should demonstrate that seeking feedback is a learning exercise, rather than a sign of weakness. Lastly, the learning process must include the full self-assessment cycle. In this regard, lecturers must not only request students to evaluate the end product of their task, but also encourage them with criteria setting, process monitoring, external feedback seeking, and reviewing the learning process.

Hence, effective self-assessment does not only concern the degree to which students can self-assess correctly, but also the degree to which they can become responsible learners in a conscious, reflective, and sustainable way. Self-assessment, with the active participation of students, lecturers, and institutions, can be made a practice that is not only academically meaningful but also personally relevant in the creation of lifelong learners.

5.5 Future Direction

The implications of the study findings have opened up various opportunities to expand the knowledge of self-assessment practices in higher education, particularly within the Indonesian context. Therefore, future studies can consider the following direction. First, the next researcher may broaden the research by carrying out a comparative study among the local cultures in different parts of

Indonesia. Being a country that is highly diverse in its socio-cultural domain, learning norms and traditions in each region might vary.

Secondly, a mixed-methods design can be a rather efficient method to investigate the aspects that cannot be touched by quantitative data. The interviews, written reflections, or the use of focus groups may provide the qualitative information needed by the researchers who want to better comprehend how students perceive self-assessment, what emotional or social barriers to this strategy they have to overcome, and how the learning process influences their attitudes towards feedback. Third, longitudinal studies are also possible and can be used to monitor the growth of the self-assessment practices among students over longer intervals of time. Monitoring students over time will allow the researchers to notice trends in how students set criteria, request feedback, and evaluate their progress in the learning process.

Moreover, it may also be valuable to introduce the distal aspects (gender, student demographical characteristics such as socioeconomic status, urban or rural upbringing, and teacher educational level) in future studies. These variables might have subtle yet meaningful interplays with the behaviors of self-assessment in students. For instance, the background of the teacher may have an impact on the nature and quality of feedback. By integrating these aspects into the analysis, one may allegedly give more detailed accounts of the rating behaviors, allowing for to retrieval of the culturally ingrained patterns, and, eventually, resulting in the more enriched and context-sensitive implications.

Altogether, the identified trends in research demonstrate that the practice of self-assessment can potentially be strengthened in the Indonesian higher education environment. Its development, however, should be done hand-in-hand with a thorough knowledge of the cultural, emotional, and social dynamics that guide the way students evaluate themselves. It is anticipated that through the extension of the methodology, context, and time coverage, subsequent studies will stimulate the practice of self-assessment that is more contextualized, inclusive, and transformative to students as lifelong learners.

REFERENCES

Alemdag, E., & Narciss, S. (2025). Promoting formative self-assessment through peer assessment: Peer work quality matters for writing performance and internal feedback generation. *International Journal of Educational Technology in Higher Education*, *22*(1), 22. https://doi.org/10.1186/s41239-025-00522-4

Ali, I. M. (2024). A Guide for Positivist Research Paradigm: From Philosophy to Methodology. *Idealogy Journal*, *9*(2). https://doi.org/10.24191/idealogy.v9i2.596

Alias, M., Masek, A., & Salleh, H. H. M. (2015). Self, Peer and Teacher Assessments in Problem Based Learning: Are They in Agreements? *Procedia - Social and Behavioral Sciences*, *204*, 309–317. https://doi.org/10.1016/j.sbspro.2015.08.157

Andersson, C., & Palm, T. (2017). 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. https://doi.org/10.1016/j.learninstruc.2016.12.006

Andrade, H. L. (2019). A Critical Review of Research on Student Self-Assessment. *Frontiers in Education*, *4*, 87. https://doi.org/10.3389/feduc.2019.00087

Andrade, H. L., & Brookhart, S. M. (2020). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*, *27*(4), 350–372. https://doi.org/10.1080/0969594X.2019.1571992

Andrade, H. L., & Cizek, G. J. (Eds.). (2010). *Handbook of formative assessment*. Routledge.

Andraden, H. L., & Brown, G. T. L. (Eds.). (2016). Student self-assessment in the classroom. In *Handbook of Human and Social Conditions in Assessment* (0 ed., pp. 319–334). Routledge. https://doi.org/10.4324/9781315749136

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory: Measuring in the Educational, Social and Health Sciences*. Springer Nature Singapore. https://doi.org/10.1007/978-981-13-7496-8

Aryadoust, V. (2015). Self- and Peer Assessments of Oral Presentations by First-Year University Students. *Educational Assessment*, *20*(3), 199–225. https://doi.org/10.1080/10627197.2015.1061989

Aryadoust, V. (2016). Gender and Academic Major Bias in Peer Assessment of Oral Presentations. *Language Assessment Quarterly*, *13*(1), 1–24. https://doi.org/10.1080/15434303.2015.1133626

Azizah, N., Suseno, M., & Hayat, B. (2021). Item Analysis of the Rasch Model Items in the Final Semester Exam Indonesian Language Lesson. *World Journal of English Language*, *12*(1), 15. https://doi.org/10.5430/wjel.v12n1p15

Baxter, P., & Norman, G. (2011). Self-assessment or self deception? A lack of association between nursing students' self-assessment and performance: Self-assessment in nursing. *Journal of Advanced Nursing*, *67*(11), 2406–2413. https://doi.org/10.1111/j.1365-2648.2011.05658.x

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, T. G., & Fox, C. M. (2012). *Applying the Rasch model: Fundamental measurement in the human science* (2nd ed). Routledge.

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch model: Fundamental measurement in the human sciences* (Fourth edition). Routledge Taylor & Francis Group.

Boone, W. J. (2020). Rasch Basics for the Novice. In M. S. Khine (Ed.), *Rasch Measurement* (pp. 9–30). Springer Singapore. https://doi.org/10.1007/978-981-15-1800-3_2

Boone, W. J., Staver, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Sciences*. Springer Netherlands. https://doi.org/10.1007/978-94-007-6857-4

Boud, D. (2000). Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, *22*(2), 151–167. https://doi.org/10.1080/713695728

Boud, D., & Falchikov, N. (1989). Quantitative studies of student self-assessment in higher education: A critical analysis of findings. *Higher Education*, *18*(5), 529–549. https://doi.org/10.1007/BF00138746

Boud, D., & Falchikov, N. (2006). Aligning assessment with long-term learning. *Assessment & Evaluation in Higher Education*, *31*(4), 399–413. https://doi.org/10.1080/02602930600679050

Boud, D., & Falchikov, N. (Eds.). (2007). *Rethinking assessment in higher education: Learning for the longer term*. Routledge.

