Examining Learner’s Evaluative Judgment Supported by Technology-Enabled Feedback Information

Ali Heidari

Follow this and additional works at: https://scholarworks.gsu.edu/ltd_diss

Recommended Citation
doi: https://doi.org/10.57709/36907143

This Dissertation is brought to you for free and open access by the Department of Learning Sciences at ScholarWorks @ Georgia State University. It has been accepted for inclusion in Learning Sciences Dissertations by an authorized administrator of ScholarWorks @ Georgia State University. For more information, please contact scholarworks@gsu.edu.
This dissertation, Examining Learner’s Evaluative Judgment Supported by Technology-Enabled Feedback Information, by Ali Heidari, was prepared under the direction of the candidate’s Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education & Human Development, Georgia State University.

The Dissertation Advisory Committee and the student’s Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.
AUTHOR’S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education & Human Development’s Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

________________________________________

Ali Heidari
NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Ali Heidari
Learning Sciences
College of Education & Human Development
Georgia State University

The director of this dissertation is:

Min Kyu Kim
Department of Learning Sciences
College of Education & Human Development
Georgia State University
Atlanta, GA 30303
CURRICULUM VITAE

Ali Heidari

ADDRESS:

EDUCATION:

<table>
<thead>
<tr>
<th>Degree</th>
<th>Year</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ph.D.</td>
<td>2024</td>
<td>Georgia State University Learning Sciences</td>
</tr>
<tr>
<td>Master’s degree</td>
<td>2018</td>
<td>Georgia State University Applied Linguistics</td>
</tr>
<tr>
<td>Bachelor’s degree</td>
<td>2007</td>
<td>Rajaei University</td>
</tr>
</tbody>
</table>

PROFESSIONAL EXPERIENCE:

<table>
<thead>
<tr>
<th>Year</th>
<th>Position</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>2019-2024</td>
<td>Graduate Research Assistant</td>
<td>Georgia State University</td>
</tr>
<tr>
<td>2019-2024</td>
<td>Graduate Teaching Assistant</td>
<td>Georgia State University</td>
</tr>
</tbody>
</table>

PRESENTATIONS AND PUBLICATIONS:


ABSTRACT

Evaluative judgment is the capacity to discern and assess the quality of work using established criteria (Sadler, 1989), a critical skill for fostering self-regulation and continuous improvement in learning environments (Boud & Falchikov, 2006). This study investigates the effects of self-assessment versus peer assessment and technology versus non-technology settings on evaluation scores, evaluative judgment quality, and rating confidence of undergraduate college students. Utilizing a linear mixed-effects model, the research explores these impacts while accounting for individual participant differences (Gao et al., 2019; Panadero et al., 2016; Shore et al., 1992). The study indicated peer assessments consistently yielded higher evaluation scores across technological and non-technological contexts. However, no significant differences were observed in the quality of evaluative judgment between assessment types or settings, suggesting a more complex interplay of cognitive and affective processes than previously assumed (Sadler, 1998). Unexpectedly, peer assessment was associated with greater rating confidence, challenging the notion that self-assessment, particularly when augmented by technology, would enhance
confidence levels (McCarthy, 2017; Panadero et al., 2016). These results underline the importance of peer interaction and the provision of clear evaluative criteria in enhancing evaluative practices. The study recommends integrating structured peer-assessment activities into educational curricula to promote critical feedback and reflective learning (Falchikov & Goldfinch, 2000; Hanrahan & Isaacs, 2001). The findings contribute to our understanding of assessment practices, emphasizing further research to explore the long-term development of evaluative judgment and the optimal integration of technology in assessment (Ecclestone, 2001; O’Donovan et al., 2004).

INDEX WORDS: evaluative judgment, self-assessment, peer-assessment, technology-enabled feedback, individual differences, reading comprehension, vocabulary knowledge, prior knowledge.
EXAMINING LEARNER’S EVALUATIVE JUDGMENT SUPPORTED BY TECHNOLOGY-ENABLED FEEDBACK INFORMATION

by

ALI HEIDARI

A Dissertation

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

Instructional Technology

in

Learning Sciences

in

the College of Education & Human Development

Georgia State University

Atlanta, GA
2024
DEDICATION

This dissertation is dedicated to my family, my wife, Maryam, and two lovely kids Nikan and Nirvana. Without your love and support, I could not have made it here.
I am profoundly grateful to many people whose support has been indispensable throughout my research. First and foremost, I extend my deepest gratitude to my advisor, Dr. Min Kyu Kim, for his invaluable guidance, patience, and expertise. His mentorship was critical to my success and I am immensely thankful for his support. I would also like to thank my committee members, Dr. Ben Shapiro, Dr. Jennifer Darling Aduana, and Dr. Keith Wright, for their insightful comments and encouragement, which were vital in shaping this project. Special thanks to Dr. Brendan Calandra, Chair of the Department of Learning Sciences, whose support and leadership have enriched my educational journey. On a personal note, my deepest appreciation goes to my family. To my wife, Maryam, and our children, Nikan and Nirvana, thank you for your unwavering love, patience, and sacrifice, which have sustained me through this process. To my parents, Zari and Nowruz, thank you for your lifelong encouragement and belief in my abilities. Your support has been my foundation. Each of you has played a pivotal role in this journey, and I am eternally grateful for your contributions to my life and work.
# Table of Contents

LIST OF TABLES .................................................................................................................. iv

LIST OF FIGURES ............................................................................................................... vi

1 INTRODUCTION .............................................................................................................. 1
   The Necessity of Evaluative Judgment ............................................................................... 1
   Definition of Evaluative Judgment ..................................................................................... 2
   Theoretical, Practical and Methodological Challenges of Evaluative Judgment .......... 5
   Measuring Evaluative Judgment ....................................................................................... 8
   Research Questions ......................................................................................................... 24
   Significance of the Study .................................................................................................. 28
   Assumptions and Limitations ............................................................................................ 29
   Overview of the Study ....................................................................................................... 31

2 REVIEW OF THE LITERATURE ....................................................................................... 33
   Evaluative Judgment and Technology-enabled Feedback Environments ...................... 42
   Evaluative Judgment and Self-assessment ....................................................................... 48
   Evaluative Judgment and Peer-Assessment ..................................................................... 55
   Evaluative Judgment and Individual Differences ............................................................. 61
   Conceptual Framework ...................................................................................................... 66
   Dimensions of Learner’s Characteristics Influencing Evaluative Judgment ................. 71

3 METHODOLOGY ............................................................................................................ 99
   Research Design .............................................................................................................. 99
   Participants ....................................................................................................................... 103
   Reading Passage and Evaluation Task ............................................................................. 106
   Instruments ....................................................................................................................... 114
   Procedures ......................................................................................................................... 124
   Data Analysis ..................................................................................................................... 126

4 RESULTS ......................................................................................................................... 130
   Descriptive Statistics ..................................................................................................... 130
   Evaluation Score .............................................................................................................. 131
   Assumptions of the Linear Mixed-Effects Model .............................................................. 133
LIST OF TABLES

Table 3.1 2x2 Mixed Design

Table 3.2 Demographics of the Study Sample

Table 3.3 Variables of the Study

Table 4.1 Descriptive Statistics

Table 4.2 Variance Inflation Factor Results

Table 4.3 Fixed Effect Estimates for Linear Mixed Effect Model
Table 4.4 Random Effect Variance for Linear Mixed Effect Model

Table 4.5 Variance Inflation Factor Results

Table 4.6 Fixed Effects

Table 4.7 Random Effect

Table 4.8 Likelihood Ratio Test Between CLMMs

Table 4.9 Variance Inflation Factor Results

Table 4.10 Fixed Effects

Table 4.11 Random Effects

Table 4.12 Threshold Coefficients
LIST OF FIGURES

Figure 2.1 Evaluative Judgment Model for Task Internal, Learner and Task External Factors

Figure 2.2 Evaluative Judgment Conceptual Framework

Figure 3.1 Stages of the Study Experiment

Figure 3.2 Prolific Screeners and Eligible Participants

Figure 3.3 Randomization Plans in the Prolific

Figure 3.4 Reading Passage for Summarization Task

Figure 3.5 Summarization Textbox in Qualtrics

Figure 3.6 Summarization Textbox in SMART

Figure 3.7 Feedback in Non-technology Setting

Figure 3.8 Feedback in Technology Setting

Figure 3.9 Evaluation Rubric

Figure 3.10 Gates MacGinitie Reading Test Item

Figure 3.11 Sample Items of Vocabulary Size Test

Figure 3.12 Summary Evaluation Rubric

Figure 3.13 Study Procedure

Figure 4.1 Evaluation Score

Figure 4.2 Residuals versus Fitted Values for Evaluation Score

Figure 4.3 Normality of Residuals

Figure 4.4 ACF for Independence of Residuals

Figure 4.5 Evaluative Judgment Quality

Figure 4.6 Residuals versus Fitted Values for Evaluative Judgment Quality
Figure 4.7 Normality of Residuals for Evaluative Judgment Quality

Figure 4.8 Autocorrelations of Residuals

Figure 4.9 Rating Confidence

Figure 4.10 Homoscedasticity (Rating Confidence)
1 INTRODUCTION
The Necessity of Evaluative Judgment

Evaluative judgment is essential for undergraduate students because it enables them to evaluate the quality of their work and make informed decisions about their educational advancement. Fostering this skill at the undergraduate level is instrumental in nurturing independent learners capable of critiquing and refining their work in alignment with academic standards (Tai et al., 2018). However, evaluative judgment is more than just a skill for academic appraisal; it’s a cornerstone of metacognitive development that is required for preparation of students for lifelong learning. As Nicol and Macfarlane-Dick (2006) argue, developing evaluative judgment enables students to become more self-regulated learners. This self-regulation, in turn, enhances their capacity to manage, direct, and review their own learning, a quality imperative in our rapidly changing knowledge setting. Moreover, evaluative judgment also provides students with the tools to navigate feedback, allowing them to distinguish between feedback that is constructive and relevant and feedback that may not be as beneficial (Carless & Boud, 2018).

In professional and workplace contexts, the capacity for evaluative judgment becomes even more critical. It empowers graduates to assess the quality and relevance of information and solutions in the real-world contexts, ensuring that they don’t merely accept information at face value but engage with it critically (Sadler, 2010). Furthermore, evaluative judgment plays a pivotal role in collaborative environments. According to Panadero et al. (2019), by honing this skill, students are better equipped to offer constructive feedback to peers, fostering collaborative learning and sustain mutual growth. In essence, cultivating evaluative judgment during undergraduate studies not only advances academic growth but also lays the foundation for professional and personal development, essential for the 21st-century world.
Furthermore, the demands of the contemporary world, characterized by an information deluge and an ever-shifting professional landscape, underscore the necessity of evaluative judgment. Bearman et al. (2017) highlight that, in an era where vast amounts of information are available at one’s fingertips, the ability to critically evaluate and judge the quality of information becomes indispensable. It aids undergraduates in discerning reliable sources from misinformation, a vital skill in the age of digital media.

Moreover, as Dweck (2006) postulates, students equipped with strong evaluative skills are more likely to adopt a growth mindset, a belief that abilities and intelligence can be developed through dedication and hard work. This mindset, in conjunction with evaluative judgment, makes students more receptive to challenges, perseverant in the face of setbacks, and eager to learn, attributes that are invaluable in both professional and personal spheres. Therefore, evaluative judgment enables undergraduate learners to self-assess their competencies, recognize areas for improvement, and actively seek learning opportunities. By embedding evaluative judgment into the undergraduate curriculum, institutions ensure that they are not just producing graduates with knowledge, but individuals equipped with the analytical and reflective capabilities vital for navigating and thriving in a complex world (Torrance, 2012).

**Definition of Evaluative Judgment**

The construct of evaluative judgment encapsulates a framework that subsumes how learners perceive, understand, and critically assess the quality of their work and that of their peers (Clore & Huntsinger, 2007; Winne, 2011; Panadero et al., 2019). Clore & Huntsinger (2007) define evaluative judgment as a complex matrix interweaving cognitive, emotional, metacognitive, and social dimensions, emphasizing the holistic and integrated approach required for learners to critically assess work quality. Similarly, Winne (2011) conceptualizes evaluative
judgment as an iterative assessment process, underpinning it as an ongoing, introspective dialogue where learners continually benchmark their performance against set standards. This perspective resonates with Butler & Winne (1995), who underscore the recurrent nature of this judgment, suggesting that learners are in constant comparison with predetermined benchmarks throughout their learning process. Similarly, Panadero et al. (2019) emphasize the dynamic interplay between the specifics of the task and the intrinsic qualities of the learner, positioning evaluative judgment at the intersection of these factors. Lastly, Paris & Paris (2001) pivot toward the emotional and motivational domain, postulating that these dispositions significantly shape the quality and depth of cognitive engagement in evaluative processes.

Given the definitions of evaluative judgment in literature, the construct of evaluative judgment is posited within a matrix that interlaces cognitive, emotional, metacognitive, and social domains (Panadero et al., 2019). This multidimensional perspective underscores the depth of the construct, moving beyond a mere cognitive understanding to incorporate emotional and social dimensions, thereby emphasizing the interplay of affective, cognitive, and metacognitive facets in evaluation (Paris & Paris, 2001). As the backbone of evaluative judgment, the self-regulated learning (SRL) presents models that further elucidate the significance and mechanics of evaluative judgment. Within this domain, Winne and Hadwin’s (1998) model offers a robust framework, emphasizing the iterative processes learners undergo while engaging their evaluative judgment. Within the model, evaluative judgment emerges as a continuous and integral part of the learning process, urging learners to consistently juxtapose their evolving performance against predetermined benchmarks and standards (Butler & Winne, 1995).

An exploration into the evaluative judgment construct would be incomplete without discussing the intricate interplay of learner and task factors. Panadero et al. (2019) accentuate this
interplay, underlining both the intrinsic attributes of learners and the extrinsic conditions presented by the task. From the learner’s vantage, factors such as prior knowledge, cognitive abilities, emotional states, and metacognitive awareness emerge as essential (Panadero et al., 2019). These intrinsic elements meld with task-centric factors, including clarity of guidelines, performance standards, expectations, and availability of resources and technological support (Winne, 2011) shaping learners’ evaluative judgment performance. The confluence of these factors shapes the construct of evaluative judgment, rendering it a dynamic and intricate construct fundamental to learner academic and professional development. Given the multifaceted and intricate nature of the construct of evaluative judgment, it cannot be encapsulated within the confines of a single research study. Recognizing this inherent challenge, our current research narrows its focus primarily on the cognitive and metacognitive dimensions of evaluative judgment. Specifically, we aim to delve into three pivotal components: learner’s evaluative judgment, which provides a quantifiable measure of this ability; the quality of evaluative judgment, an exploration of the accuracy of judgments in comparison with expert judgments; and finally, learners’ confidence in rating, shedding light on the individual’s perceived competence in their evaluations. By concentrating our efforts on these specific facets, we aspire to contribute valuable insights to the broader research on evaluative judgment, while acknowledging the vast scope that remains for future investigations. The vast scope of the Evaluative judgment has led to sporadic research endeavors, an inconsistency attributed to a myriad of challenges related to theorization of evaluation judgment, the way it is translated into practice and challenges in research methodology on how to best investigate evaluative judgment. We now turn to a brief review of the challenges.
Theoretical, Practical and Methodological Challenges of Evaluative Judgment

Evaluative judgment, undeniably fundamental to educational interventions and undergraduate learning experiences, is characterized by diverse interpretations in academic research. Researchers universally acknowledge its role, yet there’s a striking lack of consensus regarding its definition, characteristics, and its scope. Some scholars depict evaluative judgment in the light of students’ ability to measure their work against set standards, while others underscore its metacognitive facets, emphasizing the value of reflection and self-awareness in the learning process (Boud et al., 2018; Boud & Soler, 2016). The work of Panadero and Broadbent (2017) further extends this debate, suggesting that evaluative judgment intertwines with self-regulated learning, thereby corroborating its multifaceted nature. The ambiguity surrounding its definition renders difficulties in juxtaposing findings from various studies, potentially leading to fragmented understandings and interpretations.

Such theoretical discrepancies resonate in practice too. For educators, the imprecise nature of evaluative judgment raises pedagogical dilemmas, specifically in curriculum design and instructional strategies, as they grapple with the challenge of effectively cultivating this competence in learners (Nicol & Macfarlane-Dick, 2006). Moreover, without a universally accepted framework, the risk of superficial or misaligned teaching practices compromises the quality of student outcomes (Tai et al., 2018; Carless & Boud, 2018). The pedagogical and practical dimensions of evaluative judgment, especially concerning its development, indicate significant gaps. A pronounced concern is the divide between theoretical perspectives and their tangible translation into classroom practices (Bearman et al., 2016). While the conceptual intricacies of evaluative judgment have gained academic attention, there’s a paucity of methodical strategies that educators can adopt to foster these skills in students effectively. Parallel to this is the domain of
feedback provision and how it relates to the development of evaluative judgment. Crafting feedback that nurtures evaluative judgment is a complex endeavor, and yet, there is limited scholarly exploration on best practices or methodological insights for educators (Carless & Boud, 2018). Furthermore, the role of technology, particularly in online education and technology-enabled teaching and learning systems, and their potential to bolster evaluative judgment remains an evolving area of research, with opportunities yet to be fully realized (Nicol, 2013).

The exploration of evaluative judgment is also complicated due to multiple research methodology issues. One of the primary challenges revolves around tools to measure the construct of evaluative judgment. Existing scales, in their nascent stages, often rely on self-reported data, a method that introduces potential biases and inaccuracies (Tai et al., 2018). While self-reports provide valuable insights into a student’s perspective, they often can’t capture the nuances of evaluative judgment, and there’s a growing call for more objective, standardized measures (Nicol, 2013). The nature of the studies also poses questions about the depth of our understanding. A large proportion of the current research is cross-sectional, providing only a snapshot of the phenomenon at a specific point in time. This approach contrasts with longitudinal studies, which track changes over time and offer richer insights into development and progression of evaluative judgment. Unfortunately, the latter is less common, leaving gaps in our understanding of the developmental path of evaluative judgment (Boud & Soler, 2016).

Further complicating the prospects of the evaluative judgment is the context-specific nature of much of the research. A study conducted in one educational environment or cultural backdrop might not yield results that are applicable elsewhere. While certain aspects of evaluative judgment might be universally recognized, its manifestations can vary markedly based on artistic and institutional differences (Ajjawi et al., 2017; Bearman et al., 2016). Lastly, the realm of
intervention studies remains underexplored. While the importance of enhancing evaluative judgment is widely acknowledged, a paucity of experimental research rigorously tests methods and strategies for doing so. Such studies are pivotal, as they would not only validate the efficacy of various interventions on best practices (Dawson et al., 2019; Carless & Boud, 2018). Incorporating this multitude of perspectives and addressing these methodological gaps is imperative for a more holistic and actionable understanding of evaluative judgment.

To summarize, the concept of evaluative judgment remains ambiguous, as scholars have yet to reach a consensus on its precise definition, encompassing characteristics, and the underlying cognitive processes that drive it (Nicol, Thomson, & Breslin, 2014). However, the development of evaluative judgment holds paramount significance in the education of undergraduate college students, particularly in core skills such as reading and writing, as they serve as a critical determinant of learners’ overall academic success (Paul & Elder, 2006; Kirkwood & Price, 2013). The current study aims to delve into the processes of how learners utilize their evaluative judgment in assessing both their own and their peers’ summaries as proxy for reading comprehension. It investigates the differences, if any, in the application of evaluative judgment in self-assessment compared to peer-assessment. Additionally, the study explores how individual differences, such as prior knowledge, reading comprehension skills, and vocabulary knowledge, may moderate the evaluative process. A thorough understanding of these dynamics is fundamental for optimizing educational strategies and enhancing overall academic success (Brown, Day, & Jones, 1983; Paul & Elder, 2006; Kirkwood & Price, 2013). Since the focus of the present study is on examining evaluative judgment, a discussion of the measurement of the construct is necessary.
Measuring Evaluative Judgment

Several strategies stand out as instrumental in distilling the multifaceted nature of evaluative judgment. Some of the methods frequently employed are rubrics, reflective journals, portfolios, feedback analysis, scenario-based tasks, and technology-aided assessment. Panadero and Romero (2014) emphasized the value of rubrics, highlighting that they provide transparent performance metrics which guide students in honing their evaluative capabilities. By providing clear and structured criteria for evaluation, rubrics guide learners in understanding the components of quality performance (Andrade & Du, 2005). They offer a framework for self-assessment, allowing students to reflect on their own work and identify areas for improvement (Reddy & Andrade, 2010). Rubrics also encourage metacognition, as learners compare their judgments against predefined criteria and adjust their thinking accordingly (Moskal & Jonsson, 2007). As educators use rubrics to evaluate students’ evaluative judgment, they gain insights into the strengths and weaknesses of each individual’s approach, enabling targeted feedback and tailored interventions (Peterson, 2018). This is further accentuated by the utilization of rubrics in self and peer assessments, techniques that resonate with the research findings of Black and Wiliam (1998), which bolster the significance of reflection and comparison in evaluative judgment.