Boud, D., Lawson, R., & Thompson, D. G. (2013). Does student engagement in self-assessment calibrate their judgement over time? *Assessment & Evaluation in Higher Education*, *38*(8), 941–956. https://doi.org/10.1080/02602938.2013.769198

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698–712. https://doi.org/10.1080/02602938.2012.691462

Boud, D., & Soler, R. (2016). Sustainable assessment revisited. *Assessment & Evaluation in Higher Education*, *41*(3), 400–413. https://doi.org/10.1080/02602938.2015.1018133

Brown, G. T. L., Andrade, H. L., & Chen, F. (2015). Accuracy in student self-assessment: Directions and cautions for research. *Assessment in Education: Principles, Policy & Practice*, *22*(4), 444–457. https://doi.org/10.1080/0969594X.2014.996523

Brown, G. T. L., & Harris, L. R. (2013). Student Self-Assessment. In J. McMillan, *SAGE Handbook of Research on Classroom Assessment* (pp. 367–393). SAGE Publications, Inc. https://doi.org/10.4135/9781452218649.n21

Carless, D. (2023). Teacher feedback literacy, feedback regimes and iterative change: Towards enhanced value in feedback processes. *Higher Education Research & Development*, *42*(8), 1890–1904. https://doi.org/10.1080/07294360.2023.2203472

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315–1325. https://doi.org/10.1080/02602938.2018.1463354

Carless, D., & Winstone, N. (2023). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, *28*(1), 150–163. https://doi.org/10.1080/13562517.2020.1782372

Chang, C.-C., Tseng, K.-H., & Lou, S.-J. (2012). A comparative analysis of the consistency and difference among teacher-assessment, student self-assessment and peer-assessment in a Web-based portfolio assessment environment for high school students. *Computers & Education*, *58*(1), 303–320. https://doi.org/10.1016/j.compedu.2011.08.005

Cheah, P. K., Diong, F. W., & Yap, Y. O. (2018). *Peer Assessment in Higher Education: Using Hofstede's Cultural Dimensions to Identify Perspectives of Malaysian Chinese Students*. *26*(3), 1471–1489.

Chong, S. W. (2021). Reconsidering student feedback literacy from an ecological perspective. *Assessment & Evaluation in Higher Education*, *46*(1), 92–104. https://doi.org/10.1080/02602938.2020.1730765

Cleary, T. J. (2006). The development and validation of the self-regulation strategy inventory—Self-report. *Journal of School Psychology*, *44*(4), 307–322.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches* (Fifth edition). SAGE.

Dawson, P., Yan, Z., Lipnevich, A., Tai, J., Boud, D., & Mahoney, P. (2024). Measuring what learners do in feedback: The feedback literacy behaviour scale. *Assessment & Evaluation in Higher Education*, *49*(3), 348–362. https://doi.org/10.1080/02602938.2023.2240983

DeLong, T. J. (2011). *Flying Without a Net: Turn Fear of Change into Fuel for Success*. Harvard Business Review Press.

Earl, L. M. (2013). *Assessment as learning: Using classroom assessment to maximize student learning* (Second edition). Corwin Press.

Eckes, T. (2015). *Introduction to Many-Facet Rasch Measurement: Analyzing and Evaluating Rater-Mediated Assessments- 2 nd Revised and Updated Edition* (2nd, Revised ed eds.). Peter Lang GmbH, Internationaler Verlag der Wissenschaften. https://doi.org/10.3726/978-3-653-04844-5

Ehlers, U.-D. (2013). *Open Learning Cultures: A Guide to Quality, Evaluation, and Assessment for Future Learning*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-38174-4

Ehlers, U.-D. (2020). *Future skills: The future of learning and higher education*. BoD–Books on Demand.

Ehlers, U.-D., & Eigbrecht, L. (2024). *Creating the university of the future: A global view on future skills and future higher education*. Springer Nature.

Engelhard, G., & Wang, J. (2021). *Rasch models for solving measurement problems: Invariant measurement in the social sciences*. SAGE.

Engelhard, G., & Wind, S. A. (2018). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments* (First edition). Routledge.

Engelhard Jr., G. (2013). *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences* (0 ed.). Routledge. https://doi.org/10.4324/9780203073636

Engelhard, Jr., G., & Wang, J. (2024). *Invariant Measurement: Using Rasch Models in the Social, Behavioral, and Health Sciences* (2nd ed.). Routledge. https://doi.org/10.4324/9781003458746

Entwistle, N., McCune, V., & Tait, H. (2013). *Approaches and Study Skills Inventory for Students (ASSIST). Report of the development and use of the inventories (updated March, 2013)*.

Gibbs, G., & Simpson, C. (2004). Conditions Under Which Assessment Supports Students' Learning. *Learning and Teaching in Higher Education*, *1*, 3–31.

González-Betancor, S. M., Bolívar-Cruz, A., & Verano-Tacoronte, D. (2019). Self-assessment accuracy in higher education: The influence of gender and performance of university students. *Active Learning in Higher Education*, *20*(2), 101–114. https://doi.org/10.1177/1469787417735604

Gunawardena, M., Bishop, P., & Aviruppola, K. (2024). Personalized learning: The simple, the complicated, the complex and the chaotic. *Teaching and Teacher Education*, *139*, 104429. https://doi.org/10.1016/j.tate.2023.104429

Hambleton, R. K., & Lee, M. K. (2013). Methods for translating and adapting tests to increase cross-language validity. *The Oxford Handbook of Child Psychological Assessment*, 172–181.

Jimaa, S. (2011). The impact of assessment on students learning. *Procedia - Social and Behavioral Sciences*, *28*, 718–721. https://doi.org/10.1016/j.sbspro.2011.11.133

Johnson, R. B., & Christensen, L. B. (2025). *Educational research: Quantitative, qualitative, and mixed approaches* (Eighth edition). SAGE.

Johnston, L., & Miles, L. (2004). Assessing contributions to group assignments. *Assessment & Evaluation in Higher Education*, *29*(6), 751–768. https://doi.org/10.1080/0260293042000227272

Khoiriyah, U., & Roberts, C. (2025). Investigating the role of self-assessment in enhancing self-regulated learning amongst medical students in problem-based learning. *BMC Medical Education*, *25*(1), 780. https://doi.org/10.1186/s12909-025-07359-5

Khonamri, F., Kralik, R., Viteckova, M., & Petricovicova, L. (2021). Self-Assessment and EFL Literature Students' Oral Reproduction of Short Stories. *European Journal of Contemporary Education*, *10*(1). https://doi.org/10.13187/ejced.2021.1.77

Kostons, D., Van Gog, T., & Paas, F. (2012). Training self-assessment and task-selection skills: A cognitive approach to improving self-regulated learning. *Learning and Instruction*, *22*(2), 121–132. https://doi.org/10.1016/j.learninstruc.2011.08.004

Leach, L. (2012). Optional self-assessment: Some tensions and dilemmas. *Assessment & Evaluation in Higher Education*, *37*(2), 137–147. https://doi.org/10.1080/02602938.2010.515013

Lee, I., Mak, P., & Yuan, R. E. (2019). Assessment as learning in primary writing classrooms: An exploratory study. *Studies in Educational Evaluation*, *62*, 72–81. https://doi.org/10.1016/j.stueduc.2019.04.012

Lew, M. D. N., Alwis, W. A. M., & Schmidt, H. G. (2010). Accuracy of students' self-assessment and their beliefs about its utility. *Assessment & Evaluation in Higher Education*, *35*(2), 135–156. https://doi.org/10.1080/02602930802687737

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. *Language Testing*, *38*(2), 189–218. https://doi.org/10.1177/0265532220932481

Linacre, J. (2002). *A User"s Guide to Winsteps: Rasch-Model Computer Program*.

Linacre, J. M. (1999). Investigating rating scale category utility. *Journal of Outcome Measurement*, *3*(2), 103–122.

Linacre, J. M. (2025). *A user's guide to WINSTEPS ministep: Rasch-model computer programs*. Winsteps.com.

Linacre, J. M., & Wright, B. D. (2002). *Construction of Measures from Many-facet Data*. *3*(4), 486–512.

Linacre, J., & Wright, B. (2002). Understanding Rasch measurement: Construction of measures from many-facet data. *Journal of Applied Measurement*, *3*, 486–512.

Mat Roni, S., Merga, M. K., & Morris, J. E. (2020). *Conducting Quantitative Research in Education*. Springer Singapore. https://doi.org/10.1007/978-981-13-9132-3

Mendoza, N. B., & Yan, Z. (2021). Involved and autonomy-supportive teachers make reflective students: Linking need-supportive teacher practices to student self-assessment practices. In *Assessment as Learning* (pp. 173–189). Routledge.

Mills, A. J., Durepos, G., & Wiebe, E. (Eds.). (2010). *Encyclopedia of case study research*. SAGE Publications.

Mohamat, R., Sumintono, B., & Abd Hamid, H. S. (2022). Raters' Assessment Quality in Measuring Teachers' Competency in Classroom Assessment: Application of Many Facet Rasch Model. *Asian Journal of Assessment in Teaching and Learning*, *12*(2), 71–88. https://doi.org/10.37134/ajatel.vol12.2.7.2022

Mok, M. M. C., Cheong, C. Y., Moore, P. J., & Kennedy, K. J. (2006). The development and validation of the Self-directed Learning Scales (SLS). *Journal of Applied Measurement*, *7*(4), 418–449.

Murchan, D. (2017). *Understanding and applying assessment in education* (1st edition). SAGE Publications.

Myford, C. M., & Wolfe, E. W. (2003). Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M., & Wolfe, E. W. (2004). Detecting and measuring rater effects using many-facet Rasch measurement: Part II. *Journal of Applied Measurement*, *5*(2), 189–227.

Nawas, A. (2020). Grading anxiety with self and peer-assessment: A mixed-method study in an Indonesian EFL context. *Issues in Educational Research*, *30*(1), 224–244.

OECD. (2019). *TALIS 2018 Results (Volume I): Teachers and School Leaders as Lifelong Learners*. OECD. https://doi.org/10.1787/1d0bc92a-en

Panadero, E., Brown, G. T. L., & Strijbos, J.-W. (2016). The Future of Student Self-Assessment: A Review of Known Unknowns and Potential Directions. *Educational Psychology Review*, *28*(4), 803–830. https://doi.org/10.1007/s10648-015-9350-2

Panadero, E., Lipnevich, A., & Broadbent, J. (2019). Turning Self-Assessment into Self-Feedback. In M. Henderson, R. Ajjawi, D. Boud, & E. Molloy (Eds.), *The Impact of Feedback in Higher Education* (pp. 147–163). Springer International Publishing. https://doi.org/10.1007/978-3-030-25112-3_9

Panadero, E., Romero, M., & Strijbos, J.-W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. *Studies in Educational Evaluation*, *39*(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Pastore, S., Yan, Z., & Lao, H. (2025). Teacher instructional practices and strategies for student self-assessment. *Assessment in Education: Principles, Policy & Practice*, *32*(2), 152–172. https://doi.org/10.1080/0969594X.2025.2510204

Pintrich, P. R. (1991). *A manual for the use of the Motivated Strategies for Learning Questionnaire (MSLQ).*

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danish Institute for Educational Research. Expanded edition, University of Chicago Press.

Riznanda, W. A. (2024). Self-Assessment in EFL Writing Class: Students' Practices. *ENGLISH FRANCA: Academic Journal of English Language and Education*, *8*(2 November), 255–268.

Salehi, M., & Masoule, Z. S. (2017). An investigation of the reliability and validity of peer, self-, and teacher assessment. *Southern African Linguistics and Applied Language Studies*, *35*(1), 1–15. https://doi.org/10.2989/16073614.2016.1267577

Schuwirth, L. W. T., & Van Der Vleuten, C. P. M. (2011). Programmatic assessment: From assessment of learning to assessment for learning. *Medical Teacher*, *33*(6), 478–485. https://doi.org/10.3109/0142159X.2011.565828

Shen, R., & Chong, S. W. (2023). Learner engagement with written corrective feedback in ESL and EFL contexts: A qualitative research synthesis using a perception-based framework. *Assessment & Evaluation in Higher Education*, *48*(3), 276–290. https://doi.org/10.1080/02602938.2022.2072468

Sridharan, B., Tai, J., & Boud, D. (2019). Does the use of summative peer assessment in collaborative group work inhibit good judgement? *Higher Education*, *77*(5), 853–870. https://doi.org/10.1007/s10734-018-0305-7

Stiggins, R. (2008). A call for the development of balanced assessment systems. *Assessment Manifesto*.

Sumintono, B. (2015). *Aplikasi Pemodelan Rasch Pada Assessment Pendidikan* (W. Widhiarso, Ed.). Penerbit Trim Komunikata.

Susanti, Y., Nurhajati, D., Kencanawati, D., Riwayatiningsih, R., Sukmayani, F. A., & Aprilia, N. (2023). Self-peer Assessment Method in Writing English Teaching Module Based on Merdeka Curriculum. *Education and Linguistics Knowledge Journal*, *5*(2), 144–161.