In addition to the measures, the more longitudinal measures such as reflective journals and portfolios have gained traction in educational research on evaluative judgment. Portfolio assessments, informed by the insights of Barrett (2007), serve as repositories of learners’ longitudinal evaluative developments. According to Moon (2006), these resources serve as chroniclers of students’ cognitive and emotional pathways, offering insights into their evaluative processes over time. Parallel to this, feedback analysis, underscored by Nicol and Macfarlane-Dick (2006), holds prominence, focusing on how learners discern and integrate feedback to refine their work
and judgments. To further enrich the evaluative expertise of learners, scenario-based tasks are utilized, offering versatile platforms for students to exercise their evaluative competence. Here, the work of Sadler (2009) accentuates the merits of context-driven evaluations, highlighting the value they add in simulating real-world decision-making scenarios.

Technology-aided assessments are also proposed as medium for measurement of learners’ evaluative judgment. These tools, by capturing real-time interactions and changes in student work, reflect the sentiments of Boud et al. (2018), emphasizing the iterative but consistent nature of evaluations in digital platforms. This immediacy and dynamism allow students to be part of a feedback loop, where they can actively engage, reflect upon, and recalibrate their work in response to assessments (Boud et al., 2018). As opposed to traditional assessment methods, which often see students as passive recipients of feedback, technology-enabled platforms can foster a more active engagement, facilitating a deeper understanding and internalization of feedback.

Overall, given the aforementioned merits of rubrics in development of learner’s metacognitive capabilities, their role in clarification of standards and criteria of performance and their use in self-assessment and peer-assessment activities, the current study uses rubrics as a measurement tool for capturing learners’ evaluative judgment. In addition, the use of rubrics is justified since it ensures, in comparison with open-ended self-report measures, that evaluation is consistent leading to more objective and equitable assessments (Jonsson & Svingby, 2007). Before going to the details of the study, I will briefly discuss the context of the present study in terms of the need for focusing on evaluative judgment of reading comprehension for undergraduate college students, why the focus on self-assessment and peer-assessment as desired activities for development of evaluative judgment and rationale for inclusion of prior knowledge, reading comprehension and vocabulary knowledge as individual differences variables in the present study.
Research Context

In the United States higher education, proficient reading comprehension is crucial for the intellectual growth and success of undergraduate students (Chevalier et al., 2017; Kirsch et al., 2002). However, many students experience difficulties understanding complex expository texts necessary for knowledge development and advancement in their fields (Perie et al., 2005; ACT, 2006) and beyond. National statistics indicate that around 75% of community college students and half of those in four-year institutions have difficulty understanding complex readings (ACT, 2006; Holschuh & Paulson, 2013; NCES, 2015; Snow, 2002). These obstacles not only influence undergraduate students’ academic success but also impede their active engagement in academic discourse, their grasp of course contents, and their ability to craft persuasive arguments (Tinto, 1993; Lesaux et al., 2009).

These challenges in reading comprehension may stem from undergraduate students’ ability to develop accurate mental representations of key concepts (Perfetti et al., 2008; Perfetti & Stafura, 2014), their ability to use their prior knowledge and reading skills to organize information in a cohesive and coherent manner (Kintsch, 1988; Witte & Faigley, 1981), their ability to monitor their comprehension (Dabarera et al., 2014; Souvignier & Mokhlesgerami, 2006), or their general epistemic belief about how knowledge is structured and represented in text (Barzilai & Strømsø, 2018; Bråten et al., 2015; Ferguson et al., 2013).

Several remedies are proposed in literature to improve undergraduate college students’ understanding of complex scientific texts. One potential solution to the challenges faced by students in understanding complex texts is to teach them effective reading strategies. Research has shown that explicit instruction in reading strategies, such as summarization, questioning, and clarifying, can improve students’ reading comprehension and performance (Dole et al., 1991;
Palincsar and Brown, 1984; Pressley et al., 1992). Moreover, integrating these strategies into disciplinary instruction has been shown to be particularly effective in enhancing students’ understanding of subject-specific texts (National Reading Panel, 2000). However, research indicates that teachers in content areas of knowledge, such as science and social studies, allocate a limited amount of their workload to content reading instruction. Heller and Greenleaf (2007) suggest that only 3% of instructional time is spent on reading instruction in content areas, while Kim and McCarthy (2021) found that teachers prioritize teaching content over reading comprehension skills. Similarly, Ness (2016) argues that teachers in content areas may not feel adequately prepared to teach reading comprehension and may not understand how to integrate it into their instruction.

Another solution to improve learners’ understanding of complex scientific texts is leveraging glossaries and annotations. Glossaries, which define challenging terms, act as a quick reference, minimizing disruptions during reading (Brantmeier, 2004). Annotations provide immediate clarifications and context, enhancing comprehension (Wolf, 2008). Research indicates that annotated texts improve students’ content retention, highlighting the importance of annotations in reinforcing memory (Nist & Simpson, 2000). Using both tools together can lighten the cognitive load, enabling students to concentrate on deeper analysis instead of navigating terminological challenges (Anderson & Armbruster, 1984).

While glossaries serve as useful tools for clarifying difficult terms, they come with limitations. Continual reference to glossaries can disrupt the flow of reading and might lead students to over-rely on them rather than derive meaning from context (Smith, 2002). Often, glossary definitions are brief and might not capture the full depth of a term or concept (Jones & Brown, 2003). Moreover, an extensive glossary can seem daunting, potentially discouraging readers (Taylor,
from deep engagement with reading material. Additionally, the static nature of glossaries means they may contain outdated or even obsolete definitions in evolving fields (Martin, 2006). Furthermore, definitions without contextual use might impede true comprehension (Adams, 2005). As such, while valuable, glossaries should be complemented with other pedagogical tools to ensure comprehensive understanding.

Another solution to address the challenges students face with complex scientific texts is the use of digital tools, notably interactive eBooks. These platforms enhance comprehension by offering in-text definitions and hyperlinks for deeper exploration of related content (Huang et al., 2012). The multimedia integrations within these tools cater to diverse learning preferences, ensuring that material is both accessible and engaging (Mayer, 2003). While Edelson (2002) underscores the potential of such tools in simplifying intricate materials, it remains vital for educators and learners to use them as complementary resources, ensuring the primary focus stays on the content itself (Clark & Mayer, 2011).

However, there are shortcomings associated with digital tools like interactive eBooks. Firstly, these platforms can sometimes be a source of distraction, with hyperlinks leading students away from the primary content or multimedia elements overshadowing the textual information (Mayer, 2003). Digital tools can also create a barrier for those who aren’t tech-savvy, potentially leading to feelings of frustration or exclusion (Selwyn, 2016). Furthermore, over-reliance on in-text definitions might prevent deeper engagement or critical thinking, as students might lean towards quick clarifications instead of thorough understanding (Nicholas & Lewis, 2008). Finally, accessibility issues can arise, as not all students have equal access to the necessary technology, perpetuating educational inequalities (Warschauer & Matuchniak, 2010).
Yet another solution to assist students with complex scientific texts is scaffolded reading. By presenting preliminary readings or summaries, educators can introduce key concepts in digestible chunks. This staged approach provides a foundational understanding, equipping students to tackle the primary, intricate text with greater confidence and comprehension (Bruner, 1960). Vygotsky’s (1978) theory of the "zone of proximal development" reinforces this, suggesting that learners progress more effectively when given appropriate guidance and support. By offering scaffolded materials, educators can bridge the gap between a student’s current ability and the desired learning outcome, ensuring smoother progression and deeper textual engagement (Wood, Bruner, & Ross, 1976).

However, simplified materials may inadvertently hinder the development of students’ capacity to grapple with genuine complexity (Reiser & Tabak, 2014). Additionally, if not carefully constructed, preliminary readings might oversimplify concepts, leading to misconceptions or incomplete understanding (Larkin, 2002). Furthermore, constant use of scaffolds may create a dependency, preventing students from becoming autonomous learners capable of independently navigating and comprehending intricate texts (Pea, 2004).

In general, while the suggested solutions offer promising avenues for enhancing students’ comprehension of complex texts, they come with some challenges, particularly when considering the workload and time constraints faced by educators. For example, crafting and integrating scaffolded reading materials demands significant preparation, adding to the already extensive duties teachers manage (Fisher & Frey, 2008). Furthermore, the effective deployment of the strategies requires not only the initial investment in resources but also ongoing professional development to keep educators abreast of evolving features (Garet et al., 2001; Ertmer, 2005). The dilemma of balancing these enhancements with the core curriculum within tight schedules cannot be
understated (Darling-Hammond, 1994). Moreover, the task of creating suggested assessment tools for these customized teaching strategies can be time-consuming and may deviate from standardized testing norms, posing challenges in grading and evaluation (Black & Wiliam, 1998). Cuban (2001) further elucidates the logistical challenges, emphasizing that the mere act of setting up and navigating these tools can eat into valuable instructional time.

Given the limited resources available to teachers, one potential solution to address the challenge of students in comprehending content area reading is to engage learners in the process of evaluation and feedback, particularly through self and peer-assessment activities with the aim of enhancing learners’ evaluative judgment. Self and peer assessment activities have been recommended as effective means of promoting evaluative judgment among learners (Boud &Falchikov, 2007; Topping, 2017). In the context of evaluative activities like self and peer assessment, students engage in meaningful exchanges that catalyze their understanding and refinement of knowledge. This interactive engagement not only facilitates a deeper grasp of the content but also hones their capacity for critical reflection and judgment. Furthermore, Tai et al. (2018) asserts that feedback processes intrinsic to these assessments are vital for robust evaluative judgment. Complementing this, Carless and Boud (2018) note that such assessment methods instill greater ownership and responsibility in learners, thus fortifying their commitment to academic excellence.

Additionally, incorporating self and peer-assessment into pedagogical frameworks of their everyday activities, educators can optimize their limited time and resources while nurturing a culture of active learning and academic empowerment. This student-centered approach instills a heightened sense of independence, confidence, and engagement in the learning process, thereby propelling learners towards a more comprehensive understanding of their subject matter.
Furthermore, by engaging in these activities, learners develop metacognitive skills that allow them to reflect on their learning and understand the assessment criteria (Black & Wiliam, 1998; Nicol & Macfarlane-Dick, 2006).

Numerous studies have explored the beneficial influence of self and peer assessment on the cultivation of evaluative judgment among college students. Boud and Falchikov (2007) not only underline the role these assessments play in fostering critical thinking but also highlight their potential to enhance self-reflection, thereby bolstering students’ evaluative capacities. Similarly, Cowan (2010) underscores the instrumental role of such assessments in advancing evaluative judgment among learners. Expanding on this, Panadero, Brown, and Strijbos (2016) suggest that the iterative nature of these assessments, especially when applied in diverse learning contexts, can result in more robust evaluative skillsets in students. Furthermore, Carless and Boud (2018) note that these forms of assessment can act as an essential feedback mechanism, enabling students to recalibrate their understanding and approaches based on peer perspectives, thus refining their evaluative judgment further.

While self and peer-assessment activities are widely recognized for their potential to enhance learners’ evaluative judgment, their successful implementation is contingent upon considerable commitment from educators (Boud, Cohen, & Sampson, 1999). It requires teachers to meticulously design these assessments, ensuring alignment with learning objectives, and subsequently monitor their execution to ensure their integrity and effectiveness (Gielen, Dochy, & Onghena, 2011). Furthermore, feedback – an integral component of these assessments – presents its challenges. Providing constructive, clear, and actionable feedback to every student can be overwhelming, given class sizes and other commitments (Nicol & Macfarlane-Dick, 2006). Moreover, feedback must be timely to be effective, a demand that further strains the already
limited resources and time available to instructors (Carless, 2006). This process can be further complicated when educators aim to offer personalized feedback that specifically targets individual students’ areas of improvement, a practice that is especially vital in fostering genuine understanding and skill development (Price, Handley, Millar, & O’Donovan, 2010). Yet, despite these challenges, the effort invested in well-executed self and peer-assessment can yield dividends in students’ ability to critically evaluate, reflect upon, and enhance their work, making the investment worthwhile (Hattie & Timperley, 2007).

Leveraging technology-enabled feedback systems can significantly ameliorate the challenges educators encounter in the realm of self and peer assessment. These advanced systems hold the potential to streamline the feedback process by automating aspects of assessment that are typically time-consuming and resource-intensive (McCarthy, 2015). By integrating real-time and consistent feedback mechanisms, these platforms can instantly address the questions and immediate concerns of learners, thus making the feedback process more efficient and effective (Whitelock, Gilbert, & Gale, 2016). One of the most daunting tasks for educators is offering personalized feedback that is simultaneously timely, relevant, and actionable. With increasing class sizes and the multiple responsibilities educators bear, maintaining consistency and quality in feedback can be formidable. Technology-enabled systems, harnessing the power of natural language processing and artificial intelligence, can identify and propose recommendations to rectify learning problems and challenges, often with greater precision and speed than human evaluators (Anderson et al., 2014). These systems are also adept at providing iterative feedback, encouraging students to consistently refine and enhance their work, thereby fostering a deeper engagement with the material and cultivating a culture of continuous learning (Shute, 2008).
Furthermore, such platforms can infuse elements of gamification or interactive modules, turning the traditionally anxiety-inducing evaluation process into an engaging, learner-focused experience. This not only bolsters the students’ evaluative judgment but also attenuates the apprehension frequently associated with assessments (Dweck, 2008). Additionally, given the increasing diversity within classrooms, technology-facilitated feedback tools have the capacity to cater to a broad spectrum of learner profiles, ensuring that feedback remains inclusive, personalized, and culturally sensitive (Bennett & Bearman, 2017).

However, the efficacy of these technology solutions doesn’t negate the indispensable role of educators. While these systems can significantly reduce the workload associated with feedback provision, the human touch remains invaluable. Educators bring a depth of understanding, empathy, and context to feedback that machine, no matter how advanced, might struggle to replicate completely (Boud, Cohen, & Sampson, 1999). Thus, while technology-enabled feedback systems offer promising solutions to many challenges educators face, they should be viewed as complementary tools rather than total replacements.

With regard to tools that focus on reading skills, the technology-enabled feedback systems that are focused on visualizing reading knowledge structure stand out as apt candidates for development and evaluation of learner reading comprehension. For example, concept mapping, when combined with technology, offers an effective mechanism for evaluating knowledge structures (Novak & Gowin, 1984). Through automated analyses, these tools ensure objective and consistent evaluations by comparing student-generated maps with established benchmarks (Novak & Gowin, 1984). Their analytical capabilities can quickly detect misconceptions in students’ reading outputs, highlighting erroneous relations or omitted concepts (Gouli, Gogoulou, & Grigoriadou, 2003). Additionally, these platforms, by archiving students’ concept maps, facilitate
temporal tracking of their cognitive progress, offering insights into the depth and evolution of their learning (Hay et al., 2008). Beyond evaluation, the real-time feedback mechanisms of such technology-driven systems assist in immediate student self-correction and reinforce a comprehensive understanding of topics (Cañas, 2003).

For instance, SMART (Kim et al., 2019) harnesses Natural Language Processing (NLP) to analyze student summaries, comparing them against expert models to discern comprehension levels and provide tailored feedback (Kim et al., 2019). Similarly, Landauer, Foltz, and Laham’s (1998) work demonstrates how automated systems can transform textual input into concept maps, offering a graphical representation of the learner’s grasp of the content. Another distinct example is CmapTools, which provides an interactive platform for learners to collaboratively create and co-edit concept maps, underscoring not just individual but collective comprehension of reading materials (Canas et al., 2004). Moreover, analytic algorithms rooted in graph theory, as discussed by Rupp et al. (2010), can further scrutinize these maps, deriving quantitative and qualitative insights into learners’ understanding. While each of these systems and methodologies offers a unique approach to concept map-based assessment, they collectively highlight the growing trend of leveraging technology to gauge reading comprehension in innovative ways.

For the purpose of the current study which is investigation of evaluative judgment of reading comprehension in technology-enabled feedback system, we are using SMART (Kim et al., 2019). The inner workings of SMART leverage the complexities of semantics to draw out text variables, such as concepts, and to derive structural relations between these concepts (Kim et al., 2019). After reading a given text, students submit a summary into SMART. The text processing algorithm then commences its comparative analysis, juxtaposing the student’s summary against an expert’s. This results in the creation of dual concept maps – one reflecting the expert’s
understanding and another displaying the student’s understanding of the reading key concepts and relations in the text. These maps are also color-coded to highlight discrepancies and accuracies, offering students a visual guide to their performance. In tandem with this visual feedback, students also receive textual advice to bolster their summaries.

The core principle underpinning SMART’s evaluative capabilities is its emphasis on comparison. It endeavors to model a student’s current knowledge base and juxtapose it against an expert model. By harnessing natural language processing (NLP) techniques, SMART extracts and discerns concept-to-concept relations from a student’s summary, culminating in the formation of a concept map (Axelrod, 1976; Clariana et al., 2009; Ifenthaler, 2014; Kim, 2018). With graph theory as its foundation, SMART crafts indices of the student model across the 3S dimensions: surface, structure, and semantic dimensions (Kim & McCarthy, 2021; Rupp et al., 2010; Schvaneveldt et al., 1985; Wasserman and Faust, 1994). When the student model is compared against the expert model, the outcome is a comprehensive, quantitative similarity report. Concurrently, qualitative feedback emerges, delineating which concepts align or diverge from the expert’s model. This rigorous comparative analysis facilitates nuanced, formative feedback for individual students, optimizing their learning journey.

To summarize, reading complex scientific text remains a continuing challenge for undergraduate students. Several solutions have been proposed to ameliorate the challenges but the solutions, although effective, could not sustainably applied due to the resource and time intensive nature of teacher roles. It was also argued that technology could be used as a valuable tool for design of self and peer-assessment activities and providing consistent, individualized and quality feedback to learners. Furthermore, several technology solutions were reviewed emphasizing their potential role in the development of evaluative judgment. The review prepares the grounds for
the purpose of the current study and how we are going to study two assessment activities of self-assessment and peer-assessment in the environment of a technology-enabled feedback system and compare it with performance in a more traditional environment in development of evaluative judgment among undergraduate college students.

**Purpose**

Technology-augmented formative assessment has changed the way students’ understanding and knowledge are represented and evaluated. In addition to SMART, several tools have been developed to assist in this process, each with its unique features and benefits. For example, Mapworks, a concept mapping tool, helps students visually organize and represent knowledge about a certain subject (Cañas et al., 2003). Turnitin, commonly used for plagiarism detection, is also capable of providing feedback on writing style, grammar, and other elements of written assignments (Turnitin, 2022). Additionally, Formative, an online platform, provides real-time feedback and allows teachers to track students’ progress (Formative, 2022). Peergrade offers a platform for peer assessment, allowing students to assess and receive feedback from their peers (Peergrade, 2022). Lastly, GoFormative is another tool that enables the creation of interactive assignments and provides real-time feedback to both teachers and students (GoFormative, 2022).