Tan, K. H. K. (2012). How Teachers Understand and Use Power in Alternative Assessment. *Education Research International*, *2012*, 1–11. https://doi.org/10.1155/2012/382465

Tandiono, R., & Limijaya, A. (2025). Understanding Rater Bias in Self and Peer Assessment among University Students: A Cultural Psychology Perspective. *The Asia-Pacific Education Researcher*. https://doi.org/10.1007/s40299-024-00967-7

Topping, K. (2003). Self and Peer Assessment in School and University: Reliability, Validity and Utility. In M. Segers, F. Dochy, & E. Cascallar (Eds.), *Optimising New Modes of Assessment: In Search of Qualities and Standards* (Vol. 1, pp. 55–87). Kluwer Academic Publishers. https://doi.org/10.1007/0-306-48125-1_4

Tucker, R. (Ed.). (2017). *Collaboration and Student Engagement in Design Education:* IGI Global. https://doi.org/10.4018/978-1-5225-0726-0

Van Zundert, M., Sluijsmans, D., & Van Merriënboer, J. (2010). Effective peer assessment processes: Research findings and future directions. *Learning and Instruction*, *20*(4), 270–279. https://doi.org/10.1016/j.learninstruc.2009.08.004

Wiliam, D. (2011). What is assessment for learning? *Studies in Educational Evaluation*, *37*(1), 3–14. https://doi.org/10.1016/j.stueduc.2011.03.001

Williams, J. R., & Johnson, M. A. (2000). Self-supervisor agreement: The influence of feedback seeking on the relationship between self and supervisor ratings of performance 1. *Journal of Applied Social Psychology*, *30*(2), 275–292.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Lawrence Erlbaum Associates.

Wilson, M. (2023). *Constructing Measures: An Item Response Modeling Approach* (2nd ed.). Routledge. https://doi.org/10.4324/9781003286929

Wind, S. A., & Hua, C. (2022). *Rasch Measurement Theory Analysis in R* (1st ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781003174660

Winstone, N. E., & Boud, D. (2022). The need to disentangle assessment and feedback in higher education. *Studies in Higher Education*, *47*(3), 656–667. https://doi.org/10.1080/03075079.2020.1779687

Wright, B. D. (1994). Reasonable mean-square fit values. *Rasch Meas Transac*, *8*, 370.

Wright, B., & Stone, M. (1999). *Measurement Essentials 2nd Edition. Wilmington, Delaware: Wide Range*.

Wulandari, M., Sriyati, S., & Purwianingsih, W. (2020). Penerapan peer dan self assessment sebagai tolok ukur penilaian kinerja siswa pada materi sistem koordinasi kelas XI SMA. *Assimilation: Indonesian Journal of Biology Education*, *3*(2), 63–68.

Yan, Z. (2016). The Self-assessment Practices of Hong Kong Secondary Students: Findings with a New Instrument. *Journal of Applied Measurement*, *17*(3), 335–353.

Yan, Z. (2018). The Self-assessment Practice Scale (SaPS) for Students: Development and Psychometric Studies. *The Asia-Pacific Education Researcher*, *27*(2), 123–135. https://doi.org/10.1007/s40299-018-0371-8

Yan, Z. (2022). *Student Self-Assessment as a Process for Learning* (1st ed.). Routledge. https://doi.org/10.4324/9781003162605

Yan, Z., & Boud, D. (2022). Conceptualising assessment-as-learning. In *Assessment as Learning* (pp. 11–24).

Yan, Z., & Brown, G. T. L. (2017). A cyclical self-assessment process: Towards a model of how students engage in self-assessment. *Assessment & Evaluation in Higher Education*, *42*(8), 1247–1262. https://doi.org/10.1080/02602938.2016.1260091

Yan, Z., & Brown, G. T. L. (2021). Assessment for learning in the Hong Kong assessment reform: A case of policy borrowing. *Studies in Educational Evaluation*, *68*, 100985. https://doi.org/10.1016/j.stueduc.2021.100985

Yan, Z., & Carless, D. (2022). Self-assessment is about more than self: The enabling role of feedback literacy. *Assessment & Evaluation in Higher Education*, *47*(7), 1116–1128. https://doi.org/10.1080/02602938.2021.2001431

Yan, Z., Chiu, M. M., & Ko, P. Y. (2020). Effects of self-assessment diaries on academic achievement, self-regulation, and motivation. *Assessment in Education: Principles, Policy & Practice*, *27*(5), 562–583. https://doi.org/10.1080/0969594X.2020.1827221

Yan, Z., Panadero, E., Wang, X., & Zhan, Y. (2023). A Systematic Review on Students' Perceptions of Self-Assessment: Usefulness and Factors Influencing Implementation. *Educational Psychology Review*, *35*(3), 81. https://doi.org/10.1007/s10648-023-09799-1

Yan, Z., & Yang, L. (Eds.). (2022). *Assessment as learning: Maximising opportunities for student learning and achievement*. Routledge, Taylor & Francis Group. https://doi.org/10.4324/9781003052081

Yang, Y., Yan, Z., Zhu, J., Guo, W., Wu, J., & Huang, B. (2025). The development and validation of the Student Self-feedback Behavior Scale. *Frontiers in Psychology*, *15*, 1495684. https://doi.org/10.3389/fpsyg.2024.1495684

Yates, N., Gough, S., & Brazil, V. (2022). Self-assessment: With all its limitations, why are we still measuring and teaching it? Lessons from a scoping review. *Medical Teacher*, *44*(11), 1296–1302. https://doi.org/10.1080/0142159X.2022.2093704

YeşiLçinar, S., & Şata, M. (2021). Examining Rater Biases of Peer Assessors in Different Assessment Environments. *International Journal of Psychology and Educational Studies*, *8*(4), 136–151. https://doi.org/10.52380/ijpes.2021.8.4.622

Zulfikar, T. (2018). The Making of Indonesian Education: An overview on Empowering Indonesian Teachers. *Journal of Indonesian Social Sciences and Humanities*, *2*, 13–39. https://doi.org/10.14203/jissh.v2i0.19

APPENDICES

APPENDIX 1. *Research Instruments*

Original English Version of SaPS

Seeking External Feedback through Monitoring (SEFM)

1. I checked whether I have mastered the course content by doing extra exercises.
2. I checked whether I have fully understood the course content by doing past exam papers.
3. I keep track of my progress by recording my performance.
4. I ask myself questions in my head to check whether I have understood the course content.
5. I check my performance against the answers in the textbook or on a website.

Seeking External Feedback through Inquiry (SEFI)

1. I ask my teachers to give me feedback about my performance.
2. I ask my family members to give me advice on my work.
3. I ask my friends to tell me how to improve my learning.
4. I ask my fellow group members to evaluate my contributions to group work tasks.

Seeking Internal Feedback (SIF)

1. My gut feelings tell me whether my work is good or bad
2. My emotions influence my evaluation on my learning performance.
3. How my body feels tells me how well I am doing.
4. My intuition tells me if I am doing a good job or not.

Self-Reflection (SR)

1. I seek out the reasons for mistakes I made after getting back marked work.
2. I think about how much sense the comments of the other people (e.g., teachers, family members, friends) regarding my work to me.
3. Any areas I am unsure of after finishing my work, I go over again.
4. As I study, I think about whether the way I am studying is really helping me learn.
5. When I do exercise, I look at what I got wrong or did poorly on to guide me as to what I should learn next.
6. I pay attention to my assessment results in order to identify what I can do better next time.
7. I reflect on my weaknesses when I discuss study-related issues with my classmates.

APPENDIX 1. *Research Instruments*

KUISIONER PENELITIAN (Self)

Nama	:	Pekerjaan ayah/ibu	:
Asal kota	:	Pendidikan terakhir ayah/ibu	:
Anak ke-	:		
Jenis kelamin	:		

Petunjuk pengisian:

Jawablah sesuai pengalaman Anda. Tidak ada jawaban benar atau salah. Kuesioner ini dirancang untuk membantu Anda merefleksikan bagaimana Anda biasanya melakukan penilaian terhadap pembelajaran Anda sendiri. Bacalah setiap pernyataan dengan seksama, lalu beri nilai seberapa setuju Anda dengan pernyataan yang diberikan berdasarkan skala berikut:

1 = Sangat Tidak Setuju **3 = Netral** **5 = Sangat Setuju**

2 = Tidak Setuju **4 = Setuju**

No	Items	STS	TS	N	S	SS
1	Saya mengerjakan latihan tambahan untuk mengecek apakah saya telah menguasai materi perkuliahan.					
2	Saya mengerjakan soal-soal dari ujian sebelumnya untuk mengecek apakah saya sudah sepenuhnya memahami materi perkuliahan.					
3	Saya mencatat nilai atau hasil tugas kuliah saya secara rutin untuk memantau perkembangan saya.					
4	Saya bertanya-tanya pada diri sendiri untuk mengecek apakah saya telah memahami materi perkuliahan.					
5	Saya mengecek tugas/pekerjaan saya dengan membandingkannya dengan jawaban dari catatan/ppt/buku/internet.					
6	Saya meminta feedback dari dosen mengenai performa belajar saya.					
7	Saya meminta saran dari anggota keluarga (orangtua/saudara) mengenai tugas kuliah saya.					

8	Saya meminta saran dari teman-teman untuk meningkatkan pembelajaran saya.					
9	(Ketika dalam tugas kelompok) Saya meminta evaluasi dari para anggota tim terhadap kontribusi saya dalam tugas kelompok.					
10	Firasat saya mengatakan/memberitahu apakah tugas yang saya kerjakan bagus atau tidak.					
11	Emosi saya mempengaruhi penilaian saya terhadap performa belajar.					
12	Kondisi fisik saya mencerminkan seberapa baik saya belajar.					
13	Intuisi saya memberitahu apakah saya telah mengerjakan tugas dengan baik atau tidak.					
14	Saya mencari tahu sebab dari kesalahan yang saya buat setelah mendapatkan kembali/menerima hasil koreksi.					
15	Saya mempertimbangkan apakah komentar dari orang lain (dosen, keluarga, atau teman) tentang tugas yang saya kerjakan masuk akal dan bermanfaat bagi saya.					
16	Jika ada bagian yang membuat saya ragu setelah menyelesaikan tugas, saya akan memeriksanya kembali.					
17	Selama saya belajar, saya memikirkan apakah cara belajar ini benar-benar membantu saya belajar.					
18	Ketika saya melakukan latihan, saya fokus kepada jawaban yang salah untuk membantu saya menentukan apa yang harus dipelajari selanjutnya.					
19	Saya memperhatikan hasil penilaian/hasil ujian untuk mengidentifikasi apa yang bisa saya tingkatkan di kesempatan berikutnya.					
20	Saya merenungkan kembali kelemahan saya (dalam belajar) ketika berdiskusi mengenai studi dengan teman sekelas.					

KUISIONER PENELITIAN (Peer)

Nama	:	Nama teman yang di nilai	:
Asal kota	:	Jenis kelamin (teman)	:
Anak ke-	:	Berapa lama mengenal (teman ini)	:
Jenis kelamin	:		

Petunjuk pengisian:

Jawablah sesuai pengamatan Anda. Tidak ada jawaban benar atau salah. Kuesioner ini meminta Anda untuk menilai praktik *penilaian diri* dari salah satu teman sekelas Anda, berdasarkan interaksi atau pengamatan langsung (misalnya saat kerja kelompok, diskusi, atau belajar bersama). Beri nilai seberapa setuju Anda dengan pernyataan yang diberikan berdasarkan skala berikut:

1 = Sangat Tidak Setuju **3 = Netral** **5 = Sangat Setuju**

2 = Tidak Setuju **4 = Setuju**

No	Items	STS	TS	N	S	SS
1	Teman ini mengerjakan latihan tambahan untuk mengecek apakah dia telah menguasai materi perkuliahan.					
2	Teman ini mengerjakan soal-soal dari ujian sebelumnya untuk mengecek apakah dia sudah sepenuhnya memahami materi perkuliahan.					
3	Teman ini mencatat nilai atau hasil tugas kuliahnya secara rutin untuk memantau perkembangan belajarnya.					
4	Teman ini bertanya-tanya pada diri sendiri untuk mengecek apakah dia telah memahami materi perkuliahan.					
5	Teman ini mengecek tugas kuliah dengan cara membandingkannya dengan jawaban dari catatan/ppt/buku/internet.					
6	Teman ini meminta feedback dari dosen mengenai performa belajarnya.					
7	Teman ini meminta saran dari anggota keluarganya (orangtua/saudara) mengenai tugas perkuliah.					

8	Teman ini meminta saran dari saya atau teman-teman lain mengenai cara meningkatkan pembelajaran.					
9	(Ketika dalam tugas kelompok) Teman ini meminta evaluasi/masukan dari anggotanya terhadap kontribusinya dalam tugas kelompok.					
10	Teman ini mengandalkan Firasat untuk memberitahu apakah tugas yang dia kerjakan bagus atau tidak.					
11	Emosi teman ini mempengaruhi penilaiannya terhadap performa belajar.					
12	Kondisi fisik Teman ini mencerminkan seberapa baik dia belajar.					
13	Teman ini mengandalkan Intuisinya untuk memberitahu apakah dia telah mengerjakan tugas dengan baik atau tidak.					
14	Teman ini mencari tahu sebab dari kesalahan yang ia buat setelah mendapatkan kembali/menerima hasil koreksi.					
15	Teman ini mempertimbangkan apakah komentar dari orang lain (dosen, keluarga, atau teman) tentang tugas yang ia kerjakan masuk akal dan bermanfaat bagi dirinya.					
16	Jika ada bagian yang membuatnya ragu setelah menyelesaikan tugas, dia akan memeriksanya kembali.					
17	Selama dia belajar, teman ini memikirkan apakah cara belajarnya benar-benar membantunya belajar.					
18	Ketika dia melakukan latihan, teman ini fokus kepada jawaban yang salah untuk membantunya menentukan apa yang harus dipelajari selanjutnya.					
19	Teman ini memperhatikan hasil penilaian/hasil ujian untuk mengidentifikasi apa yang bisa ia tingkatkan di kesempatan berikutnya.					
20	Teman ini merenungkan kembali kelemahannya (dalam belajar) ketika berdiskusi mengenai studi dengan teman sekelas.					