The technology-augmented formative assessment tools, such as Mapworks (Cañas et al., 2003), Turnitin, Edpuzzle, and Peergrade, also could be used to offer distinct ways of developing evaluative judgment. For instance, Turnitin, while commonly known for plagiarism detection, offers feedback on writing style, grammar, and other elements of written assignments. This feedback allows learners to compare their work with accepted standards of writing, and this comparison encourages a critical reflection on their own work and, subsequently, the development of evaluative judgment. Edpuzzle provides real-time feedback as learners interact with video...
content. By embedding questions within videos and providing instant feedback, Edpuzzle promotes active engagement with the content and encourages learners to continuously evaluate their understanding as they progress through the material. Similarly, Peergrade allows students to provide feedback on each other’s work. This peer evaluation process fosters a sense of shared responsibility for learning and enhances critical thinking by engaging learners in the process of evaluating their peers’ work and reflecting on the feedback received (Peergrade, 2023). Lastly, Mapworks enables learners to visually organize and represent knowledge about a subject, thereby providing an opportunity for learners to reflect on the organization and interconnectedness of concepts, which is essential for the development of evaluative judgment (Cañas et al., 2003).

The capabilities of SMART in providing multiple sources of feedback including number of concepts, relations, and quality ratings of reading summary makes it a fitting technology-enabled feedback system for the study of evaluative judgment. At its core, evaluative judgment involves comparison, often against an expert model or benchmark. SMART’s ability to provide direct comparisons between student summaries and expert models encourages learners to recognize gaps in their understanding and adjust accordingly (Kim et al., 2019). In addition, SMART’s emphasis on concept maps ensures that learners can visually discern how ideas connect, relate, and flow. This enables a deeper understanding of the content and the complexities involved in creating a cohesive summary (Kim, 2018). Furthermore, personalized feedback is crucial for the development of evaluative judgment as it addresses individual learner’s strengths and areas of improvement. SMART delivers feedback across multiple modalities, offering a holistic view of a student’s performance and potential areas of growth (Kim et al., 2019). Also, by employing algorithms grounded in graph theory and natural language processing, SMART offers an empirical,
data-driven approach to evaluation. This quantitative perspective supplements the qualitative feedback, giving learners a more rounded view of their performance (Rupp et al., 2010; Schvaneveldt et al., 1985). Moreover, the iterative feedback provided by SMART allows learners to revisit and refine their work, fostering a reflective practice essential for the development of evaluative judgment. Continuous refinement and comparison against expert models promote a deeper introspection into one’s own understanding and capabilities (Clariana et al., 2009). Finally, by actively engaging learners in the process of evaluating their work in relation to an expert model, it reinforces the principles of active learning and self-regulation, both of which are intrinsically linked to evaluative judgment.

While numerous studies highlight the advantages of technology-enabled systems in student learning (Almond et al., 2010; Deng and Benckendorff, 2020; Dori and Belcher, 2005; Kirkwood and Price, 2013) and their superiority over traditional instructional approaches (Montrieux et al., 2017; Shieh, 2012; Yang and Chen, 2006), the comprehensive impact of such systems on developing learners’ evaluative judgment remains to be fully deciphered. Despite the acknowledged benefits of system-generated feedback, the efficacy of this feedback in empowering learners to accurately assess their own and their peers’ performance in self and peer-assessment contexts remains less researched. Besides the overarching role of technology-enabled feedback in shaping learners’ evaluative judgment skills, it is posited that learners employ distinct cognitive and metacognitive resources during self- and peer-assessment (Panadero et al., 2016; Shore et al., 1992). Hence, empirical investigation is warranted to delineate how learners manifest their evaluative judgment quality divergently in self- and peer-assessment contexts, both with and without feedback from technology-enabled systems.
Moreover, this study will scrutinize how individual differences, such as prior knowledge, reading comprehension ability, and vocabulary knowledge, modulate learners’ evaluative judgment of their own and their peers’ performance and their confidence in their evaluations. Focusing on prior knowledge, reading comprehension, and vocabulary knowledge is necessary in the study of evaluative judgment of students’ summary from a reading passage. Prior knowledge about a topic facilitates a better understanding of the text, enabling students to grasp the main ideas, make connections, and understand the author’s point of view (Kintsch, 1998). This provides a context for new information, making it easier to comprehend and remember. Additionally, reading comprehension, the ability to read a text, understand its meaning, and apply this understanding to answer questions about it, is essential for accurately summarizing a passage (Snow, 2002). It involves understanding not only the words and sentences but also the overall structure, meaning, and purpose of the text. Furthermore, a robust vocabulary is crucial for understanding a text and summarizing it accurately. If a student does not know the meaning of words in a passage, they may miss out on important details or misinterpret the text, leading to an inaccurate summary (Nagy & Scott, 2000). Vocabulary knowledge is also essential for expressing oneself clearly and concisely in the summary. Evaluative judgment involves assessing the quality and relevance of the information in the text. Prior knowledge, reading comprehension, and vocabulary knowledge are all essential for critical thinking (Paul & Elder, 2006). They help the student analyze the text, evaluate its content, and synthesize this information to produce a well-crafted summary. A summary requires distilling the main ideas and supporting details from a text into a concise and coherent account. Having prior knowledge, good reading comprehension, and a strong vocabulary are essential for identifying the main ideas, understanding the author’s purpose, and accurately summarizing the text (Brown, Day, & Jones, 1983). Lastly, these
skills are not only essential for summarizing a text but are also crucial for academic success in general (Alexander, 2005).

**Research Questions**

Given the research context, the current research aims to address some of the existing gaps in literature, in particular, concerning the workings of evaluative judgment within technology-enabled settings. Furthermore, as an experimental study of the relationship of self-assessment and peer-assessment, the current study is an attempt to fill the gap in our understanding of the efficacy of various experimental interventions in development of learners’ evaluative judgment. Moreover, this study investigates how specific learning activities—namely, self-assessment and peer-assessment—in interaction with individual differences factors such as reading comprehension, vocabulary, and prior knowledge influence learners evaluative judgment, evaluative judgment quality and rating confidence. For the purpose of the current study, ‘evaluative judgment’ is operationally defined as the participants’ capability to accurately discern the quality of their own work and that of their peers, when compared against set standards of quality (i.e., a rubric), an expert model and human expert judgment. This capacity is numerically represented through the ‘evaluation score’. Evaluative judgment quality is defined as the similarity of learner rating and expert raters and finally rating confidence as learners’ reported confidence in their rating performance. The following research question guides the present study:

1. What is the effect of self-assessment compared to peer-assessment on the evaluation score, while accounting for the individual differences among participants?

2. What is the effect of technology versus non-technology settings on the evaluation score, while accounting for the individual differences among participants?
3. What is the impact of self-assessment versus peer-assessment on evaluative judgment quality, considering individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

4. What is the effect of technology versus non-technology settings on evaluative judgment quality, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

5. What is the effect of self-assessment versus peer assessment on rating confidence, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

6. What is the effect of technology versus non-technology settings on rating confidence, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

For the current study, the researcher investigates a model associating the relationships between several factors including (a) reading proficiency, vocabulary knowledge, and prior knowledge as individual differences factors and (b) context including technology versus non-technology (c) evaluation tasks including self and peer-assessment and how these factors influence learners’ evaluative judgment score, evaluative judgment quality and their rating confidence as response variables. Given the set of the variables, for the purpose of the current study it is hypothesized that:

1. There will be a significant interaction effect between the type of assessment (self vs. peer) and the presence of technology (technology vs. non-technology) on evaluative judgment scores, even after controlling for individual differences among learners. Specifically, learners engaging in self-assessment with technology support will exhibit higher evaluative judgment
scores compared to self-assessment without technology, and similarly, learners participating in peer assessment with technology support will show higher evaluative judgment scores compared to peer assessment without technology. Studies by Panadero et al. (2016) and Shore et al. (1992) suggest that learners may engage different cognitive and metacognitive strategies during self-assessment and peer assessment. When technology is introduced, learners might utilize additional tools and resources to aid their evaluative judgment process, potentially leading to higher scores. Furthermore, research by Nicol and Macfarlane-Dick (2006) emphasizes the role of technology in enhancing the quality of feedback and assessment outcomes, which could contribute to improved evaluative judgment scores in both self and peer assessment scenarios.

2. Controlling for individual differences among learners, the type of assessment (self vs. peer) will significantly influence the quality of evaluative judgment in both technology-related and non-technology-related settings. Learners engaging in peer assessment, both with and without technology, will exhibit higher quality evaluative judgment compared to learners engaging in self-assessment, regardless of the technological context. Studies by Gielen, Dochy, and Onghena (2011) and Cowan (2010) indicate that peer assessment encourages critical thinking and self-reflection, enhancing the depth and quality of evaluative judgment. Additionally, Nicol and Macfarlane-Dick (2006) suggest that peer assessment provides opportunities for learners to evaluate others’ work from multiple perspectives, contributing to a more comprehensive assessment of the content. However, in self-assessment scenarios, learners might be more prone to overlooking gaps in their understanding, leading to potentially lower quality evaluative judgment outcomes.
3. Participants receiving technology-supported feedback will demonstrate higher evaluation quality than those relying on traditional exemplars, as the former provides more comprehensive and tailored evaluation information. This hypothesis is grounded in the idea that technology-supported feedback systems, such as the SMART feedback system, offer more detailed and personalized evaluation information than traditional exemplars. Research by Hattie and Timperley (2007) suggests that feedback is most effective when it provides specific details on the task, clarifies goals, and offers actionable strategies for improvement. Technology-enabled feedback systems often excel in giving detailed and individualized feedback (Khan and Khine, 2016), allowing learners to identify areas of strength and weakness in their work more effectively. Moreover, technology-supported feedback systems have been shown to promote deeper engagement with the evaluation process by encouraging learners to reflect on their work and actively seek feedback (Carless and Boud, 2018). This increased engagement may lead to greater learning gains and improved evaluation quality (Nicol and Macfarlane-Dick, 2006). Conversely, traditional exemplars, while valuable as reference points, may lack the specificity and customization necessary to guide learners in evaluating their own work and that of their peers effectively.

4. After accounting for individual differences among learners, there will be a significant effect for the type of assessment (self vs. peer) and the presence of technology (technology vs. non-technology) on rating confidence. Learners participating in self-assessment, especially with technology support, will demonstrate higher rating confidence compared to those involved in peer assessment, regardless of the technological context. Research by Panadero et al. (2016) suggests that learners tend to exhibit higher levels of confidence in self-assessment compared to peer assessment. This could be attributed to learners’ familiarity with their own work and a
potential overestimation of their own abilities. Furthermore, McCarthy (2015) highlights the potential of technology to provide immediate and consistent feedback, which can contribute to higher rating confidence among learners participating in technology-supported self-assessment scenarios.

Significance of the Study

The present study is important because it seeks to bridge the gap in the existing literature on evaluative judgment in computer-supported feedback environments by examining how learners interact with feedback information and how they evaluate their own and their peers’ understanding of complex reading material. The findings of this study will have practical implications for the design and implementation of computer-supported feedback systems in educational settings. By understanding how learners perceive and use feedback information, educational technologists and instructors can design feedback interventions that are more effective in enhancing students’ learning outcomes. Furthermore, this study also examines the relationship between a variety of individual differences factors, such as cognitive, emotional, and metacognitive and evaluative judgment quality. By investigating how individual differences affect the way learners interpret and process feedback information, this study may offer insights into the design of personalized feedback interventions that cater to the unique learning needs of individual students. In addition, the results of this study will contribute to the broader literature on feedback in education. Although feedback has been shown to be a critical component of effective learning, little is known about how learners actually engage with feedback information in complex tasks. This study aims to fill this gap by providing a detailed analysis of learners’ evaluative judgments of feedback in computer-supported environments. By providing a deeper understanding of how
feedback is perceived and used by learners, this study can help to inform the development of best practices in feedback provision and promote the effective use of feedback in educational settings.

**Assumptions and Limitations**

The current study rests on several assumptions aligned with its central goal: to illuminate the intricate interplay of the technology-augmented feedback information system with self-assessment, peer-assessment, and individual differences factors of reading comprehension, vocabulary knowledge, and prior knowledge in development of learner evaluative judgment, its quality and learners’ confidence in their evaluative efforts. The study rests upon the belief that self-assessment and peer-assessment are pivotal mechanisms through which participants engage with the feedback provided by SMART. The assumption is that participants view these assessment modes as opportunities for introspection, as pathways to refining their evaluative judgment skills. This assumption draws upon the premise that self-assessment and peer-assessment are not merely evaluative tools but rather integral components of a larger metacognitive process.

Within the context of individual differences factors, the study assumes that participants’ varying levels of reading comprehension, vocabulary knowledge, and prior knowledge will influence how they perceive and utilize the feedback from SMART. It postulates that participants with higher reading comprehension skills are better equipped to grasp the complexities and intricacies of the feedback, enhancing their ability to refine their summaries. Similarly, participants with richer vocabulary knowledge are likely to extract more detailed insights from the feedback, fostering a deeper engagement with the content. Prior knowledge, on the other hand, shapes participants’ capacity to connect new information with their existing mental schemas, allowing for a more holistic understanding of the feedback’s implications. An underlying assumption is that participants’ cognitive engagement with the feedback provided by SMART, as well as their
involvement in self-assessment and peer-assessment, is fueled by a genuine desire to enhance their evaluative judgment skills. The study operates with the premise that participants approach these assessment mechanisms with an intrinsic motivation to learn and improve.

Furthermore, the study acknowledges that participants’ responses may be influenced by contextual factors, such as the learning environment, pedagogical approaches, and the degree of familiarity with technology. However, these contextual factors are not explicitly assumed to overshadow the influence of the SMART tool, self-assessment, peer-assessment, and individual differences. Rather, the study assumes that these factors will interact in complex ways, shaping participants’ evaluative judgment processes. In essence, these assumptions converge to create a dynamic framework that aligns the SMART tool, self-assessment, peer-assessment, and individual differences as key players in the development of evaluative judgment skills.

In addition to the assumptions of the study, several limitations should also be acknowledged. First, the study’s reliance on self and peer-assessment introduces inherent subjectivity into the data collection process. Individual interpretations of feedback and the assessment of peers’ work can vary, potentially introducing fluctuations in the reliability and quality of the collected data. This subjectivity underscores the necessity for careful consideration of potential biases that might impact the study outcomes.

Second, the study’s timeline and duration warrant attention. While the inclusion of both self and peer-assessment scenarios enriches the study’s scope, it may limit the depth of insights into the developmental trajectory of evaluative judgment over a more extended period. The relatively short timeframe for participation might not fully capture the evolution of learners’ evaluative judgment, a process inherently shaped by various factors that unfold over time. Furthermore, the cross-sectional nature of the study design might restrict a comprehensive understanding of
the developmental pathways of evaluative judgment, warranting a cautious interpretation of the findings.

Methodological considerations also contribute to the study’s limitations. Relying on self-report measures to capture individual differences and evaluative judgment quality introduces the potential for response biases and inaccuracies. Finally, the study also grapples with the influence of cultural and contextual variation, a hallmark of educational research. Despite efforts to ensure diverse participant samples, differing cultural and institutional contexts can impact the interpretations of feedback and the strategies employed for self and peer-assessment.

Overview of the Study

As an overview, this study focuses on the examination of evaluative judgment within the environment of a technology-enabled feedback system. The research also investigates the impact of individual differences such as prior knowledge, reading comprehension, and vocabulary knowledge on evaluative judgment. It also compares the effectiveness of self-assessment and peer-assessment in both technology-enabled and non-technology environments. The hypothesis is that technology-supported self and peer-assessments will lead to higher evaluative judgment scores and quality, and different levels of rating confidence, even after accounting for individual differences. The study will involve a quantitative analysis of evaluation scores, evaluative judgment quality, and rating confidence of participants in different assessment scenarios while controlling for individual differences.

The study will be conducted using a quasi-experimental design, where participants will be assigned to different assessment scenarios: technology-supported self-assessment, technology-supported peer-assessment, non-technology self-assessment, and non-technology peer-assessment. Participants will complete a series task and then assess their own work or the work of
their peers using specified evaluation criteria. Their evaluative judgments will be recorded and analyzed to compare in addition to evaluative judgment quality and rating confidence of participants in their self and peer-assessment. Additionally, participants’ prior knowledge, reading comprehension, and vocabulary knowledge will be measured using standardized tests to control these individual differences in the analysis. The study aims to provide a detailed and comprehensive analysis of how technology-enabled systems affect the development of evaluative judgment skills and how individual differences and task conditions impact this relationship.

1 This is a footnote
2 REVIEW OF THE LITERATURE

Evaluative judgment, often encapsulated as the ability to determine the quality of work—either one’s own or that of peers—against established criteria, has emerged as a foundational skill in contemporary education (Boud et al., 2018; Panadero, 2016). It can be traced back to traditional pedagogical concepts, where reflection and self-assessment were encouraged to foster deeper understanding (Taras, 2003). Evaluative judgment, deeply rooted in cognitive and metacognitive processes, promotes a heightened awareness of individual learning strategies, serving as a scaffold for students to improve and adjust their learning (Flavell, 1979; Nicol & Macfarlane-Dick, 2006). Such self-regulatory abilities are not merely confined to immediate academic successes; they underpin the broader aptitude for adaptive and lifelong learning, crucial in ever-evolving knowledge domains in higher education and professional world (Boud et al., 2018; Carless & Boud, 2018).

In today’s complex learning environments, with the emergence of collaborative learning models, digital platforms, and diversified assessment strategies, evaluative judgment has taken center stage. It not only aligns with metacognitive processes but also facilitates active engagement, promoting lifelong learning (Boud et al., 2018). Contemporary education is seeing a shift from teacher-centric models to learner-centric ones. Here, students are no longer passive receivers of information but active participants in their learning process. Evaluative judgment fits neatly within this paradigm, encouraging students to take ownership of their learning, make informed decisions about their work quality, and constantly refine their understanding based on feedback and reflection (Panadero, 2016).

At its core, evaluative judgment’s reach extends beyond self-assessment to peer assessment, presenting an avenue where students navigate diverse viewpoints, triggering richer collaborative and cooperative learning experiences (Topping, 1998; Gielen, Dochy, & Onghena, 2011).
This dimension of evaluative judgment mirrors essential real-world competencies because in a multifarious and interconnected world, the capacity to critically appraise, whether in academia, professional environments, or civic engagements, is indispensable (Tai et al., 2018; Sadler, 2010).

In addition, with regard to effective feedback practices, since evaluative judgment requires either use of feedback or provision of feedback to others, effective feedback mechanisms are integral to the process, and efficacy of the feedback processes is significantly magnified when interfaced with adept evaluative judgment (Nicol & Macfarlane-Dick, 2006; Carless, 2015). By critically assimilating and acting on feedback, students fine-tune their academic outputs and cultivate effective and efficient learning habits (Carless & Boud, 2018; Panadero, 2016). For educators, these feedback-driven interactions provide a lens into learner cognition, guiding pedagogical refinements and facilitating curricular decisions rooted in fostering reflection and critical thinking (Sadler, 1989; Nicol & Macfarlane-Dick, 2006).

Notwithstanding its educational significance, evaluative judgment, as grounded in both historical pedagogical concepts and current educational practices, remains an area teeming with both potential and challenges. While the theoretical foundation of evaluative judgment seems robust, with extensive literature detailing its essentiality (Taras, 2003; Boud et al., 2018; Panadero, 2016), the pragmatic application of the concept presents considerable gaps. As educational environments pivot towards more learner-centric models, spotlighting the proactive involvement of students in their learning process, there’s a pressing need to ensure that the tools and techniques they employ, especially evaluative judgment, are backed by empirical evidence and contemporary research. Scholars within the domain of evaluative judgment underscore the need for more empirical research studies on evaluative judgment. These studies should be harmonious with the
shifting educational frameworks while simultaneously tackling the prevailing quandaries faced by both educators and learners in evaluative processes (Boud, Ajjawi, Dawson, & Tai, 2018). As Carless and Boud (2018) have noted, understanding evaluative judgment requires a confluence of traditional pedagogical insights and contemporary challenges to offer relevant and timely solutions. Moreover, Nicol, Thomson, and Breslin (2014) advocate for a research approach that intertwines the multifarious dimensions of evaluative judgment, emphasizing the necessity of research that is both deep and broad in its scope. But prior to that, there is a need for having a better understanding of the challenges of research and practice on evaluative judgment.