KUISIONER PENELITIAN (Teacher)

Nama dosen	:	Mahasiswa yang dinilai	:
Lama mengajar	:	Lama mengajar siswa	:
Mengajar di semester	:	Matakuliah yg di ajar	:
Jenis kelamin	:		

Petunjuk pengisian:

Jawablah sesuai pengamatan Anda. Kuesioner ini bertujuan mengumpulkan penilaian Anda sebagai dosen terhadap praktik *penilaian diri* seorang mahasiswa, berdasarkan pengamatan profesional Anda (misalnya saat pembelajaran, tugas yang dikumpulkan, atau saat siswa meminta umpan balik). Beri nilai seberapa setuju Anda dengan pernyataan yang diberikan berdasarkan skala berikut:

1 = Sangat Tidak Setuju **3 = Netral** **5 = Sangat Setuju**

2 = Tidak Setuju **4 = Setuju**

No	Items	STS	TS	N	S	SS
1	Mahasiswa ini mengerjakan latihan tambahan untuk mengecek apakah ia telah menguasai materi perkuliahan.					
2	Mahasiswa ini mengerjakan soal-soal dari ujian sebelumnya untuk mengecek apakah ia sudah sepenuhnya memahami materi perkuliahan.					
3	Mahasiswa ini mencatat nilai atau hasil tugas kuliahnya secara rutin untuk memantau perkembangan belajarnya.					
4	Mahasiswa ini bertanya-tanya pada diri sendiri (merenung) untuk mengecek apakah dia telah memahami materi perkuliahan.					
5	Mahasiswa ini mengecek tugas kuliah dengan cara membandingkannya dengan jawaban dari catatan/ppt/buku/internet.					
6	Mahasiswa ini meminta feedback dari dosen mengenai performa belajarnya.					
7	Mahasiswa ini meminta saran dari anggota keluarganya (orantua/saudara) mengenai tugas perkuliah.					

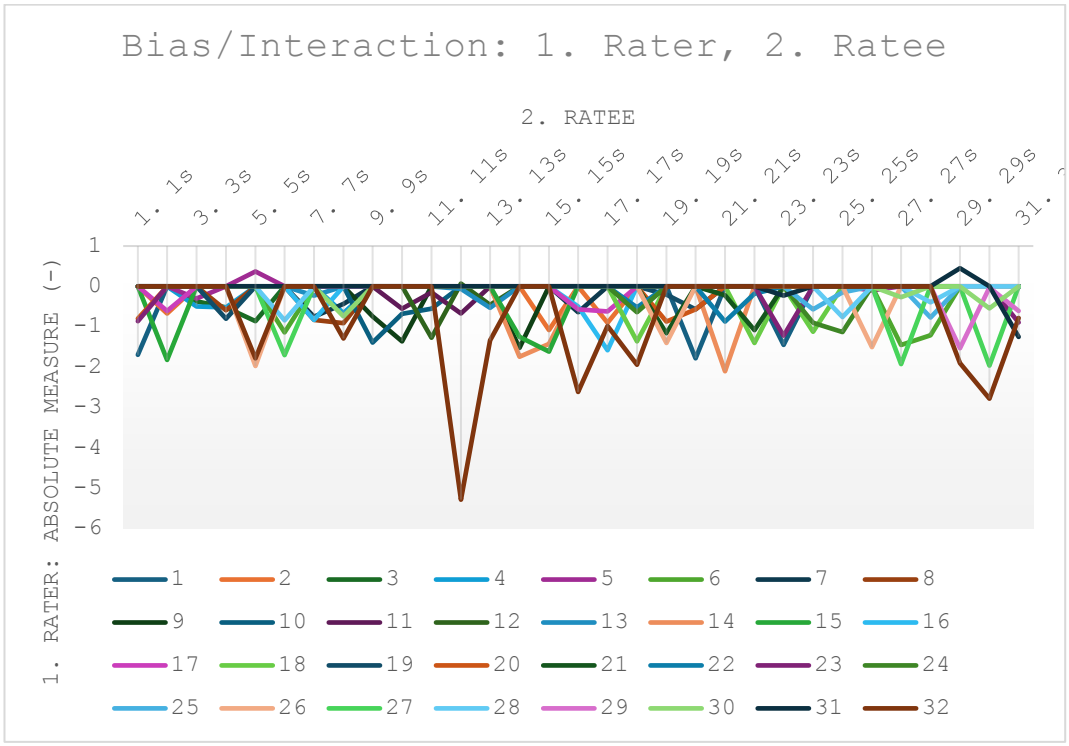
8	Mahasiswa ini meminta saran dari teman-temannya terkait cara meningkatkan pembelajaran.					
9	(Ketika dalam tugas kelompok) Mahasiswa ini meminta evaluasi dari rekan satu timnya terhadap kontribusinya dalam tugas kelompok.					
10	Mahasiswa ini mengandalkan Firasat untuk memberitahu apakah tugas yang dia kerjakan bagus atau tidak.					
11	Emosi mahasiswa ini mempengaruhi penilaiannya terhadap performa belajar.					
12	Kondisi fisik mahasiswa ini mencerminkan seberapa baik dia belajar.					
13	Mahasiswa ini mengandalkan Intuisinya untuk memberitahu apakah dia telah mengerjakan tugas dengan baik atau tidak.					
14	Mahasiswa ini mencari tahu sebab dari kesalahan yang ia buat setelah mendapatkan kembali/menerima hasil koreksi.					
15	Mahasiswa ini mempertimbangkan apakah komentar dari orang lain (dosen, keluarga, atau teman) tentang tugas yang ia kerjakan masuk akal dan bermanfaat bagi dirinya.					
16	Jika ada bagian yang membuatnya ragu setelah menyelesaikan tugas, dia akan memeriksanya kembali.					
17	Selama ia belajar, mahasiswa ini memikirkan apakah cara belajarnya benar-benar membantunya belajar.					
18	Ketika dia melakukan latihan, mahasiswa ini fokus kepada jawaban yang salah untuk membantunya menentukan apa yang harus dipelajari selanjutnya.					
19	Mahasiswa ini memperhatikan hasil penilaian/hasil ujian untuk mengidentifikasi apa yang bisa ia tingkatkan di kesempatan berikutnya.					
20	Mahasiswa ini merenungkan kembali kelemahannya (dalam belajar) ketika berdiskusi mengenai studi dengan teman sekelas.					