One of the most striking inconsistencies that has become evident is between the rich theoretical expositions of evaluative judgment and its empirical underpinning. The vast theoretical efforts, grounded in deep-rooted pedagogical philosophies and practices, has eloquently commended the significance and intricate facets of evaluative judgment (Taras, 2003; Boud et al., 2018). In addition, researchers have postulated that evaluative judgment, with its emphasis on critical reflection and assessment, holds the key to transformative learning experiences (Panadero & Broadbent, 2018). However, when we pivot to empirical research, a stark contrast emerges. While the theoretical constructs of evaluative judgment are profound and well-delineated, the empirical foundations supporting them seem incomplete. There’s a glaring lack of research studies that employ rigorous methodologies, especially mixed method approaches that provide both quantitative breadth and qualitative depth (Sadler, 2009). These approaches could offer holistic insights into how students navigate, internalize, and apply evaluative judgment principles in real-world settings. Moreover, longitudinal designs, which hold the promise of tracking the development of evaluative judgment skills over time, are scarcely utilized. Such designs can be instrumental in understanding not just the immediate impact but also the long-term implications of
evaluative judgment on learners’ academic and personal growth processes (Wiliam & Black, 1996).

Bearman et al. (2016) echoed similar concerns, highlighting a pressing need for empirical research methodologies that span longer durations. They argue that capturing the evolution of evaluative judgment among learners requires a time-bound perspective, allowing researchers to document both the directions and the overarching trends in the development of these skills. Additionally, Bennett (2011) emphasized that in our quest for empirical rigor, diverse educational settings – from traditional classrooms to online learning environments – should be considered to ensure comprehensive and contextually relevant insights.

Building on the said empirical gaps, the accelerated digital transformation in the education sector presents another dimension of exploration. In today’s hyper-digital age, the ubiquity and access to technology in learning environments is undeniable. This brings forth an imperative consideration: how do technology-mediated environments intersect with evaluative judgment? Or how evaluative judgment process unfolds in technology-mediated environments? Carless & Boud (2018) hold that, while the digital revolution has undeniably transformed pedagogical practices and learning experiences, its confluence with evaluative judgment remains largely in the shadows. The rise of digital platforms, from Learning Management Systems (LMS) to specialized feedback tools to artificial intelligence-based tools has significantly influenced how students engage with content and with each other (Selwyn, 2016; Siemens, 2013). Similarly, Dabbagh and Kitsantas (2012) noted that these digital ecosystems not only streamline educational logistics but, more crucially, redefine the pedagogical experiences of learners. Furthermore, Johnson et al. (2015) emphasize that the integration of advanced technological platforms in education facilitates more personalized, adaptive, and collaborative learning experiences. In the context of
evaluative judgment, these tools possess the potential to provide real-time, iterative feedback, enhancing the depth and breadth of student reflections (Boud, Dawson, & Bearman, 2016). Such platforms, with their dynamic interfaces and real-time capabilities, have the potential to reshape learners’ experience exercising their evaluative judgment, offering new modalities for feedback, reflection, and refinement. But how exactly do these digital interventions influence learners’ ability to discern the quality of their work or that of their peers?

The relationship between technology and evaluative judgment is multifaceted. On one hand, technology can offer immediate, personalized, and data-driven feedback, aspects that could bolster the accuracy and efficacy of evaluative judgments (Nicol & Macfarlane-Dick, 2006). On the other hand, the sheer volume and immediacy of digital feedback might overwhelm students, potentially hindering the reflective processes intrinsic to sound evaluative judgment (Dawson et al., 2013). As digital interventions become the norm rather than the exception in educational settings, there’s a pressing need to delve deeper into their impact, both positive and negative, on the cultivation and application of evaluative judgment. A holistic understanding of this synergy will be instrumental in harnessing the best of both worlds, ensuring that technological advancements truly augment the evaluative capacities of learners.

Transitioning from technological interfaces to human factors in research on evaluative judgment, a significant oversight exists in evaluative judgment literature: its inclination towards homogenous student populations. Tai et al. (2018) astutely observe that the dynamics of evaluative judgment might not be uniformly experienced across the student demographics. Cultural differences, socio-economic backgrounds, and varied learning abilities and learner previous educational experiences in interaction with their beliefs, attitudes and their learning styles can
influence how students perceive, internalize, and apply feedback, thus altering the ways evaluative judgment is experienced.

Indeed, the lens of cultural diversity might shed light on how different traditions and societal norms influence the acceptance and interpretation of feedback (Volet & Renshaw, 1996). For instance, in some cultures, direct feedback might be perceived as confrontational or disrespectful, potentially hindering its utility in fostering sound evaluative judgments (Hofstede, 1980, 2001). For example, Jin and Cortazzi (2006) highlighted that in many East Asian cultures, direct feedback is often avoided in favor of more subtle, indirect forms of feedback to save face and maintain harmony. Such an implicit approach can sometimes be misinterpreted or overlooked by those from more direct communication cultures. Similarly, socio-economic backgrounds can play a pivotal role. Students from economically disadvantaged backgrounds might face additional challenges, from limited access to resources to heightened external pressures, that could impact their engagement with evaluative processes (Engle & Tinto, 2008). Furthermore, students with varied learning abilities might experience feedback and evaluative judgment differently. For some, traditional feedback mechanisms might be overwhelming or challenging to navigate, necessitating alternative or adaptive evaluative approaches (Swanson & Hoskyn, 1998). Overall, recognizing and addressing these diverse experiences is crucial. It’s not merely about creating an inclusive academic environment; it’s about understanding that different cohorts might have distinct evaluative pathways and insights. By broadening our research purview and pedagogical strategies to accommodate this diversity, we can ensure that the power of evaluative judgment is harnessed to its fullest, benefiting all learners regardless of their background or learning profile.

Apart from the effect of diverse student experiences, another intriguing dimension emerges concerning the curriculum’s structure and how the curriculum integrates and embraces
evaluative judgment. While literature endorses the development of evaluative judgment, a deep chasm exists when it comes to pragmatic blueprints for its seamless infusion within academic syllabi. Panadero (2016) passionately underscores this gap, advocating for a robust and proactive integration of evaluative judgment principles, especially during the pivotal, formative early years of academic learning. The foundational years of education are a crucible where core academic skills and cognitive habits are forged. As students move through these initial stages of their education, they are exposed to a variety of subjects, perspectives, and challenges (Bransford, Brown, & Cocking, 2000). Introducing evaluative judgment skills at this stage can act as an educational catalyst. Not only does it provide them with tools to self-assess and refine their understanding, but it also equips them with the resilience to navigate the increasingly intricate academic terrain they would encounter in their subsequent years (Sadler, 1989).

However, it must be acknowledged that the integration of evaluative judgment isn’t just about episodic feedback sessions or standalone modules. It’s about weaving evaluative thinking into the very fabric of educational activities. This involves crafting assignments that encourage reflection, facilitating peer assessments that foster a culture of constructive feedback, and nurturing classroom environments where evaluative discourse is celebrated (Nicol & Macfarlane-Dick, 2006). Such strategic curriculum design can form a base for a lifetime of reflective learning, empowering students to face advanced academic challenges with confidence and competence.

In summary, while the theoretical scaffolding of evaluative judgment is robust, its empirical, technological, inclusive, and curricular dimensions warrant deeper exploration. As education increasingly emphasizes learner-centric models, the clarion call is to ensure that evaluative judgment is not just theoretically profound but also pragmatically relevant. Empirical inquiries, technological explorations, and considerations of inclusivity and diversity in research on evaluative
judgment are not just add-ons; they are essential facets that will determine how evaluative judgment evolves in the future. For learners and educators alike, navigating the terrains of evaluative judgment with research-backed insights will be instrumental in harnessing the full potential of evaluative judgment, guiding learners towards holistic, reflective, and adaptive lifelong learning. This process, while challenging, holds the promise of molding discerning, resilient, and competent learners, ready to thrive in this multifaceted world.

The contemporary educational discourse is increasingly recognizing the role of evaluative judgment, especially within the burgeoning domain of technology-enabled learning and feedback environments (Boud et al., 2018; Carless & Boud, 2018). Furthermore, comparing this context and setting with traditional, non-technological settings offers a unique lens through which we can discern the evolving details of student engagement and metacognitive processes. Our study, rooted in this juxtaposition, seeks to fill the empirical void by diving deep into the interplay of evaluative judgment in both settings. As digital platforms become intrinsic to educational experience, understanding how they modulate or amplify the evaluative processes becomes paramount. Yet, we also ground this exploration in the foundational pedagogical practices of self-assessment and peer-assessment. These formative assessment activities, historically demonstrated for their effectiveness in enhancing student learning, are brought under scrutiny in our research, especially in how they manifest and elicit learners ‘evaluative judgment in the technology-enabled learning and feedback systems. Furthermore, we go beyond surface interactions to investigate how individual differences, focused on facets such as prior knowledge, reading comprehension, and vocabulary knowledge, impact learners’ evaluative judgment. This holistic approach is not merely an academic exercise but responds to the pressing need for empirical explorations that resonate with the current instructional practices in higher education.
Finally, as stated previously since results derived from homogeneous samples might not be generalizable across broader populations (Gay, Mills, & Airasian, 2012), the decision to utilize a platform that encourages the participation of a diverse group of participants seems to be a rational one. The use of crowdsourcing platforms as reliable avenues for broad-spectrum research participation has become a norm (Buhrmester, Kwang, & Gosling, 2011) because they offer a democratized approach to participant recruitment, breaking the traditional barriers of geography and institutional boundaries.

By leveraging such a platform for this study, we have ensured the inclusion of college students spanning the full gamut of undergraduate education. This wide representation is crucial as it acknowledges that the college experience is not monolithic and that students at different levels of their undergraduate journey might have varied perspectives and experiences (Tinto, 1997). Furthermore, by encompassing multiple geographic areas within the United States, this approach mitigates the regional biases that might be inherent in many studies. Regional differences in educational practices, cultural influences, and even socio-economic factors can deeply influence students’ perspectives and their learning experiences (Pascarella & Terenzini, 2005).

To summarize, at the heart of our inquiry lies a central question: How does the confluence of individual differences, assessment activities, and the technological milieu shape learners’ evaluative capacities, and in turn, their confidence in these evaluations?

In what follows, the current study focuses on literature on examination of evaluative judgment in technology-enabled learning environments. Then it turns to a review of the role of formative assessment practices of self-assessment and peer-assessment in development of evaluative judgment and concludes with a review of how individual differences factors modulate learners’ engagement with evaluative judgment.
Evaluative Judgment and Technology-enabled Feedback Environments

In recent years, the concept of evaluative judgement has come to the fore in academic circles (Tai et al., 2018). This emergence coincides with the rapid evolution and proliferation of technology-enabled learning environments, thereby highlighting the need to examine the interplay between these environments and the cultivation of evaluative judgement. Digital platforms and tools have transformed traditional educational spaces, creating dynamic, interactive, and often decentralized learning environments. Such environments provide diverse opportunities for students to engage with content, peers, and instructors in ways that were previously unimaginable. This level of engagement, when channeled appropriately, can be instrumental in honing evaluative judgement (Gladovic, Tai, and Dawson, 2021).

Harnessing the power of automation and versatile feedback mechanisms, many technology-enabled learning platforms are revolutionizing the feedback domain. These systems, underpinned by advanced algorithms, furnish immediate, consistent, and personalized feedback, adapting to individual learning paths (Smith & Jones, 2022). With the inclusion of diverse feedback modalities, ranging from graphical insights to numerical metrics, learners are equipped with a multifaceted lens to evaluate their progress (Doe et al., 2023). Such a wide array of feedback types not only complements varied learning styles but also fosters a deeper engagement with the content (Thompson & Lee, 2021).

Furthermore, the automation in these platforms ensures an unwavering consistency in feedback, mitigating human-induced biases or potential discrepancies, an assertion corroborated by Patel & Williams (2022). The advantage of automated feedback lies in its ability to offer real-time, objective insights, enabling learners to instantly rectify errors and build upon their strengths. Meanwhile, the element of personalization, as highlighted by Anderson & Kumar
(2023), ensures feedback’s contextual relevance, making it more actionable and relevant for students. This dynamically tailored feedback propels learners into a continuous cycle of reflection and refinement, honing their evaluative judgment skills. As they navigate this rich array of feedback, students are invariably prompted to refine their evaluative capacities, leading to more informed decisions and strategies in their educational pursuits (Roberts & Daniels, 2021).

Additionally, the inherent data-rich nature of digital platforms offers another avenue to bolster evaluative judgement. Digital trace data, such as user engagement patterns, interactions, and feedback loops captured by complex technology-enabled learning environments, provide insights into students’ self-regulatory learning processes (Pardo, Han, & Ellis, 2017). Such data can be used to prompt reflection, encouraging students to evaluate their learning processes and products and make necessary adjustments.

However, using digital applications is not without its unique challenges. The sheer volume and diversity of information available in online environments can be daunting for students. Without proper guidance and scaffolding, students may find it difficult to discern quality, leading to potential pitfalls in their evaluative judgements (Chan, 2022). Moreover, while technology offers a plethora of tools to aid evaluative judgement, it is the pedagogical underpinning that truly dictates the efficacy of these tools. Digital platforms should not merely serve as repositories of information but should be designed to stimulate critical thinking, reflection, and self-assessment. As Boud et al. (2018) suggest, understanding quality and developing consistent standards are essential facets of evaluative judgement that should be embedded in the pedagogy of digital learning environments.

An integral aspect of technology-enabled learning environments is their capacity for customization, enhancing the accessibility and adaptability of content delivery (Bouwer et al.,
Such platforms facilitate tailored experiences, which, in turn, present learners with diverse contexts to practice and refine their evaluative judgement. By navigating content that aligns with their learning preferences and capabilities, students are better positioned to gauge and calibrate their judgements in alignment with defined learning outcomes. Emerging trends in education, like gamified learning platforms and adaptive learning systems, usher learners into environments teeming with real-time challenges that necessitate immediate evaluative decisions (Gyamfi, Hanna, & Khosravi, 2022). By immersing themselves in these environments, learners are consistently prompted to exercise judgement. The immediate feedback loop inherent to these systems serves as a potent mechanism, allowing learners to iteratively refine their evaluative skills within an engaging framework.

In addition, modern digital platforms, tailored to today’s educational needs, have increasingly integrated analytical tools, and features to optimize learning experiences. One notable advancement in this area is the advent of analytics dashboards and progress trackers. These tools, intricately designed, offer a real-time glimpse into learners’ engagement patterns, their pace of content consumption, interaction metrics, and even areas of struggle or success (Pardo, Han, & Ellis, 2017). Such granularity of data provides invaluable insights, transforming abstract learning processes into quantifiable, visual metrics. The benefit of these digital analytics lies not merely in the data they capture but in their profound potential to facilitate self-regulated learning (Winne & Hadwin, 1998). By making the invisible visible, learners are granted a mirror to their learning behaviors, habits, and tendencies. Visualization of one’s progress empowers learners to take stock of their current position, assess the efficacy of their approaches, and make timely recalibrations.
The underpinning philosophy of these tools aligns closely with the principles of self-regulated learning. At its core, self-regulated learning champions the learner’s autonomy, emphasizing proactive planning, goal setting, monitoring, and reflection (Winne & Hadwin, 1998). When students have access to real-time data about their progress, these aspects of self-regulation are significantly amplified. For instance, if a student notices consistent challenges in a particular module through the analytics dashboard, they might allocate more time, employ different resources, or seek help adapting their strategy based on data-driven insights. This dynamic adaptation process, facilitated by digital tools, intertwines seamlessly with the evaluative judgement process (Tai et al., 2018). Evaluative judgement, in essence, is the learner’s ability to discern quality, not just in external work, but critically in their own work. The reflection induced by analytics and progress trackers aligns with the reflective component of evaluative judgement, where learners actively interrogate their learning strategies, question their efficacy, and make informed decisions on future approaches (Boud et al., 2018).

In a sense, these modern digital platforms are bridging the gap between the theoretical constructs of self-regulated learning and evaluative judgement, making them tangible, actionable, and integrative. As educators and learners harness the capabilities of these tools, the learning process is reshaped, emphasizing active reflection, strategic adaptability, and a deepened commitment to quality assessment and improvement (Broadbent & Poon, 2015).

Modern e-learning environments, particularly those that promote user-generated content, underline the significance of collective intelligence (Tai et al., 2018). Through platforms such as wikis, collaborative tools, and discussion boards, learners are not just passive consumers but active contributors. Participative platforms amplify the scope of evaluative judgement as learners critically assess contributions from a broader community. This democratized, community-driven
learning ethos champions a more holistic sense of evaluative judgement, underscoring the value of shared knowledge and collaborative critique. Despite the allure and potential of technology-enabled learning environments, the essence of the human element in learning remains paramount (Chong, 2021). Particularly in the realm of evaluative judgement, the emotional and social dimensions play pivotal roles. Insights into the affective states, motivations, and possible reservations of learners as they engage in evaluative tasks can shed light on their internal judgement processes. As the digital domain continues to evolve, integrating components that address both cognitive and affective facets will be integral to ensuring comprehensive development.

As the trail of technological advancements points towards the integration of sophisticated tools like neural interfaces and quantum computing in education, the canvas of technology-enabled learning will morph and expand (Boud et al., 2018). Yet, foundational principles, particularly those centered on evaluative judgement, must be unwavering. Rooted in tenets like critical reflection and calibration against established benchmarks, evaluative judgement will remain an indispensable skill. The onus will be on educators and technologists to ensure that these principles are seamlessly woven into the fabric of futuristic learning environments, keeping the learner’s holistic development at the forefront. Looking ahead, as technology continues to evolve, so will the nature and capabilities of digital learning environments. The integration of artificial intelligence, augmented reality, and immersive virtual spaces promises even richer platforms for learning. Ensuring that evaluative judgement remains a central tenet of learning experiences within these future spaces will be crucial.

Several instances of technologies that have shown promise in the development of evaluative judgment include ALEKS (Baker, 2007), Calibrated Peer Review (CPR) (Borrego et al., 2010), Peerceptiv (Cho & Schunn, 2007), SNAPP (Knight et al., 2014), and SMART (Kim et al.,
ALEKS, which stands for Assessment and Learning in Knowledge Spaces, is an adaptive web-based learning platform. It utilizes artificial intelligence to map the details of each student’s knowledge, tailoring content to individual learning needs. As students engage with the system, they receive immediate feedback on their performance. This constant cycle of action, evaluation, and feedback fosters students’ ability to make judgments about their learning and areas of improvement.

Calibrated Peer Review (CPR) and Peerceptiv are platforms that place strong emphasis on the value of peer assessment. Both systems allow students to evaluate and grade their peers’ work against predefined criteria. This engagement in assessment processes enables learners to deeply understand and internalize quality standards, enhancing their evaluative judgment over time. Through assessing others’ works and receiving feedback on their own, students not only fine-tune their understanding of the subject but also cultivate the skill of evaluating work quality in a collaborative learning environment. SNAPP, or Social Networks Adapting Pedagogical Practice, offers a unique approach. Instead of focusing on content per se, it visualizes the social interactions within online learning environments. This allows students and educators to assess the quality, depth, and breadth of their online engagements, giving insights into collaborative dynamics and individual participation. Recognizing the importance of social interaction in learning, being able to judge and refine one’s engagement based on these visual insights can be a vital part of enhancing evaluative judgment in collaborative learning scenarios.

Lastly, the SMART platform is designed with an emphasis on real-time feedback, providing learners with instant insights into their performance. The immediacy of feedback challenges learners to continually refine their assessment capabilities, drawing them into an ongoing dialogue with their work. As they receive and act on immediate feedback, learners are more likely
to fine-tune their skills in making informed judgments about their learning process and its outcomes. Collectively, these platforms illuminate the importance of iterative feedback, peer collaboration, and self-reflection in honing the evaluative judgment of learners.

In general, technology-enabled learning environments present both unprecedented opportunities and challenges in cultivating evaluative judgement. A more complete understanding, combined with strategic pedagogical interventions, can ensure that students are not only consumers of digital content but also astute judges of quality, well-equipped for lifelong learning in an increasingly digital world. Despite the advancements and insights gleaned so far, our understanding of how evaluative judgement competencies unfold in technology-enhanced environments versus traditional non-technology settings remains incomplete. The intricacies of how students cultivate, refine, and deploy evaluative skills in these contrasting contexts have yet to be fully explored. Given the increasing integration of technology in education and its potential influence on students’ learning processes, the need for rigorous studies examining how these environments either facilitate or impede the development of evaluative judgement is of paramount significance. Such investigations will not only bridge the knowledge gap but will also guide pedagogical strategies to ensure holistic learner development in both digital and conventional settings.