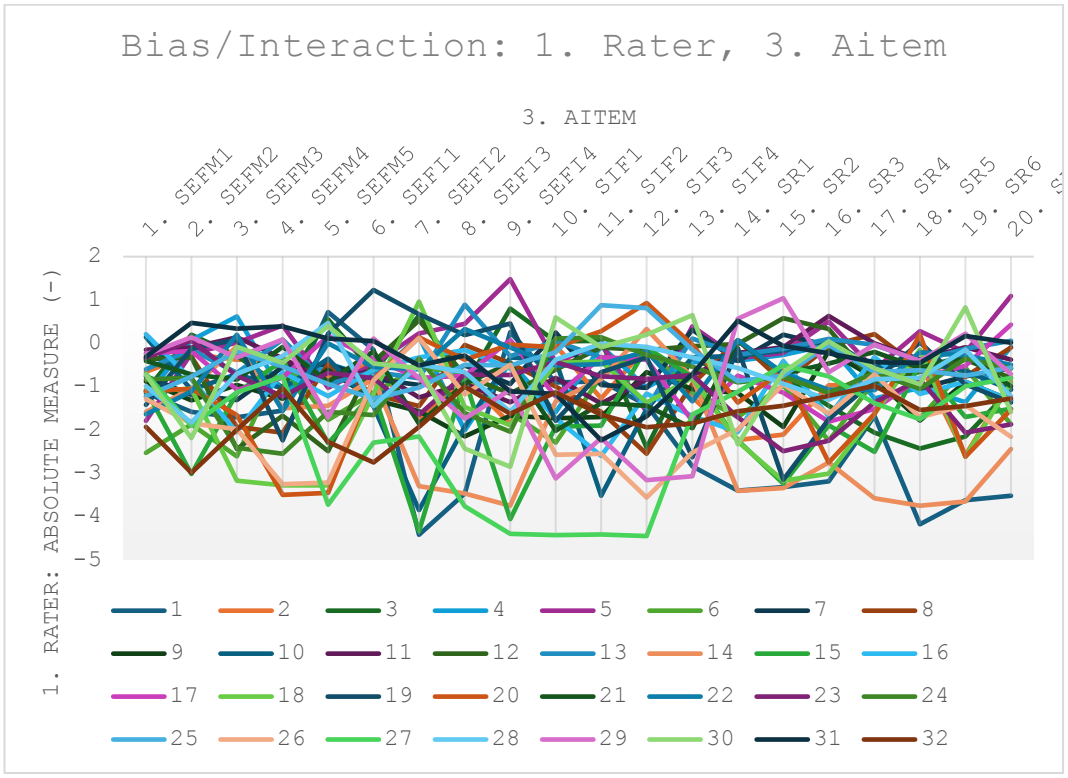
APPENDIX 2. *Unexpected Responses*

Cat	Score	Exp.	Resd	StRes	Nu	Ra	Nu	Rat	Nu	Aitem
1	1	4.0	-3.0	-3.8	3	3	3	3s	5	SEFM5
1	1	4.0	-3.0	-3.8	29	29	30	30s	15	SR2
1	1	3.9	-2.9	-3.6	26	26	26	26s	7	SEFI2
2	2	4.3	-2.3	-3.4	1	1	1	1s	5	SEFM5
1	1	3.8	-2.8	-3.4	30	30	8	8s	13	SIF4
2	2	4.3	-2.3	-3.3	1	1	20	20s	11	SIF2
1	1	3.8	-2.8	-3.3	30	30	8	8s	10	SIF1
1	1	3.8	-2.8	-3.3	30	30	8	8s	11	SIF2
2	2	4.3	-2.3	-3.2	14	14	15	15s	12	SIF3
1	1	3.7	-2.7	-3.2	19	19	19	19s	5	SEFM5
1	1	3.8	-2.8	-3.2	30	30	8	8s	12	SIF3
1	1	3.7	-2.7	-3.1	3	3	3	3s	9	SEFI4
1	1	3.6	-2.6	-3.0	18	18	18	18s	7	SEFI2
2	2	4.2	-2.2	-3.0	32	32	18	18s	11	SIF2
2	2	4.2	-2.2	-2.9	15	15	14	14s	12	SIF3
2	2	4.2	-2.2	-2.9	15	15	15	15s	12	SIF3
1	1	3.6	-2.6	-2.9	31	31	29	29s	5	SEFM5
3	3	4.6	-1.6	-2.9	32	32	8	8s	4	SEFM4
2	2	4.1	-2.1	-2.8	3	3	5	5s	10	SIF1
2	2	4.1	-2.1	-2.8	3	3	5	5s	11	SIF2
2	2	4.1	-2.1	-2.8	3	3	5	5s	13	SIF4
1	1	3.5	-2.5	-2.8	30	30	30	30s	19	SR6
1	1	3.5	-2.5	-2.8	31	31	29	29s	4	SEFM4
2	2	4.1	-2.1	-2.8	32	32	31	31s	4	SEFM4
2	2	4.1	-2.1	-2.7	1	1	23	23s	9	SEFI4
1	1	3.5	-2.5	-2.7	3	3	3	3s	2	SEFM2
3	3	4.6	-1.6	-2.7	27	27	30	30s	15	SR2
2	2	4.0	-2.0	-2.6	5	5	5	5s	20	SR7
1	1	3.4	-2.4	-2.6	19	19	19	19s	9	SEFI4
2	2	4.0	-2.0	-2.6	32	32	31	31s	17	SR4
3	3	4.5	-1.5	-2.5	1	1	20	20s	5	SEFM5
1	1	3.4	-2.4	-2.5	3	3	3	3s	3	SEFM3
1	1	3.4	-2.4	-2.5	30	30	27	27s	5	SEFM5
2	2	3.9	-1.9	-2.4	1	1	1	1s	1	SEFM1
1	1	3.3	-2.3	-2.4	3	3	3	3s	7	SEFI2
2	2	4.0	-2.0	-2.4	9	9	14	14s	12	SIF3
2	2	3.9	-1.9	-2.4	14	14	21	21s	3	SEFM3
2	2	3.9	-1.9	-2.4	17	17	17	17s	5	SEFM5
2	2	4.0	-2.0	-2.4	18	18	24	24s	10	SIF1
2	2	3.9	-1.9	-2.4	19	19	20	20s	8	SEFI3
2	2	3.9	-1.9	-2.4	28	28	28	28s	5	SEFM5
5	5	2.6	2.4	2.4	31	31	31	31s	11	SIF2
5	5	2.6	2.4	2.4	31	31	31	31s	12	SIF3
3	3	4.5	-1.5	-2.4	32	32	13	13s	13	SIF4
3	3	4.5	-1.5	-2.4	32	32	17	17s	8	SEFI3
2	2	3.9	-1.9	-2.3	1	1	1	1s	3	SEFM3
2	2	3.9	-1.9	-2.3	21	21	21	21s	14	SR1
2	2	3.9	-1.9	-2.3	29	29	30	30s	14	SR1
5	5	2.7	2.3	2.3	31	31	31	31s	9	SEFI4
3	3	4.5	-1.5	-2.3	32	32	8	8s	9	SEFI4
3	3	4.4	-1.4	-2.3	32	32	8	8s	10	SIF1
3	3	4.5	-1.5	-2.3	32	32	13	13s	12	SIF3
2	2	3.8	-1.8	-2.2	5	5	5	5s	9	SEFI4