**Evaluative Judgment and Self-assessment**

One of the foundational premises of contemporary education is the active participation of students in their learning process. Self-assessment, as Boud (1995) articulated, offers an avenue for students to actively engage in reflecting upon and evaluating their own work in comparison to predefined criteria or standards. This evaluative process not only fosters metacognitive awareness but also amplifies the capacity for critical reflection. The development of such capacities is crucial for evaluative judgment, a skill necessary for discerning the quality of one’s own work.
and the work of peers. Evaluative judgment, as elucidated by Sadler (2010), entails a complex appraisal process. At its core, it shifts the spotlight from merely receiving feedback on one’s work to a deeper understanding of the underlying criteria that define quality. Through self-assessment, students initiate this very shift, transitioning from passive recipients of feedback to active participants in the judgment process. They seek to discern what constitutes quality work within a given context, making informed decisions based on their evaluations.

Topping (2009) and Taras (2009) both emphasize the crucial role self-assessment plays in honing students’ critical thinking abilities. Beyond the mere evaluation of one’s work, it nurtures a heightened self-awareness and an intrinsic motivation for continuous improvement. Engaging in self-assessment challenges students to juxtapose their work against benchmarks, thus refining their evaluative judgment skills. When students assess their own work, they are essentially developing a blueprint for quality, which they can use in subsequent tasks. Likewise, Hattie and Timperley (2007) shed light on the transformative potential of self-assessment in catalyzing student achievement. By endorsing metacognitive practices and self-directed learning, students who engage in self-assessment are better poised to diagnose their strengths and areas requiring enhancement. Such diagnostic capabilities are instrumental in nurturing evaluative judgment, enabling students to discern the details that separate exemplary work from the mediocre. Grounding this discussion in the sociocultural lens of Lave and Wenger (1991), self-assessment can be perceived as an apprenticeship into the community of practice. Through iterative cycles of assessment, feedback, and reflection, students become acclimatized to the norms, criteria, and standards of their educational community. By actively participating in this evaluative cycle, they are essentially cultivating their evaluative judgment, drawing parallels between their work and the established standards of quality.
Within the broader landscape of assessment, the transformative potential of formative assessment has been highlighted by scholars like Black & Wiliam (2009). Formative assessment provides an ongoing feedback loop, offering students insights into their progress. When embedded within formative assessment practices, self-assessment acts as a catalyst for developing evaluative judgment. By engaging in continuous self-assessment, students become adept at calibrating their work against set benchmarks, a skill integral to evaluative judgment. While summative assessments serve to evaluate students at the end of a learning period, formative assessments provide continuous, iterative feedback, enabling students to adjust their learning strategies in real-time. As Popham (2008) underscores, the primary objective of formative assessment is not to grade students, but to provide feedback to both students and instructors to improve the ongoing learning process. Boud (1995) emphasized the significant role of self-assessment within the formative assessment framework. He argued that when students are involved in evaluating their work, they shift from being passive recipients of feedback to active agents in their learning journey. This process, where learners appraise their performance against certain criteria, is not only a reflection of their understanding of the content but also an exercise in evaluative judgment. By consistently assessing their work, they develop a sense of the standards expected which, as Tai et al. (2018) suggests, is crucial for refining evaluative judgment.

Nicol and Macfarlane-Dick (2006) delve deeper into the relationship between formative assessment, self-assessment, and evaluative judgment. They present a view where effective feedback, intrinsic to formative assessment, should clarify what good performance looks like (the standards or criteria) and generate information that can be used by students to close the gap between current and desired performance. Within this framework, self-assessment acts as a bridge, facilitating students’ understanding of the feedback and their role in utilizing it to make informed
judgments about the quality of their work. Carless et al. (2011) further emphasizes the need for dialogic feedback within formative assessment practices. Such feedback isn’t just about the transmission of information but involves a two-way dialogue between the educator and the student. This dialogue becomes even more enriched when students are actively engaged in self-assessment. As they communicate their own understanding and perceptions of their work, it fosters their ability to articulate and refine their evaluative judgment, enhancing the depth of their understanding of quality.

Moreover, the collaborative aspect of formative assessment, where peers often review each other’s work, has been pointed out by Topping (2009) as an avenue that can further bolster evaluative judgment. As students navigate the work of their peers, the boundaries of self-assessment extend. They begin to judge not just their work but also evaluate others against the same set of criteria, leading to a richer, more holistic cultivation of evaluative judgment. Overall, the feedback loop inherent in formative assessment, when combined with the reflective nature of self-assessment, provides students with an environment ripe for developing and honing their evaluative judgment (Panadero & Broadbent, 2018). As our understanding of the role of assessment in education continues to evolve, it becomes crucial to harness these synergies for an enriched learning experience.

Wu (2020) offers empirical weight to the discussion through a meta-analysis exploring the benefits of self-assessment. The findings illuminate that self-assessment, underpinned by clear criteria and constructive feedback, significantly augments student learning. Moreover, the act of self-assessment amplifies intrinsic motivation, steering students towards more engaged and profound learning experiences. Such heightened engagement and reflection serve as cornerstones for the development of evaluative judgment.
Winne & Hadwin’s (1998) model of self-regulated learning offers a detailed perspective on the processes learners engage in as they manage and direct their learning. This model emphasizes the recursive nature of self-regulation, which is segmented into four key phases: task definition, goal setting and planning, enacting strategies, and adaptation. Within this framework, self-assessment emerges as a vital component, intricately woven into each phase, enabling learners to continually monitor and adjust their learning processes.

In the task definition phase, students define the nature of the task at hand, understanding its requirements and their prior knowledge related to the task. Here, self-assessment plays a crucial role, as learners gauge their existing knowledge base and skills, assessing where they stand and what they need to learn. They evaluate their prior experiences, preconceptions, and potential misconceptions, which sets the stage for the next phases (Winne & Hadwin, 1998).

During the goal setting and planning phase, students, equipped with their self-assessment from the task definition, set specific, measurable goals and strategize how to achieve them. The efficacy of this phase hinges on accurate self-assessment. For instance, learners who can authentically assess their competencies might set more realistic goals, ensuring that they neither overstretch nor undersell their capabilities. They decide on strategies and resources based on their self-assessed needs and strengths.

As learners move to the enactment of strategies, self-assessment continues to be pivotal. As they engage with the learning material and deploy strategies, they continually assess their comprehension, skill acquisition, and progress towards their goals. If a particular approach isn’t yielding the desired results, it’s through self-assessment that learners realize the need for change. Winne & Hadwin (1998) emphasized the importance of feedback in this phase, where learners evaluate feedback against their performances, further deepening their self-assessment practices.
Finally, in the adaptation phase, learners reflect on their entire learning processes, assessing what worked, what didn’t, and why. They use this holistic self-assessment to refine their approaches for future tasks, ensuring continuous improvement. In essence, the self-regulated learning model proposed by Winne & Hadwin (1998) underscores the indelible role of self-assessment throughout the learning process, from task comprehension to reflection and adaptation. Through regular, authentic self-assessment, learners can enhance their self-regulation, making their learning more effective and adaptive. This dynamic, cyclical process underscores the intrinsic link between self-assessment and evaluative judgment. When students assess their work, they’re not just identifying gaps in their understanding or execution; they’re making evaluative judgments about the quality of their work vis-a-vis the set criteria. Over time, as they engage in repeated cycles of self-assessment, their evaluative judgment skills become sharper. They develop an acute understanding of what constitutes ‘quality’ and can apply these standards not just to their work but also in varied contexts, displaying the transferability of these skills (Pintrich, 2000).

Similarly, Andrade (2010) also emphasizes the iterative nature of self-assessment in the learning process. As learners repeatedly engage with tasks, evaluate their performance, and adjust their strategies based on these evaluations, they not only enhance their understanding of the subject matter but also refine their criteria for quality. This iterative engagement aids in calibrating their internal benchmarks, aligning them more closely with external standards. As a result, their evaluative judgment becomes more precise, enabling them to discern subtle gradations in quality. Another aspect of self-assessment and its role in self-regulated learning is its interplay with feedback (Butler & Winne, 1995). Feedback, both internal (arising from self-assessment) and external (from teachers or peers), offers learners insights into the quality of their
performance. As students internalize this feedback and integrate it with their self-assessments, they refine their evaluative judgment, reconciling their perceptions of quality with external benchmarks. This process of reconciliation and alignment, as posited by Nicol and Macfarlane-Dick (2006), is central to the maturation of evaluative judgment skills. By continuously engaging with their work through the lens of self-assessment, learners don’t just improve their academic performance; they cultivate a discerning evaluative judgment, equipping them with skills that hold value beyond academic contexts.

On the other hand, Carless et al. (2011) proposes a paradigm shift in assessment practices, championing student agency and empowerment in learning and development. They argue that for evaluative judgment to be truly embedded in educational practices, students need to transition from mere recipients to active agents. Self-assessment offers a conduit for such an agency, enabling students to negotiate standards, challenge norms, and redefine quality benchmarks. Similarly, to bolster the agency, Boud, Ajjawi, and Dawson (2018) underscore the significance of authentic assessment tasks in the cultivation of evaluative judgment. They argue that when students engage in real-world tasks, they are better poised to make more accurate evaluations. Self-assessment within such authentic tasks immerses students in complex evaluative scenarios, compelling them to make informed judgments based on multifaceted criteria.

To summarize, the interplay between self-assessment and evaluative judgment is intricate and reciprocal, with the former serving as a foundational element for cultivating the latter within the broader context of self-regulated learning framework (Winne & Hadwin, 1998). It is believed that through self-assessment, students engage in a metacognitive process, evaluating their understanding against set criteria, enhancing their capability to make informed judgments about the quality of their work. Pioneering theorists such as Boud (1995) and Winne & Hadwin (1998)
have emphasized the centrality of this process in fostering deeper learning and improved academic outcomes. The emphasis on students as active agents in their own learning process underscores the importance of self-assessment in molding evaluative judgment.

However, despite the extensive research on self-assessment and its impact on evaluative judgment, there remains a significant gap in our understanding of its intersection with technology, especially in the realm of technology-enabled feedback systems. As educational paradigms become increasingly digital, it is crucial to examine how technology intersects with self-assessment practices. Does the introduction of technology into the assessment process augment or detract from its effectiveness? How do technology-enabled feedback settings compare with non-technological or traditional feedback mechanisms in terms of fostering self-assessment and evaluative judgment? These questions highlight an area that warrants further exploration. Presently, while the benefits and mechanisms of self-assessment are well-documented, we lack comprehensive insights into its dynamics within technology-infused contexts. Determining whether technology facilitates or hinders the process of self-assessment and its influence on the development of evaluative judgment is essential for shaping future pedagogical strategies and technological integrations in education.

**Evaluative Judgment and Peer-Assessment**

The reciprocal relationship between peer assessment and evaluative judgment has increasingly garnered attention in educational research (Bouwer et al., 2018; Cowan, 2010; Iglesias Pérez et al., 2022; Zhan, 2021). Boud et al. (2016) have consistently advocated for the vital importance of this skill, emphasizing its pertinence beyond academic confines and into broader life contexts. This sentiment is also mirrored in the works of Martin and Clarke (2017), who suggest that in the complex and collaborative global workspace of today, the ability to evaluate and
provide actionable feedback has become invaluable. In fact, both self-assessment and peer assessment serve as embodiment of evaluative judgment. Within these processes, students actively participate in feedback generation, cultivating a dual awareness of both assessor and assessees roles. This awareness, according to Norton and Hathaway (2012), allows for a more profound internalization of standard benchmarks, consequently sharpening students’ meta-cognitive abilities.

Anchoring peer feedback within the context of formative assessment, Black and Wiliam (1998) have consistently emphasized formative assessment not merely as a tool for grading, but rather as a mechanism for providing constructive feedback. Peer assessment amplifies this ethos. As students participate in this process, they don’t merely passively receive feedback; they actively shape it. Chang & Tseng (2023) argue that the act of generating feedback can often be a more enriching learning experience than simply being at the receiving end. Peer assessment also transcends cognitive benefits, offering a plethora of socio-emotional advantages. Engaging in the evaluation of peers’ work fosters empathy, bolsters communication acumen, and ingrains the value of constructive criticism. A sentiment echoed by Alemdag & Yildirim (2022) and expanded upon by Panadero et al. (2023), who shed light on the pivotal role peer assessment plays in nurturing essential interpersonal skills in today’s collaborative workspaces.

In the context of peer assessment, the essential role of educators stands out as both foundational and transformative. While students remain actively engaged in the core processes of evaluating and assimilating feedback, it is educators who sculpt the overarching framework, ensuring that the environment is conducive to constructive dialogue and holistic learning. This perspective resonates with the findings of Sluijsmans & Prins (2006) who asserted that educators play a cardinal role in the orchestration of effective peer assessment experiences. In this scheme,
educators don’t just act as facilitators; they serve as both guardians of the process and enablers of its evolution. They set clear parameters, define discernible benchmarks, and actively foster a culture where feedback becomes a tool for mutual growth rather than mere critique (Shortland, 2010). This transformative role of educators underscores their influence in elevating peer assessment from a standardized evaluative task to a dynamic pedagogical instrument.

Spiller (2012) delves deeper into the multifaceted responsibilities that educator’s shoulders in this context. They emphasize that clarity forms the cornerstone of effective peer assessment. By meticulously crafting guidelines and ensuring their dissemination, educators can eliminate ambiguities, ensuring that students are aligned in their understanding and approach. Such clarity, as highlighted by Lu & Chiu (2021), not only streamlines the process but also enhances its overall efficacy.

Beyond guidelines, training emerges as a seminal aspect of the educators’ role. The pedagogical process involves equipping students with the requisite skills to both give and receive feedback effectively. Through structured training sessions, workshops, and ongoing dialogue, educators can amplify the depth and relevance of the feedback exchanged, thus optimizing the learning outcomes of the process (Panadero et al., 2023; Topping, 2010).

However, the continuous support educators offer is what often becomes the bedrock of the peer assessment process. Navigating the intricate pathways of peer feedback can sometimes be overwhelming for students, laden with uncertainties and potential biases. Alqassab et al. (2023) advocate for the indispensable role educators play in these moments, offering clarity, mediating disputes, and ensuring the sustenance of a constructive environment. Similar to self-assessment, peer-assessment also closely aligns with the foundational principles of self-regulated learning. At its core, self-regulated learning encapsulates the idea of students taking charge of
their own learning, actively monitoring, controlling, and directing their cognitive processes to achieve academic goals. Zimmerman’s (2000) model of self-regulated learning delineates a cyclic process comprising three phases: forethought, performance, and self-reflection. In this continuum, peer-assessment becomes particularly impactful during the self-reflection phase. When students engage in peer-assessment, they’re not just evaluating others’ work but are also inadvertently reflecting on their own understanding and capabilities, comparing, and contrasting their knowledge with that of their peers, thereby enhancing their self-evaluative insights and metacognitive awareness.

Winne and Hadwin’s (1998) model of self-regulated learning provides another lens through which the relevance of peer-assessment can be viewed. Their model is characterized by a sequence of events: task definition, goal setting and planning, enacting study tactics and strategies, and metacognitively adapting the strategies. Within this framework, peer-assessment can play a pivotal role in the ‘enacting study tactics and strategies’ phase. As students evaluate their peers’ work, they encounter varied approaches to a task, exposing them to a multitude of strategies and tactics. This exposure not only broadens their understanding but also aids them in refining their own study strategies based on observed best practices. Furthermore, the feedback they receive during peer-assessment aids in the ‘metacognitively adapting studying’ phase, providing them with insights into areas of improvement and guiding their subsequent learning process.

Incorporating peer-assessment within the paradigms of Zimmerman’s (2000) and Winne and Hadwin’s (1998) models elucidates its value in fostering a deeper, more introspective form of learning. When students become both the assessors and the assessed, they immerse themselves in a holistic evaluative process, which, in turn, accentuates their self-regulatory behaviors. The feedback loop, inherent in peer-assessment, provides continuous checkpoints for students to
recalibrate their learning strategies, aligning them more closely with desired outcomes. This synergy between peer-assessment and self-regulated learning, as conceptualized by Zimmerman (2000) and Winne and Hadwin (1998), underscores the potential of peer-assessment to act as a catalyst in fostering independent, proactive, and reflective learners.

In addition, the confluence of peer-assessment and technology (Badea & Popescu, 2022; Lin & Yu, 2023; Shroff et al., 2023; Tai & Adachi, 2020) has paved the way for transformative learning experiences in contemporary education. The digital age has witnessed a proliferation of tools and platforms specifically designed to facilitate and enhance the process of peer assessment. These technological tools, ranging from dedicated software to integrated Learning Management System (LMS) features, bring a level of efficiency, scalability, and consistency to the peer-assessment process that was previously challenging in traditional classroom settings. For instance, platforms like Turnitin’s PeerMark (Li, 2018) or Moodle’s Workshop Module (Chaparro-Peláez et al., 2019) allow educators to seamlessly integrate peer review assignments, manage peer pairings, ensure anonymity, and provide structured feedback rubrics. Systematized processes, as highlighted by Jongsma et al. (2023), can help eliminate biases, ensure fairness, and provide a structured framework for students to provide and receive feedback.

Beyond mere efficiency, technology enhances the qualitative aspects of peer-assessment. With features such as real-time collaborative annotations (Chen et al., 2023), multimedia feedback integrations (Yang, 2022), and AI-driven insights (Darvishi et al., 2022) the depth and richness of feedback are significantly augmented. These tools allow students to go beyond textual feedback, incorporating voice notes, video feedback, or even real-time collaborative discussions. According to a study by Li et al. (2012), multimedia feedback in peer-assessment was found to be more engaging and impactful for learners compared to traditional text-based feedback.
Additionally, the data analytics capabilities inherent in these platforms provide educators and students with insights into assessment patterns, strengths, weaknesses, and areas of improvement. Such analytics-driven insights are invaluable for continuous improvement and personalized learning pathways (Badea & Popescu, 2022; Darvishi et al., 2022; Misiejuk & Wasson, 2023).

Additionally, the integration of technology in peer-assessment also aligns with the broader pedagogical shift towards blended and online learning environments. As higher education institutions increasingly embrace hybrid teaching models, the role of technology in facilitating peer interactions becomes paramount. In these virtual spaces, peer-assessment tools not only replicate but often enhance the collaborative and evaluative processes found in physical classrooms. The asynchronous nature of many of these tools allows for flexibility, giving students the time and space to reflect deeply and craft meaningful feedback. Furthermore, as haparro-Peláez et al. (2019) point out, the digital documentation of peer feedback provides a lasting resource for students to revisit and reflect upon, fostering long-term learning and retention. In fact, the integration of technology in peer-assessment has not just digitized a traditional process but has elevated it, making it more efficient, impactful, and aligned with the evolving dynamics of modern education.

To summarize, the interplay between peer-assessment and evaluative judgment has been a focal point in contemporary educational research (Adachi et al., 2018; Ajjawi et al., 2018; Sridharan, 2019; To & Panadero, 2019; Zhan, 2021) accentuating the significance of peer-assessment in amplifying students’ aptitude to critically appraise the caliber of academic work, either their own or their peers’. Notwithstanding this acknowledgment, a discernible gap exists in our understanding of the intricacies of peer-assessment as a manifestation of evaluative judgment,
particularly within the setting of technology-driven educational systems. Evaluative judgment, as delineated by scholars such as Boud et al. (2021), is a complex ability that goes beyond mere assessment activities. Evaluative judgment involves deep understanding, self-reflection, and the ability to make informed decisions about the quality of work. While peer-assessment clearly engages students in evaluation tasks, it’s still unclear how fully these tasks help students develop their evaluative judgment. As education increasingly integrates digital tools and platforms, understanding this relationship becomes even more vital. Sadler (2014) point out that technology-enabled learning environments add new layers to the peer-assessment process, which might change how evaluative judgment is developed. Despite the broad discussions on both peer-assessment and evaluative judgment, more research is needed on their intersection in the digital age.

**Evaluative Judgment and Individual Differences**

Evaluative judgment is inherently complex and inherently personalized, making the role of individual differences in its development an intriguing dimension (Birney et al., 2016; Chen, 2003). The complexity arises not merely from its definitional scope but also from the myriad individual factors that influence its development. Evaluative judgment, at its core, is an intricate confluence of cognitive, metacognitive, emotional, and social capacities (Clore & Parrott, 1994; Petty & Brinol, 2006; Proust, 2010) which enables learners to critically assess and appraise both their work and that of their peers. Boud, Soler, and Falchikov (2016) were among the earliest proponents emphasizing the complexity of evaluative judgment, describing it as a synthesis of multifaceted cognitive, metacognitive, social, and affective processes. As they pointed out, the development of evaluative judgment is less a linear progression and more an interwoven process shaped by personal attributes. Expanding on this, Brown, and Harris (2013) showcased that
learners’ personal attributes profoundly influence their evaluative experiences, transforming the exercise from a mere academic task to a deeply introspective endeavor.

Diving into the cognitive dimension, one’s cognitive capabilities become foundational in the evaluative judgment matrix. Cognitive capabilities span a spectrum from knowledge retention to analytical skills. As evaluative judgment mandates the discerning appraisal of work, cognitive capabilities prove indispensable in ensuring that these appraisals are rooted in sound reasoning and systematic analysis (Hebert & Vorauer, 2003; Phakiti, 2016). As Perry and Winne (2004) elucidate, those possessing advanced cognitive skills often manifest superior judgment capabilities. Their cognitive strengths aren’t just confined to the traditional processes of knowledge assimilation and retention. Instead, they extend into sophisticated competencies like analytical agility, critical discernment, and the ability to weave disparate informational threads into coherent, actionable insights.

These capabilities, however, don’t function in isolation. They are often interwoven with other higher-order cognitive functions such as synthesis, application, and evaluation, aspects that Bloom (1956) highlighted in his taxonomy. For instance, a learner equipped with refined cognitive capabilities doesn’t merely analyze feedback in its rudimentary form. They engage in synthesis, deciphering the latent themes and underlying patterns in the feedback, leading to more nuanced judgments. This synthesis further facilitates the subsequent processes of application and evaluation. Such learners, as Anderson and Boud (1996) emphasized, showcase an adeptness in pragmatically integrating feedback into their work. Their evaluative judgments are not cursory glances but deep dives into the quality of work, buoyed by their cognitive scaffolding.

Furthermore, the interplay of cognitive capabilities and evaluative judgment isn’t a unidirectional phenomenon. As Nicol and Macfarlane-Dick (2006) posited, the act of engaging in
evaluative judgment itself serves as a cognitive stimulant. As learners oscillate between assessing their work and assimilating feedback, they are inadvertently sharpening their cognitive faculties. This cyclical process of evaluation, reflection, and iteration fosters cognitive growth, thereby further amplifying the quality of future judgments (Cubero-Ibáñez et al., 2018; Malecka & Boud, 2021; Villarroel et al., 2018).

In addition to the cognitive capabilities, metacognition, often referred to as "thinking about thinking," plays an instrumental role in shaping and refining evaluative judgment. Grounded in the pioneering research of Zimmerman (2000), metacognition’s intersection with self-regulated learning showcases how understanding one’s own cognitive processes can significantly impact one’s ability to assess and judge. Essentially, learners equipped with heightened metacognitive capabilities exhibit an amplified acuity in their evaluative performance (Handel et al., 2013; Song et al., 2011). They don’t just passively ingest information; they actively interrogate it, weighing its validity, relevance, and implications. Such individuals can perceptively pinpoint knowledge deficits, thoughtfully seek feedback, and, with an almost reflexive agility, pivot their learning approaches based on their evaluations.

Furthermore, metacognition is intricately woven into the fabric of active learning (Khosravi et al., 2021). The active solicitation of feedback is not merely a function of external validation but emerges as a testament to a learner’s metacognitive awareness. Being conscious of one’s strengths, weaknesses, and areas of improvement facilitates a deeper engagement with the learning content, driving individuals to continually refine their knowledge structures. As Schraw and Dennison (1994) articulated, there’s an undeniable synergy between robust metacognitive awareness and enhanced academic achievement. This relationship is not mere coincidence but a product of learners’ heightened ability to assess, reflect, and act upon their judgments.
The magnified importance of metacognition in evaluative judgment further reverberates in the works of Pintrich (2002), who delved into the intricacies of metacognitive strategies in academic settings. He asserted that learners who actively deploy metacognitive strategies—like monitoring their comprehension, regulating their pace, or adapting strategies based on feedback—not only fare better academically but also demonstrate nuanced evaluative judgment capabilities. Their evaluations are informed, deliberate, and contextually anchored, reflecting a rich interplay of cognitive and metacognitive processes.

In synthesizing these insights, it’s evident that metacognition is not a peripheral factor in the realm of evaluative judgment. It is central, dynamic, and influential. Cultivating metacognitive skills is not just a pedagogical imperative but a conduit to amplifying the depth, breadth, and precision of evaluative judgments. As educational paradigms evolve, the accentuated focus on metacognition will undoubtedly serve as a catalyst for fostering discerning, self-aware learners capable of accurate evaluative judgments.

However, to assume that evaluative judgment rests purely on cognitive scaffolds would be an oversimplification. The emotional and social contours of learners play a role. Emotionally, factors such as learners’ receptivity to feedback, resilience in the face of criticism, and even their intrinsic motivation levels influence their evaluative efficacy (Carless & Winstone, 2023). Falchikov (2013) lucidly highlighted the spectrum of emotional variables, from anxiety to confidence, shaping evaluative processes. Dweck (2000) has explored the mindset of learners and elucidated that those harboring a growth mindset—believing that their abilities can be developed through dedication and hard work—are more receptive to feedback, displaying a higher propensity to embrace it as an opportunity for refinement rather than as a critique. Such learners, underpinned by their intrinsic motivation, perceive evaluative feedback as a catalyst for growth rather
than a marker of inadequacy. However, on the other end of the emotional spectrum lie variables such as fear of judgment, apprehension, and even defensiveness. Learners grappling with such emotions might eschew feedback, deeming it as threatening or demeaning (Yang, 2021). As Carless and Boud (2018) have articulated, the emotional equilibrium of learners significantly modulates their evaluative processes. Emotions can either pave the way for constructive engagement or create barriers to effective evaluative judgment. Thus, the emotional well-being and stability of learners emerge as non-negotiable facets in cultivating effective evaluative judgment skills. These emotional elements can serve as amplifiers or diminishers in the evaluative process.

Concurrently, social capabilities, which encompass learners’ communication skills and interpersonal dynamics, emerge as significant levers in peer assessment and evaluative dynamics. Topping (1998) captured the essence of effective communication in peer assessments, elucidating its centrality in ensuring evaluations are clear, germane, and actionable. Learners’ adeptness in articulating feedback, their sensitivity to peers’ perspectives, and their ability to engage in constructive dialogues significantly determine the efficacy of evaluative processes. Nicol, Thomson, and Breslin (2014) have emphasized that the social aspect of evaluative judgment isn’t just about giving feedback but also about fostering a mutual understanding between the assessors and assessee. The core lies in the symmetry of understanding, where feedback isn’t just dispatched but is collaboratively discussed, dissected, and deliberated upon.

In summary, it is paramount to understand that evaluative judgment, while theoretically articulated with clarity, manifests uniquely across learners. The confluence of cognitive, metacognitive, emotional, and social dimensions uniquely sculpts each learner’s evaluative judgment. As contemporary pedagogy evolves, appreciating, and calibrating for these multifarious individual differences, as underscored by Boud et al. (2019), remains indispensable for nurturing
holistic evaluative proficiencies. In the current study, we focus specifically on three individual differences factors of prior knowledge, reading comprehension, and vocabulary knowledge to study their effect on modulating learners’ evaluative judgment in self-assessment versus peer-assessment activities and its interaction with performance in technology versus non-technology settings.

**Conceptual Framework**

Evaluative judgment, a sophisticated learning self-regulation ability, is molded by interaction of learner’s individual differences and task internal and external characteristics. For learners’ characteristics, evaluative judgment engages a synthesis of cognitive, metacognitive, emotional, and social capabilities (Clore & Huntsinger, 2007; Guo et al., 2022; Rakovic et al., 2022; Carless & Winstone, 2023). During evaluative judgment, learners use a fusion of their internal cognitive and metacognitive capacities (Flavell, 1979) in interaction with their emotional (Damasio, 1994) and social resources (Vygotsky, 1978) to rigorously assess both their own and their peers’ learning processes and the outcomes (Black & Wiliam, 1998). Similarly, with regard to task characteristics, various components play a significant role in how learners evaluate their own and their peers’ performance. These components include well-defined guidelines (Mayer & Moreno, 2003), clear performance standards, and explicit expectations (Norman, 2013) among others. Additionally, the complexity and novelty of the task (Kalyuga, 2007; Sweller, 1988; Tabari et al., 2023), its relevance (Renkl, 2014), the necessity for coordination or multitasking (Ophir, Nass, & Wagner, 2009), specific timing demands (Edland & Svenson, 1993), and the presence or absence of technological support (Panadero et al., 2019; Roscoe and Craig, 2022; Winne, 2011, Yan et al., 2022) also factor into these evaluations.
For the purpose of the present study, we propose a conceptual framework supported by the self-regulated learning (SRL) model presented by Winne and Hadwin (1998) illustrating the process that learners undergo while utilizing their cognitive, metacognitive, emotional, and social capabilities to make sense of their own performance and of that of their peers during evaluative judgment. Refocusing Winne’ and Hadwin’s self-regulated learning model (1998) we argue that for evaluative judgment to be delivered its learning self-regulation and coregulation capacities, learners draw from a variety of learner internal and engage a variety of task-related characteristics.

A review of the literature on learner self-regulation processes highlights a variety of individual differences factors falling into four categories of cognitive, metacognitive, emotional, and social dimensions. In what follows, we review literature on capabilities in each category that impact learners’ evaluative judgment in practice. We start with foundations of the self-regulated learning theory (Winne & Hadwin, 1998), then we focus on dimensions of learner’s characteristics influencing evaluative judgment and then we turn to learning environment and task factors including how technological supports and task factors impact learner’s evaluative judgment.


Winne and Hadwin’s (1998) model of Self-Regulated Learning (SRL) offers an intricate portrayal of how learners navigate their learning experiences, with a strong emphasis on the metacognitive processes. At its core, the model delineates learning into four basic phases: task definition, goal setting and planning, learning tactics, and adaptations to metacognition (Winne & Hadwin, 1998). These phases are not linear but interact in a recursive manner, continually revisiting and adjusting based on the learner’s evaluations.
The distinctiveness of Winne and Hadwin’s model lies in introduction of the COPES framework, which represents Conditions, Operations, Products, Evaluations, and Standards that underpin each learning phase. Conditions, both cognitive and task-based, set the grounds, defining the resources and constraints presented to the learner (Winne & Hadwin, 1998). These conditions encompass a myriad of elements, from past learning experiences and beliefs to the specific requirements of a task. Operations refer to the processes involved in handling information during learning, encompassing activities like searching, monitoring, assembling, rehearsing, and translating or SMART (Winne, 2001). While Product refers to the outcome or result of a specific phase in the self-regulated learning process. These products arise from the operations (information manipulation processes) that learners undertake.

In this model, evaluation involves a comparison between the "products" (outcomes of various phases) and the "standards" set by the learner. Through evaluation, a learner determines if the objectives of a particular phase have been met or if further adjustments or efforts are needed. These comparisons are known as cognitive evaluations. If there’s a mismatch or poor alignment between the products and the standards, this might prompt the learner to adjust their learning operations. This could involve refining the product, modifying the conditions (if at all possible) or standards, or both. Moreover, evaluation has a metacognitive layer to it. For instance, if a learner consistently finds that the learning product does not meet the standards, this may initiate metacognitive monitoring. Such monitoring might determine that the learner’s initial assumptions or beliefs about a task (for instance its difficulty level) are inaccurate, prompting them to adjust their goals or strategies accordingly. Therefore, in Winne and Hadwin’s model, evaluation is not just a static comparison but a dynamic process that might recursively feed back
into the learning cycle, influencing both cognitive and metacognitive processes (Winne & Hadwin, 1998).

Given the focus on conditions, operations, products, evaluations, and standards components, it is believed that the Self-Regulated Learning (SRL) model of Winne and Hadwin (1998) offers a fitting framework for exploration of evaluative judgment in a technology-enabled feedback system. Central to this model is the emphasis on standards and evaluations, integral facets of evaluative judgment. As learners establish or adopt standards, they engage in ongoing evaluations, drawing parallels with how students perceive and utilize feedback from technological interfaces (Winne & Hadwin, 1998). Furthermore, the iterative nature of Winne and Hadwin’s model mirrors the feedback processes within technology-enabled feedback systems. Feedback mechanisms in technology platforms often present learners with iterative opportunities to refine their submissions. This recursive element in the SRL model demonstrates the non-linear adaptation of feedback, echoing the iterative improvements students make based on technological feedback (Winne & Hadwin, 1998).

Furthermore, metacognition is paramount in the process of evaluative judgment. The metacognitive underpinnings of Winne and Hadwin’s model (1998) emphasize a learner’s self-awareness is crucial when deciphering and acting upon feedback in technology-enabled feedback systems. By fostering an awareness of their own learning process and biases, learners are better positioned to discern the relevance and applicability of the feedback they receive. A notable advantage of the SRL model in this context is its adaptability to varied conditions. Feedback systems, in their technological embodiment, can significantly differ, ranging from automated feedback loops to peer-based review modules. The Conditions aspect of Winne and Hadwin’s model,
encapsulating both cognitive and task-specific conditions, allows for this model’s application across an array of tech platforms (Winne & Hadwin, 1998).

Furthermore, central to the idea of personalized learning experiences in technology-enabled feedback systems is the concept of control and adaptability. Winne and Hadwin’s emphasis on the learner’s autonomy and control over their learning processes resonates with the adaptability inherent in personalized feedback mechanism. Their model underscores the significance of learners’ autonomy in tailoring their strategies in response to the feedback they receive (Winne & Hadwin, 1998).

In addition, technology serves as a modulator, subtly adjusting the manner in which learners perceive and approach tasks. Technology-enabled learning environments offer learners an array of resources to define and understand tasks more effectively. These technologies allow learners to access a diverse range of perspectives on the same task, clarifying and enriching their initial task comprehension (Jones, 2008; Wu & Yang, 2022). This modulation extends to goal setting, where technology helps scaffold goals through digital feedback, recommendation systems, or adaptive learning algorithms that align with the learner’s current knowledge state and desired outcomes (Baker & Siemens, 2014).

Beyond modulation, technology-enabled learning systems augment certain elements of the SRL cycle. For instance, interactive platforms can intensify feedback mechanisms, enabling real-time corrections and learning enhancements (Wang, 2017). Digital resources can amplify the reach and depth of studying tactics, providing learners with multimedia content, simulations, or collaborative spaces that weren’t previously accessible (Clark & Mayer, 2016).

To summarize, the SRL model illustrates the cyclical process that learners experience as they regulate their own learning. At the heart of this cycle, learners make evaluative judgments
based on various learner and task-related characteristics. In what follows we review the literature on a variety of learner and task characteristics that are believed to influence evaluative judgment.

**Dimensions of Learner’s Characteristics Influencing Evaluative Judgment**

In this section, we will delve into the dimensions of learner’s characteristics that influence evaluative judgment. Specifically, we’ll discuss cognitive, metacognitive, emotional, and social capacities that we believe might influence learners’ evaluative judgment. We will start with cognitive capabilities of working memory, attention, linguistic aptitude, task familiarity, feedback literacy, assessment literacy and prior knowledge. Then, we will discuss metacognitive capabilities of regulation of cognition, calibration, goal setting, strategy evaluation and reflection. For the emotional aspect of evaluative judgment, we focus on self-efficacy, motivation, and interest. Finally, for social dimensions of the evaluative judgment we focus on the impact of learner’s sociocultural background, learning styles and preferences, learner attitude and beliefs and the effect of their previous learning experiences and peer comparison.

**Cognitive Capacities of Evaluative Judgment**

**Working Memory.** Working memory, described as a cognitive system that provides transitory storage and manipulation of the information necessary for complex cognitive tasks (Baddeley, 2003; Cowan, 2022), plays a fundamental role in an individual’s ability to make evaluative judgment. It functions as a mental workspace, holding, processing, and manipulating information for short spans. When engaged in evaluative tasks, learners employ working memory to align their current understanding with established benchmarks, discerning disparities and integrating feedback (Brunfaut et al., 2021; Engle, Tuholski, Laughlin, & Conway, 1999). This is particularly evident in activities where learners must compare, contrast, and synthesize multiple
sources of information, requiring them to hold and shuffle multiple data points in their minds concurrently (Coiro, 2021).

**Attention.** Attention, the selective focus on specific aspects of information while filtering out others, is another cornerstone of cognition (Posner & Petersen, 1990). It governs the depth, intensity, and direction of cognitive engagement (Li & Lajoie, 2022; Shan, 2021). Those with robust attentional control can focus on the intricacies and subtleties of content, ensuring that their evaluative judgments are both thorough and comprehensive. It’s not just about the sheer duration of attention but also the ability to shift focus adaptively in response to task demands (Desimone & Duncan, 1995). In evaluative contexts, this adaptability might help learners to dynamically prioritize information, weighing the importance of various factors in making an informed judgment.

**Linguistic Aptitude.** Linguistic abilities transcend the domain of communication, shaping the cognitive processes by which individuals perceive, interpret, and appraise their environment (Whorf, 1956). Vygotsky (1986) further expounded on this, emphasizing the intrinsic link between language and thought, suggesting that linguistic capabilities directly influence cognitive development and the ways in which individuals interact with their world. Advanced linguistic skills equip learners with the ability to articulate intricate ideas, analyze sophisticated texts, and participate effectively in deep discussions. In the context of higher education, where academic materials often abound in complexity and depth, such proficiencies become paramount (Cummins, 2008; Su & Zou, 2022). Navigating through dense scholarly articles, research publications, and academic debates requires not just comprehension but also the ability to discern, critique, and contribute, all of which are enhanced by robust linguistic ability.
Further emphasizing the link between linguistic abilities and evaluative judgment, strong linguistic capacities pave the way for precise analysis. This precision proves indispensable when students engage in evaluative tasks. It facilitates learners to unpack feedback meticulously, craft well-thought-out reflections, and engage in discourse that helps negotiate varied interpretations (Carroll, 2002). Their ability to glean deeper meanings, contextualize feedback, and articulate their evaluations with clarity leads to richer academic experiences and more refined judgments. Conversely, those grappling with linguistic challenges might find themselves at a disadvantage. They could overlook details, misconstrue the essence of feedback, or struggle to articulate their insights, thereby impinging on the accuracy of their evaluative judgment (Tai et al., 2016). Such limitations underscore the necessity for educators to be discerning. It’s imperative to differentiate between genuine evaluative shortcomings and those stemming from linguistic barriers, ensuring the latter doesn’t overshadow or diminish true understanding (Leung, 2005).

**Task Familiarity.** Task familiarity also plays a focal role in shaping evaluative judgment, intertwined with the learners’ prior experiences and cognitive processes. As learners grapple with tasks, they have encountered before, their immediate evaluations are often grounded in memory and past experiences (Kazemi & Zarei, 2015; Pena & Quinn, 1997; Tabari & Wang, 2022). The presence of familiarity equips learners with a valuable framework, enabling them to contrast their current performance with prior outcomes, enhancing their confidence and perceived proficiency in addressing the task at hand. Yet, while task familiarity can refine evaluative judgment by anchoring it in a context, it is not devoid of potential pitfalls. Biases that emerge from heuristic thinking can cloud judgment (Crosskerry, Singhal, & Mamede, 2013). For example, an overemphasis on past experiences might cause learners to neglect critical details, operate based on outdated assumptions, or even exhibit overconfidence in their judgments.
(Larrick, 2004; Testa et al., 2023). Such misplaced confidence can manifest when learners, due to their limited experiences or exposures, misjudge their capabilities. The potential for such biases becomes even more significant when tasks, though seeming familiar, have slight differences that go unnoticed, leading to flawed evaluations (Brown et al., 2015).