Cat	Score	Exp.	Resd	StRes	Nu	Ra	Nu	Rat	Nu	Aitem
2	2	3.8	-1.8	-2.2	5	5	26	26s	16	SR3
5	5	2.8	2.2	2.2	22	22	21	21s	7	SEFI2
4	4	2.0	2.0	2.2	28	28	6	6s	6	SEFI1
1	1	3.2	-2.2	-2.2	28	28	28	28s	7	SEFI2
1	1	3.2	-2.2	-2.2	29	29	31	31s	19	SR6
1	1	3.1	-2.1	-2.2	30	30	27	27s	19	SR6
3	3	4.4	-1.4	-2.2	32	32	17	17s	17	SR4
2	2	3.8	-1.8	-2.2	32	32	31	31s	10	SIF1
5	5	3.0	2.0	2.1	5	5	3	3s	1	SEFM1
3	3	4.4	-1.4	-2.1	9	9	9	9s	17	SR4
2	2	3.8	-1.8	-2.1	11	11	11	11s	3	SEFM3
2	2	3.8	-1.8	-2.1	16	16	16	16s	1	SEFM1
1	1	3.0	-2.0	-2.1	19	19	19	19s	3	SEFM3
2	2	3.8	-1.8	-2.1	19	19	19	19s	14	SR1
2	2	3.8	-1.8	-2.1	21	21	21	21s	4	SEFM4
1	1	3.1	-2.1	-2.1	29	29	30	30s	7	SEFI2
5	5	2.9	2.1	2.1	29	29	31	31s	10	SIF1
5	5	3.0	2.0	2.1	29	29	31	31s	11	SIF2
5	5	2.9	2.1	2.1	29	29	31	31s	12	SIF3
1	1	3.0	-2.0	-2.1	30	30	30	30s	3	SEFM3
4	4	4.8	-.8	-2.1	32	32	5	5s	16	SR3
3	3	4.4	-1.4	-2.1	32	32	29	29s	4	SEFM4
3	3	4.4	-1.4	-2.1	32	32	29	29s	8	SEFI3
2	2	3.7	-1.7	-2.0	2	2	15	15s	12	SIF3
2	2	3.7	-1.7	-2.0	5	5	5	5s	18	SR5
2	2	3.7	-1.7	-2.0	20	20	20	20s	12	SIF3
2	2	3.7	-1.7	-2.0	20	20	20	20s	18	SR5
1	1	3.0	-2.0	-2.0	22	22	22	22s	7	SEFI2
2	2	3.7	-1.7	-2.0	28	28	28	28s	19	SR6
5	5	3.0	2.0	2.0	29	29	29	29s	7	SEFI2
5	5	3.0	2.0	2.0	29	29	31	31s	13	SIF4
2	2	3.7	-1.7	-2.0	31	31	29	29s	15	SR2

Appendix 3. *Bias Interaction Graphs*





ATTACHMENT 1. *Research Permit (Faculty)*

Universitas
Islam Internasional
Indonesia

Kementerian Agama Republik Indonesia
Universitas Islam Internasional Indonesia
Jalan Raya Bogor KM. 33.5
Cisalak, Sukmajaya, Depok, Jawa Barat 16416
secretariat@uiii.ac.id
www.uiii.ac.id

Ref. No : 211/Dek.FIP/UIII/UM.02/6/2025
Attachment : -
Subject : Request for Research Permit

Depok, June 2, 2025

Dear Sir/Madam,

Assalamu'alaikum Wr. Wb.

We hereby certify the following student:

Name	:	Fitri Amalia
Student ID Number	:	04212310005
Faculty	:	Faculty of Education
Study Program	:	MA in Education

is conducting research for her thesis with the following details:

Thesis Title	:	Self-Assessment in Higher Education: Examining Rater Factors in EFL Classroom through Many-Facet Rasch Model
Research Location	:	Universitas Islam Negeri Syarif Hidyatullah Jakarta
Research Duration	:	May-June, 2025

We respectfully request your assistance in granting research permission to the student. Thank you for your kind attention and cooperation.

Wassalamu'alaikum Wr. Wb.,

Kind Regards,
Dean of the Faculty of Education

Prof. Nina Nurmila, PhD

ATTACHMENT 2. Research permit (University)

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI (UIN)
SYARIF HIDAYATULLAH JAKARTA
FAKULTAS ILMU TARBIYAH DAN KEGURUAN

Jl. Ir. H. Juanda No. 95, Ciputat 15412, Indonesia
Telp. (62-21) 740 1925 Ekt. 1701, (62-21) 744 3328, Fax. (62-21) 744 3328
Website : http://fitk.uinjkt.ac.id, E-mail : fitk@uinjkt.ac.id

Nomor : B- 2143 /F1/HM.01.06/4/2025
Lamp. : -
Hal : Jawaban Izin Penelitian

Jakarta, 28 April 2025

Kepada Yth,

Dekan Fakultas Pendidikan
Universitas Islam International Indonesia

Di Tempat

Assalamu'alaikum Wr.Wb.

Dengan hormat, menindaklanjuti surat Dekan Fakultas Pendidikan Universitas Islam International Indonesia nomor : 086/Dek.FIP/UIII/UM.02/2/2025 dan 095/Dek.FIP/UIII/UM.02/3/2025 perihal permohonan izin penelitian atas nama Muhamad Maulana NIM. 04212310007 dan Fitri Amalia NIM. 04212310005, maka dengan ini kami sampaikan bahwa mahasiswa tersebut diijinkan untuk melakukan penelitian di Fakultas Ilmu Tarbiyah dan Keguruan UIN Syarif Hidayatullah Jakarta.

Demikian surat jawaban ini disampaikan, atas perhatiannya diucapkan terima kasih.

Wassalamu'alaikum Wr.Wb.

Dekan

Prof. Siti Nurul Azkiyah, M.Sc., Ph.D.
NIP. 197605112005012003