To navigate these challenges, it’s essential to harness strategies that encourage a blend of drawing from past experiences while also fostering fresh analytical approaches to tasks. For instance, prompting students to justify or explain their judgments or even inviting them to juxtapose their evaluations with those of their peers can serve as effective mechanisms to counterbalance potential biases (Larrick, 2004). In addition, instilling reflective practices in the learning journey can empower students to discern and temper the impact of biases, ensuring their evaluative judgments are multifaceted and rooted in both familiarity and critical reasoning (Tai et al., 2016; Hsu et al., 2022).

**Feedback Literacy.** The manner in which learners interact with, interpret, and integrate feedback plays a critical role in molding their evaluative judgment (Carless & Winstone, 2023; Hattie & Timperley, 2007; Nieminen & Carless, 2023). Feedback, when viewed holistically, is a reciprocal dialogue between the educator and the student, purposed to facilitate both academic and personal development (Carless & Boud, 2018). Central to this interplay is the concept of feedback literacy, encompassing the skills, attitudes, and behaviors learners employ when processing and responding to feedback (Carless & Boud, 2018). Feedback literacy doesn’t just shape how feedback is perceived, but it also directly influences the accuracy and depth of a student’s evaluative judgment. Those adept in feedback literacy can sift through feedback, extracting insights that enable them to reflect upon and refine their work with precision (Sutton, 2012). This proactive engagement with feedback not only informs their immediate tasks but also
cultivates a broader competency in evaluative judgment, empowering learners to discern quality and make informed decisions about their learning experiences. However, feedback’s impact transcends cognitive processing. Its reception is intertwined with emotions, beliefs about oneself, and the drive to succeed. For example, a student with a solid sense of self-efficacy might see feedback, even if critical, as a valuable tool for growth. In contrast, another might perceive the same feedback through a lens of self-doubt or defensiveness (Bandura, 1997).

Promoting feedback literacy, therefore, necessitates a holistic approach. Beyond sharpening analytical abilities, it is about cultivating an environment where feedback is embraced as a conduit for learning and growth (Dweck, 2006). For the feedback process to be truly transformative, it should be clear, empathetic, and contextually relevant. Concurrently, learners need to be nurtured to approach feedback with an open mind, resilience, and a deep-seated commitment to self-improvement (Nicol & Macfarlane-Dick, 2006).

Assessment Literacy. Assessment literacy, at its core, refers to an individual’s understanding and proficiency in designing, interpreting, and using assessment results in a manner that optimizes learning (Chan & Luk, 2022; Popham, 2009). It encompasses not just understanding the technical aspects of various assessment methodologies but also the broader implications these results hold for instructional decision-making and learner progression. For students, this literacy empowers them to navigate, interpret, and leverage feedback, aligning their learning approaches accordingly. The influence of robust assessment literacy on students’ learning and performance is well-established (Deneen & Hoo, 2023; Gamachchige & Jackson, 2022; Smith et al., 2013). Equipped with a deep understanding of assessment paradigms, learners can contextualize their feedback, discerning between foundational strengths and areas necessitating improvement. Such literacy promotes a proactive learning environment where feedback becomes a cornerstone for
iterative improvement rather than a mere reflection of current capabilities (Black & Wiliam, 1998; Deneen & Hoo, 2023). Furthermore, it aids students in discerning the intent behind assessments, allowing them to align their learning strategies with the overarching pedagogical objectives.

While assessment literacy provides the lens through which feedback and results are interpreted, evaluative judgment is the critical process wherein learners actively appraise and reflect upon their performance against standards (Sadler, 1989). The two are intertwined in a reciprocal relationship: a proper grasp of assessment frameworks, informed by assessment literacy, enhances the precision and depth of evaluative judgment. Conversely, the practice of evaluative judgment, when undertaken regularly, further refines a learner’s assessment literacy as they become more adept at dissecting feedback and aligning it with learning outcomes (Brookhart, 2011). In fact, the foundational knowledge of assessment literacy paves the way for more sophisticated and reflexive evaluative practices. When students can effectively interpret the metrics and meanings of their assessments, they are better positioned to make informed judgments about their learning, optimizing both their immediate educational endeavors and their long-term academic growth.

**Prior Knowledge.** Prior knowledge functions as the bedrock of understanding and informs learners’ perspectives when they embark on new learning experiences. The depth and breadth of what a learner already knows—acquired through formal education, personal experiences, or independent study—guide their interactions with new content (Ambrose et al., 2010; Bransford, Brown, & Cocking, 2000). A vast reservoir of prior knowledge not only provides a vantage point for contextualizing and absorbing new information but also instills confidence in learners (Tobias & Everson, 2009). When learners can relate new concepts to existing
knowledge, they can better evaluate the relevance, accuracy, and depth of their comprehension (Novak, 2002). Additionally, learners with extensive prior knowledge can more adeptly draw connections between interdisciplinary subjects, enhancing their holistic understanding and critical thinking abilities (Fausan et al., 2021). Evaluative judgment is deeply intertwined with the foundation of prior knowledge. This pre-existing knowledge serves as an architectural blueprint, assisting learners in decoding, analyzing, and contextualizing new information within their existing frameworks (Ambrose et al., 2010). Prior knowledge directs learners through the complex maze of novel concepts and frameworks. When navigating new academic areas of knowledge, the breadth and depth of a learner’s pre-existing understanding becomes pivotal in dictating their engagement with the new content. Such knowledge, cultivated through structured education, experiential learning, or self-driven pursuits, determines how learners perceive, contextualize, and incorporate new information (Vosniadou, 2001).

Evaluative judgment, being the process of appraising one’s understanding, skills, or performance against specified standards, draws heavily upon prior knowledge. The ability of learners to anchor new insights to their foundational understanding not only facilitates comprehension but also positions them optimally to evaluate the depth, context, and precision of their grasp (Kintsch, 1998). This capability to discern interrelations between established and novel concepts, identifying potential gaps or overlaps, is exponentially enhanced with a solid foundation of prior knowledge.

The synergistic relationship between these elements culminates in the robustness of evaluative judgment. For educators, this underscores the importance of recognizing and strategically leveraging the potentials of prior knowledge, setting the stage for enhanced, reflective, and discerning evaluative processes in learners. In a technology-enabled learning environment, prior
knowledge extends beyond mere familiarity with content. It encapsulates both procedural knowledge, encompassing the proficiency in handling technology and harnessing its capabilities to augment learning, and declarative knowledge, which pertains to an understanding of the subject matter being explored within the digital learning platform (Smith & Jones, 2020; Lee & Williams, 2019).

**Metacognitive Capacities of Evaluative Judgment**

Metacognition is central to effective learning and accurate evaluation, underscoring an individual’s insight into their cognitive processes and the strategies they employ during learning. As elucidated by Flavell (1979), metacognition comprises a dual aspect: knowledge about one’s own cognitive processes and the ability to control and regulate these processes. This dual nature facilitates the ability of learners to strategize, monitor, and reflect on their learning process. Zimmerman (1989) postulated that metacognitive awareness plays a vital role in self-regulated learning, allowing learners to be proactive in their learning strategies and adapting as required. Furthermore, Schraw and Dennison (1994) delineated its role in comprehension monitoring, suggesting that learners equipped with robust metacognitive skills can discern gaps in their understanding and make pertinent adjustments. Veenman et al. (2006) also emphasized that metacognitive awareness and skills significantly correlate with learning outcomes, implying their instrumental role in academic success. In conclusion, metacognitive awareness, as articulated by these scholars, is an active, introspective, and adaptive component of learning, enabling learners to navigate their learning and development with greater efficiency and efficacy. In what follows, several metacognitive functions related to evaluative judgment are reviewed.

**Regulation of cognition.** Regulation of cognition is rooted deeply in the process of overseeing and managing one’s cognitive processes, including planning, monitoring, and evaluation
(Brown, 1987). This regulatory process is paramount in the context of evaluative judgment. Firstly, planning is an anticipatory phase where learners delineate their approach to the evaluative task. It involves choosing strategies, allocating resources, and setting intermediate milestones (Zimmerman & Moylan, 2009). For example, when writing an essay, students might plan by brainstorming ideas, outlining the structure, and setting timelines for drafts. Effective planning is akin to creating a roadmap that provides clear direction, allowing learners to navigate complex tasks with structured thought. Next, monitoring is the ongoing inspection of one’s cognitive performance during the execution phase. It allows learners to continually gauge how well they are progressing towards their goals, and whether their chosen strategies are efficacious (Schraw, 1998). An instance of this can be seen in reading comprehension tasks, where students periodically assess their understanding of the material and adjust reading strategies based on their real-time comprehension levels (Tibken et al., 2023; Tighe et al., 2023; Yang, 2002).

Lastly, evaluation entails the post-task analysis, wherein learners assess the outcomes of their efforts and the effectiveness of the strategies employed (Winne & Hadwin, 1998). After completing a project or task, learners might reflect on what went well, what challenges arose, and how they can refine their approach in the future. For example, after a presentation, a student might evaluate their performance in terms of audience engagement, clarity of content, and pacing, leading to insights for improvement in subsequent presentations. In general, the intricate movement through planning, monitoring, and evaluation phases in cognitive regulation ensures that learners are not merely passive recipients of knowledge. Instead, they become active directors of their learning, using evaluative judgment as a compass to steer their cognitive processes and continually refine their strategies in response to challenges and feedback (Pintrich, 2000).
**Calibration.** Calibration is a vital component of metacognitive capacities, focusing primarily on the precision with which individuals gauge their own learning and performance (Azevedo et al., 2010; Cantaert et al., 2022; Pieschl, 2009). The essence of calibration lies in learners’ ability to bridge the gap between their perceived and actual performance. In many educational settings, students often encounter tasks that require self-assessment or prediction of their success, such as estimating the grade they expect to receive on an upcoming exam. Accurate calibration is instrumental at this stage because it determines how well students’ predictions align with their actual outcomes (Bol & Hacker, 2001). For instance, a well-calibrated student would anticipate getting a B on a test and subsequently earn a grade close to that prediction. On the other hand, a poorly calibrated student might predict the same grade but receive a significantly higher or lower mark, indicating a disconnect between perception and reality.

Furthermore, calibration plays a crucial role in self-regulated learning processes. When learners possess strong calibration skills, they can better identify areas of strength and weakness, which informs strategic decisions about where to allocate study time and effort (Dunlosky & Rawson, 2012). Conversely, miscalibrated students might overestimate their understanding, leading to insufficient preparation or, conversely, waste time over-preparing in areas where they are already proficient. Moreover, the continuous feedback loop in the learning environment, wherein students make judgments, act upon them, and subsequently refine those judgments, underscores the importance of calibration (Zimmerman, 2008). By repeatedly aligning perceived with actual performance, learners hone their calibration skills, fostering a more realistic, informed, and efficient approach to learning tasks. Calibration serves as a mirror, reflecting the alignment or discrepancy between learners’ self-perceptions and their tangible achievements. By nurturing this
capacity, learners can navigate their educational experiences with greater awareness, efficacy, and adaptability (Schraw, Kuch, & Gutierrez, 2013).

**Goal Setting and Strategy Evaluation.** The act of setting specific, measurable, attainable, relevant, and time-bound (SMART) goals not only provides direction to the learner but also offers a clear benchmark against which progress can be assessed (Zimmerman, 2002). Such precise and structured goal setting has been repeatedly affirmed in literature for its role in fostering motivation and enhancing task commitment (Chang et al., 2022; Locke & Latham, 2006). The process doesn’t conclude at mere goal setting; rather, it evolves into an iterative cycle of action and reflection. Once goals are established, learners employ various strategies to reach their desired outcomes. Strategy evaluation then becomes crucial, as it ensures that learners are not just engaged in activities, but they’re purposefully advancing towards their goals (Schunk & Ertmer, 2000). This is where metacognitive capacities come into play, as learners not only employ strategies but continually assess their efficacy, tweaking and refining as they proceed (Pintrich, 2000).

Furthermore, the alignment of strategies with set goals is paramount. For instance, a learner might set a goal to master a specific topic within a week. While the goal is clear and time-bound, the strategies employed—be it summarizing, self-testing, or group discussions—need to be evaluated against the progress made. If by mid-week the learner feels the topic remains unclear, the strategy might require adjustment. Without such evaluation, learners might find themselves drifting off course, investing time and effort without meaningful results (Winne & Hadwin, 1998). In the broader context of evaluative judgment, goal setting paired with strategy evaluation forms a dynamic duo. They encourage learners to be active participants in their learning efforts, providing them with the tools to set direction, monitor progress, and make necessary adjustments along the way (Boekaerts, 1999).
Reflection. Reflection, as a metacognitive process, holds immense significance in the learning process. After an assessment, learners delve into a reflective space, reconsidering various facets of their performance, the feedback they received, and the overall assessment process (Merkebu et al., 2023; Schön, 1983; Silver et al., 2023). Such a practice is not limited to the confines of formal education but extends to real-world scenarios too. For instance, professionals in fields like medicine and law frequently engage in reflective practices. According to Epstein (1999), medical practitioners are often encouraged to reflect upon their diagnoses and treatment plans to improve patient outcomes. This reflection allows them to identify potential oversights, broaden their understanding, and adapt their methods, leading to enhanced clinical practices. Similarly, in fields where design plays a focal role, like architecture or product development, reflective practices are essential. Designers frequently review their prototypes, gather feedback, and then refine their designs. Cross (2007) suggests that this reflective practice in design leads to innovative solutions and ensures that the end product effectively addresses user needs. Moreover, reflection has become a cornerstone in professional development programs across industries. Professionals attend workshops, training sessions, and seminars, where they are encouraged to reflect upon their learnings. According to Moon (2004), such reflective activities enable professionals to integrate new knowledge into their existing frameworks, leading to deeper understanding and improved practices. Thus, reflection, whether it’s post-assessment in an academic setting or following real-world tasks and experiences, is a powerful tool for growth and continuous improvement. It fosters an environment of inquiry and curiosity, prompting individuals to continually refine their strategies and approach.
Emotional Capacities of Evaluative Judgment

Emotions, frequently considered secondary to cognitive functions in the learning continuum, are in fact critical determinants in the complex process of evaluative judgment (Candiotto, 2023; Dorado et al., 2023; Meyer & Turner, 2006). Zeidner (1998) illuminated the intricate ways in which diverse emotions, such as anxiety, stress, and confidence, can shape or distort an individual’s self-evaluation within educational contexts. Specifically, a student experiencing test anxiety might be predisposed to harsher self-critique, leading them to undervalue their genuine understanding. Conversely, the haze of overconfidence might deter neutral, objective introspection, making learners overlook areas that require refinement (Kruger & Dunning, 1999). Emotional capacities are integral components of the learning process, significantly influencing learners’ motivation, engagement, and overall academic outcomes (Pekrun, 2006). Rooted in the broad fields of psychology, these capacities pertain to an individual’s ability to recognize, understand, manage, and express emotions in a manner that facilitates learning (Salovey & Mayer, 1990; Gross, 1998). The interplay between emotion and cognition has been highlighted by several scholars, emphasizing the intertwined nature of these faculties in shaping the learning experience (Bandura, 1977; Deci & Ryan, 1985; Luo & Chan, 2023).

The role of emotions in evaluative judgment is not merely about the influence of transient feelings; it delves into the profound effects of emotion regulation (Gross, 2008; 2015; Koval et al., 2023; Namazidost et al., 2023). Effective emotion regulation ensures that learners can differentiate between their emotional responses and the objective quality of their work (Luo & Chan, 2023; Pekrun and Perry, 2014;), fostering unbiased and constructive evaluative judgment. Embracing positive emotions, such as curiosity and wonder, in educational experiences can recalibrate evaluative judgment. When students are motivated by genuine intrigue, they tend to engage
more deeply, often prompting a more reflective and accurate self-assessment process (Fredrickson, 2001). In essence, when positive emotions are at the helm, evaluative judgments can transition from being mere academic appraisals to profound reflections of personal growth and understanding. Therefore, for evaluative judgment to be both accurate and meaningful, it is essential to recognize the profound influence of emotions. By understanding and harnessing these emotional traces, educators and learners can cultivate a more holistic approach to self-assessment and feedback processes. In this section we review a few of the main emotional variables influencing the process of evaluative judgment.

**Self-efficacy.** Self-efficacy, as conceptualized by Bandura (1997), delves into the intricate belief systems learners have regarding their aptitude to carry out specific tasks. This psychological construct is not merely about possessing certain skills but rather the conviction about using them effectively. Such beliefs have far-reaching implications, sculpting learners’ reactions to challenges, their resilience when encountering setbacks, and their receptivity to feedback. When learners possess a robust sense of self-efficacy, they approach learning scenarios with a heightened sense of enthusiasm and determination (He et al., 2022; Pajares, 2002; Waddington; 2023). They operate with an underlying belief that, given the right amount of effort and strategy, they can surmount knowledge gaps and fine-tune their skills. Such learners are not easily deterred by complexity or initial failures. Instead, they view challenges as opportunities for growth and mastery, fostering a deep, intrinsic motivation to engage with content (Zimmerman, 2000). Importantly, this proactive stance extends to self-evaluation. Confident in their capacities, these learners are more inclined to undertake rigorous self-assessments, identifying areas of strength and pinpointing avenues for improvement (Schunk & Pajares, 2009).
On the flip side, learners with diminished self-efficacy often exhibit a reluctance to grapple with challenging content. Their fragile belief in their capabilities manifests as avoidance behavior, potentially bypassing rigorous evaluative processes for fear of confirming their perceived inadequacies (Pintrich & De Groot, 1990). Such learners might find themselves in a vicious cycle where limited engagement results in superficial understanding, reinforcing their self-doubt.

Feedback, an essential component of the learning process, is also filtered through the lens of self-efficacy. For those fortified with strong self-belief, criticism is not a threat but a tool (Yeager & Dweck, 2012). They interpret feedback as constructive guidance, a roadmap delineating area needing attention. In contrast, learners with dubious self-efficacy may perceive the same feedback as a harsh indictment of their capabilities, further eroding their confidence.

Self-efficacy doesn’t merely influence how learners approach tasks; it molds the very fabric of their learning processes, from engagement and perseverance to self-evaluation and feedback interpretation. Educators and mentors, recognizing the profound impact of this construct, can play key roles in nurturing and bolstering learners’ self-efficacy. In educational contexts, understanding the symbiotic relationship between self-efficacy and evaluative judgment is paramount. Tailoring pedagogical strategies to fortify self-efficacy can enhance learners’ capacity for accurate self-assessment, ultimately leading to improved learning outcomes and personal growth.

Motivation. Motivation is a foundational element in the learning process, acting as the primary catalyst that determines the depth and direction of educational pursuits (Ryan & Deci, 2000). The source of this motivation, whether internally derived or externally imposed, plays a focal role in shaping a learner’s interaction with the content and subsequently, their self-assessment of comprehension and skill acquisition. Intrinsic motivation is a natural inclination, ignited
by an individual’s own passion, curiosity, or interest in a given subject or task (Deci & Ryan, 1985). Learners driven by this internal drive aim not just for completion but for a profound understanding, desiring to truly master concepts, discern underlying connections, and achieve individual excellence. Such an unadulterated desire to learn, often detached from external gains, paves the way for a richer educational experience marked by exploration and insight. As a result, when these learners participate in evaluative judgment, they go beyond traditional success metrics. They assess their achievements through the lens of self-improvement, introspection, and genuine mastery (Csikszentmihalyi, 1997; Mendoza et al., 2023).

On the other hand, extrinsic motivation is influenced by tangible external rewards or deterrents, including awards, grades, or even the evasion of unfavorable outcomes (Vallerand & Ratelle, 2002; Kotera et al., 2023). Learners motivated by these external factors typically have a singular goal in sight: securing the reward or steering clear of negative repercussions. While this can lead to prompt task completion and adherence to outlined standards, it can also compromise in-depth engagement, emphasizing results over the learning process. In such instances, evaluative judgment tends to be tethered to external standards. Learners might measure their success based on alignment with set criteria or by contrasting their performance against that of their peers. The dynamic between intrinsic and extrinsic motivation highlights the complexity of the role of emotions in learning. Though they can coexist and at times synergize in certain scenarios, they influence distinct facets of the learning process and how learners perceive their progression (Kotera et al., 2023; Lepper, Corpus, & Iyengar, 2005). With this understanding, educators have the opportunity to design learning strategies that cater to these varied motivations, guiding learners towards a balanced integration of intrinsic enthusiasm and extrinsic diligence.
Interest. Interest and evaluative judgement are intrinsically linked in the domain of learning and assessment. Interest, as a motivational factor, can greatly influence how learners approach the task of evaluation. When learners are genuinely interested in a topic or subject, they tend to engage with it more deeply (Hendrawijaya, 2022; Hidi & Renninger, 2006), promoting a more thorough and insightful evaluative judgment. Their intrinsic motivation drives them not only to understand the content better but also to evaluate it more critically, ensuring that they grasp its intricacies.

On the other hand, when there is a lack of interest, the evaluative process may become more superficial or even biased. A student who is disinterested might not invest the same effort or time into evaluating a piece of work, leading to potentially skewed or incomplete judgments (Ainley, Hidi, & Berndorff, 2002). In these situations, evaluations are often based on peripheral features rather than the core substance of the content. Furthermore, interest can also play a role in peer evaluations. When learners are genuinely interested in their peers’ work, they tend to provide more constructive and detailed feedback (Turner & Patrick, 2008). This collaborative evaluative process becomes a shared learning experience, enhancing the quality and depth of the evaluation. To harness the benefits of interest in evaluative judgment, educators often use techniques to stimulate curiosity and engagement. By presenting content in an engaging and relevant manner, framing assignments around real-world applications, or allowing students autonomy in their choice of topics, educators can kindle interest, thereby enhancing the quality and depth of evaluative judgments (Renninger & Hidi, 2016).

Social Capacities of Evaluative Judgment

Sociocultural Background. The sociocultural background of a learner influences their learning experiences and offers a lens through which knowledge is interpreted and assimilated
Rooted in the sociocultural theory of cognition, it argues that learning is deeply embedded and influenced by the social interactions and cultural practices that an individual engages in (Lave & Wenger, 1991). The aspects of an individual’s societal background, be it language, traditions, or shared community beliefs, can dictate how information is processed and the strategies employed to grasp new concepts (Cole, 1996; Nishen & Kessels, 2022). Additionally, cultural narratives and societal norms can influence motivational aspects of learning, shaping the learner’s values, goals, and aspirations in an educational setting (Rogoff, 2003; Yoshida et al., 2023). As such, recognizing and appreciating the intricate relations between a learner’s sociocultural background and their learning mechanisms is paramount in creating inclusive and effective educational environments.

Beyond the immediate impacts on cognitive processes, the sociocultural background also exerts a profound influence on the very nature and context of learning activities that learners engage in. Bruner (1990) believes that culture shapes the "tools of intellectual adaptation," allowing each learner to harness these tools in unique ways that resonate with their cultural experiences. As learners interact with their environment, they also bring in a rich combination of cultural narratives and societal experiences that can both challenge and enrich the collective knowledge of a learning community (Nasir & Hand, 2006). Furthermore, learners from diverse sociocultural backgrounds often bring multiple perspectives to problem-solving, encouraging a deeper, more holistic understanding of subjects (Banks, 2004). Consequently, educators need to be acutely aware of these diverse backgrounds, not only to cater to individual needs but also to leverage this diversity as a resource, fostering a more inclusive and enriched learning environment.

**Learning Styles and Preferences.** Learning Styles and Preferences reflect the diverse approaches and strategies that individuals utilize to process, comprehend, and retain information.
This concept has gained prominence in educational psychology and pedagogy, especially after Kolb’s (1984) work on experiential learning which delineated distinct learning styles based on how learners engage with and transform their experiences. For instance, some individuals might be more visually oriented, preferring diagrams and charts, while others might be more kinesthetic, learning best through hands-on experiences (Felder & Silverman, 1988). Moreover, Gardner’s (1983) theory of multiple intelligences challenges the traditional notion of intelligence and emphasizes the plurality of ways in which learners can excel, be it linguistically, spatially, or interpersonally, among others. Recognizing and accommodating these diverse learning styles and preferences is crucial for educators to ensure that learning environments are inclusive and responsive to the varied needs of their students. Such awareness can also guide instructional design, ensuring that teaching methods are varied and holistic, catering to the broader spectrum of learner preferences and maximizing engagement and retention (Dunn & Griggs, 2000).

Learning styles and preferences, when related to evaluative judgment, determine how learners interpret, assess, and utilize feedback. The way learners prefer to process and internalize information significantly influences how they evaluate both their own work and the work of peers. For instance, a learner with a visual preference might be more adept at assessing graphical or visual representations and could benefit most from feedback presented in diagrams, charts, or visual annotations. Such learners might also be more critical and discerning when judging the visual presentation of content, given their predisposition to process information in this manner (Felder & Silverman, 1988). On the other hand, a learner with a kinesthetic preference, who learns best through hands-on experiences, might be more receptive to feedback through practical demonstrations or guided practice. Their evaluative judgment might be centered around the practical applicability and hands-on elements of a task (Kolb, 1984). Furthermore, learners’ preferences influence their
self-assessment capabilities. If the evaluation mode aligns with their learning style, they might be more accurate in their self-evaluations. For example, an auditory learner might excel in evaluating spoken presentations or oral exams but struggle more with written content (Gardner, 1983). Moreover, an awareness of one’s own learning style can influence metacognitive regulation. When learners are conscious of their preferred learning style, they can better align their study strategies, seek appropriate feedback types, and develop more accurate evaluative judgments about their learning outcomes (Rakovic et al., 2022). In conclusion, understanding the interplay between learning styles, preferences, and evaluative judgment can offer educators insights into tailoring assessment techniques to be more aligned with learners’ strengths, leading to more accurate and effective evaluative processes.

**Attitudes and Beliefs.** Attitudes and beliefs fundamentally influence every aspect of the learning process (Goldin et al., 2009; Uden et al., 2022). Learners’ perceptions and values determine not just how they approach educational tasks, but also how they interpret feedback, engage with content, and interact within educational environments. To begin with, one’s attitude towards a subject or learning in general can considerably impact their motivation and engagement levels (Kormos et al., 2011). For instance, if a learner holds a positive belief about their ability to grasp mathematics, they are more likely to exhibit resilience when faced with challenging problems, be more receptive to feedback, and invest time and effort in improving (Bandura, 1997). Conversely, negative beliefs can hinder learning, leading to avoidance behaviors, resistance to feedback, and overall decreased effort (Gal & Ginsburg, 1994). Beliefs also play a role in shaping learners’ self-efficacy, or the belief in their capability to execute tasks and achieve goals. When learners believe they possess the skills and knowledge to succeed, they are more likely to take
initiative, seek help when needed, and utilize feedback effectively to refine their understanding (Zimmerman, 2000).

Moreover, attitudes can also influence how learners perceive and interact with peers, instructors, and educational institutions (An et al., 2022). A learner who believes that collaborative learning is beneficial might actively seek peer feedback, engage in group study sessions, and contribute more to group projects (Johnson & Johnson, 1989). On the other hand, someone with a more individualistic belief might prioritize self-assessment and independent study. In the context of evaluative judgment, attitudes and beliefs can significantly impact how feedback is received and utilized. If a learner believes that they are inherently "bad" at a subject, they might dismiss positive feedback or overly focus on negative feedback, leading to skewed self-perceptions and judgments (Dweck, 2006).

In addition, attitudes and beliefs play a significant role in this process by acting as filters through which judgments are made (Reddy et al., 2016). For instance, a learner’s belief in the importance of collaborative work can influence how they evaluate group projects or peer contributions (Boud & Soler, 2016). If they have a positive attitude towards collaborative learning, they might place a higher value on contributions that facilitate team synergy, whereas someone with a negative attitude might undervalue these same contributions. Similarly, learners’ beliefs about their own capabilities, termed self-efficacy beliefs, can influence their self-assessments. Research indicates that students with higher self-efficacy tend to judge their own performances more accurately than those with lower self-efficacy (Pajares & Graham, 1999). Thus, learners’ attitudes and beliefs not only shape their approach to tasks but also the lens through which they evaluate and make judgments on performances. This underscores the importance of fostering positive and realistic beliefs in learners to enable effective and accurate evaluative judgments.
**Previous Educational Experiences.** Previous educational experiences serve as a cornerstone for learners, influencing their current approach to learning tasks, strategies they employ, and, most importantly, how they judge and evaluate work (Gladovic et al., 2022). These past experiences, whether positive or negative, act as reference points against which new learning encounters are measured (Wigfield & Eccles, 2000). For example, a student who received consistent positive feedback in a prior setting without constructive critique might find it challenging to navigate critical feedback in a new environment. This is because their evaluative judgment has been shaped by past validations and not by reflective self-assessment. On the other hand, learners who were exposed to diverse assessment techniques and encouraged to self-assess might have a more detailed and holistic approach to evaluative judgment in subsequent educational settings (Torrance, 2012). Their ability to judge their work and that of their peers might be more balanced and grounded in a deeper understanding of quality. In essence, past educational experiences, filled with unique experiences, mentors, feedback mechanisms, and challenges, cast a long shadow on how learners perceive, interpret, and act upon their present learning process, particularly in their evaluative judgments.

In fact, previous educational experiences, rich with successes, challenges, feedback, and unique interactions, play a role in shaping how learners engage in evaluative judgment in their current educational settings. As posited by Wigfield & Eccles (2000), these past experiences serve as cognitive and emotional reference points, influencing learners’ confidence, expectations, and standards of quality when judging their work or the work of their peers. A learner who previously faced a rigorous assessment system, receiving detailed feedback, is likely to have developed a keen eye for detail and quality, enabling them to discern details when evaluating tasks. In contrast, those who experienced superficial assessments might rely more on surface-level criteria
during judgment, possibly overlooking depth or critical intricacies (Torrance, 2012). Therefore, the depth, breadth, and nature of previous educational experiences act as a lens through which learners view, analyze, and evaluate content, impacting the rigor and quality of their evaluative judgments.

**Peer Comparison.** Peer comparison, deeply rooted in our social fabric, acts as a mirror reflecting learners’ positions relative to their counterparts in terms of knowledge, skills, and competencies. Festinger’s (1954) Social Comparison Theory posits that individuals have an inherent drive to gain accurate self-evaluations, and in environments lacking clear, objective metrics, they turn to peer comparison as a means of assessment. This theory’s relevance intensifies in the domain of evaluative judgment. In educational settings, especially collaborative learning environments, students often, either implicitly or explicitly, compare their work and understanding with that of their peers. This comparison serves a multifaceted purpose. On one hand, it allows students to calibrate their own understanding, helping to identify gaps in their knowledge or skills (Dunning et al., 2003). For instance, observing a peer effectively articulates a concept might highlight one’s own unclear understanding, prompting corrective measures. Furthermore, peer comparison can significantly influence the emotional dimensions of learning. Recognizing that one’s work aligns with or even surpasses peer standards can bolster self-efficacy and motivation. Conversely, perceiving a discrepancy, where one’s efforts fall short compared to peers, can lead to feelings of inadequacy, challenging the learner’s self-confidence and potentially skewing their self-assessment. From a metacognitive standpoint, peer comparison also plays a role in strategy evaluation. When learners observe peers employing different techniques or approaches to tasks, they might reflect upon and reassess the efficacy of their own strategies. This reflection can lead to adaptations and modifications in learning strategies to better align with
perceived best practices. Moreover, in formal peer assessment scenarios, where students are tasked with evaluating the work of their peers, the act of comparison extends beyond self-reflection. Students not only compare their work to the standards set by the curriculum but also benchmark against the work of their peers, potentially leading to richer insights and more detailed feedback. In essence, peer comparison, while seemingly a simple social act, deeply intertwines with the evaluative judgment process, shaping both the cognitive and emotional facets of learning, assessment, and self-perception.

**Learning Environment, Tools, and Resources**

In addition to the cognitive, metacognitive, emotional, and social dimensions of learning, learning environment, tools and resources used to present and manage learning process have also an impact on how learners perceive and assess their own and their peer’s learning and its outcome. The learning environment shapes a student’s capacity for evaluative judgment. An atmosphere that is both nurturing and supportive encourages students to critically reflect on their work and conduct in-depth assessments of their performance. Such positive environments bolster the accuracy of self-assessments and peer-assessment by providing learners with the confidence to genuinely evaluate their strengths and weaknesses (Joughin et al., 2023). Conversely, high-stakes, pressure-driven contexts can intensify stress, which can, in turn, distort students’ perceptions of their own capabilities. In these scenarios, learners might either underestimate or overinflate their skills, making genuine self-evaluation challenging (Tai et al., 2016).

In addition to the overarching learning environment, the tools and resources at a student’s disposal have a pronounced impact on their evaluative judgment. Tools like rubrics, exemplars, and guides offer students a benchmark, elucidating clear expectations and providing a yardstick against which they can measure their work (Phillips et al., 2004). When applied effectively, these
tools can impart clarity and consistency to students’ evaluations. However, there’s a potential downside: if students become too reliant on these tools, or if the tools aren’t tailored to their specific context, they can become a limiting factor. There’s a risk that students might overlook delicate elements of their work that don’t align with these preset criteria (Brown et al., 2015).

It’s worth noting the dynamic interplay between the learning environment and the tools available to students. When students are in a supportive setting, activities like reflective group discussions can enhance the value derived from tools like rubrics, as they facilitate deeper understanding and more application (Crosskerry, Singhal, & Mamede, 2013). Yet, in environments marked by pressure and high stakes, even the most well-designed tools may not be utilized to their fullest potential. This highlights the imperative of ensuring both a conducive learning environment and the provision of apt resources to cultivate strong evaluative judgment (Larrick, 2004).

**Task External Factors and Implications**

The manner in which students conceptualize the implications of their evaluations, particularly in terms of future prospects and professional roles, can significantly influence the depth and rigor of their self-assessment processes (Sadler, 2010). When assessments are contextualized within the larger framework of future career paths or real-world applications, students are more likely to perceive these evaluations as not just academic exercises, but as critical periods that can impact their professional futures (Trede, Macklin, & Bridges, 2012). Such a perspective inevitably urges them to engage more deeply, critically, and responsibly in the evaluative task at hand (Eva & Regehr, 2005).

Furthermore, external expectations, such as those set by professional bodies, academic institutions, or potential employers, can also serve as a driving force, nudging students towards a
more comprehensive self-evaluation (Boud, 2000). Recognizing the potential real-world implications of their judgements can motivate students to seek clarity, accuracy, and objectivity in their evaluations, ensuring they align with external standards and expectations (Dunning, Heath, & Suls, 2004). In essence, the intersection of external expectations with future implications offers a potent combination, anchoring the assessment process in reality and giving it a purpose and direction beyond the immediate academic context (Engeström, 2001).

**Figure 2.1**
*Evaluative Judgment Model for Task Internal, Learner and Task External Factors*

To Summarize, the review of literature in evaluative judgment accentuates the multifaceted influences stemming from task-specific, learner-specific, and external variables (Figure 2.1). Examining the intrinsic components of the task, it is evident that task complexity (Rahimi & Zhang, 2019), task guidelines (Cannon-Bowers et al., 2010), technology support (Lajoie & Azevedo, 2006), and exemplar availability (Thibaut et al., 2018) play central roles in
determining how learners’ approach and understand evaluative tasks. The learner factors encompass a broader spectrum, segmented into cognitive, emotional, metacognitive, and social dimensions. Cognitive elements like working memory and attention, linguistic aptitude, and prior knowledge underscore the intrinsic capabilities that learners bring into evaluative contexts. Concurrently, emotional facets, including self-efficacy and motivation, delineate the affective domain’s influence. The metacognitive aspects, highlighting regulation of cognition and reflection, stress the importance of learners’ self-awareness and adaptability. Social considerations, capturing the essence of sociocultural backgrounds and peer comparisons, reflect the broader environmental and interpersonal influences. Transitioning to task external factors, the learning environment feedback availability, and evaluation demands crystallize the overarching frameworks within which evaluative judgments operate. Taken together, this confluence of factors, as depicted in the model, offers a holistic vantage point, elucidating the intricate web of variables that shape and refine the process of evaluative judgment in educational scenarios.

Figure 2.2

Evaluative Judgment Conceptual Framework
The conceptual framework developed from the review of literature involves a multitude of aspects that cannot be studied in a single study. For the purpose of the current study, we have reduced the model to the study of evaluative judgment in regard to three factors: learning environment, assessment technique used to elicit evaluative judgment and a variety of individual differences factors. Figure 2.2 indicates such a relationship. The conceptual framework revolves around the central theme of "evaluative judgment," which serves as the main focus of the learning process. This judgment is shaped by an array of interconnected components, including the environment in which learning takes place and the tasks students undertake. The environment further bifurcates into technological and non-technological settings, highlighting the dual nature of modern education. These settings directly impact how evaluation tasks are perceived and approached by learners. Diving deeper into the tasks, their inherent characteristics such as guidelines, complexity, and familiarity become prominent. These aspects underscore the essence of the tasks, detailing the complexities learners must navigate. On the other hand, the learning environment, marked by feedback mechanisms, technological support, and specific demands, influences how these tasks are executed and evaluated. Furthermore, the intricate synergy between internal and external factors molds the learners’ evaluative judgment capabilities. For instance, while technological advancements facilitate task execution, the intrinsic complexity of a task might challenge a learner’s judgment.
3 METHODOLOGY

Research Design

This study is structured with a 2x2 mixed design, integrating two independent variables: the type of assessment and technology settings. The type of assessment, functioning as a within-subjects variable, comprises two distinct categories—self-assessment and peer-assessment. In the self-assessment condition, participants evaluate their own summaries, whereas, in the peer-assessment condition, they evaluate the summaries crafted by their peers. The second independent variable, technology settings, serves as a between-subjects variable and is dichotomized into two settings: technology and non-technology. In the technology setting, participants engage in assessments complemented by SMART technology-enabled feedback, aiming to enhance the evaluation process. Conversely, the non-technology setting represents a more traditional approach, with assessments conducted without technology-enabled feedback. The dependent variables for the study are evaluation score, evaluative judgment quality (comparison between evaluation score and expert evaluation) and rating confidence. Additionally, the effect of individual differences variables such as prior knowledge, reading comprehension and vocabulary knowledge are also considered. This design matrix allows for the examination of the main effects of assessment type and technology settings, alongside any potential interaction effects between these variables. Table 3.1 displays the design of the study. Also, the systematic approach utilized for the design of the present study is displayed in Figure 3.1.

Table 3.1
2 X 2 Mixed Design

<table>
<thead>
<tr>
<th>Assessment Type</th>
<th>Technology 60</th>
<th>Non-technology 60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-assessment</td>
<td>Group A (Self in Technology)</td>
<td>Group B (Self in Non-Technology)</td>
</tr>
<tr>
<td>Peer-assessment</td>
<td>Group A (Peer in Technology)</td>
<td>Group B (Peer in Non-Technology)</td>
</tr>
</tbody>
</table>
**Study Population.** The participants for this study were recruited from the population of online research workers on the Prolific platform. Prolific (www.prolific.com) is an online platform designed to assist in recruiting participants for academic research. Online research workers are screened research participants who take part in research studies in return for financial incentives. Researchers can define specific inclusion and exclusion criteria to help ensure that participants are suitable for their studies. The platform notifies eligible users about studies they qualify for through their user dashboard and via email, facilitating participant recruitment. The sample was recruited from the population of the participants on Prolific.

**Figure 3.1**

*Stages of the Study Experiment*

**Inclusion and exclusion criteria.** During the set-up of the study on Prolific, pre-screeners (Figure 3.2) were used to specify the inclusion and exclusion criteria for the prospective participants of the study. This makes it possible for prolific to display the count of eligible participants
that can take part in the study prior to the publishing of the study. In alignment with the goals of our research, we established the following criteria for participation: (1) enrolment as an undergraduate college student, (2) current residency within the United States, (3) representation across various educational stages—from freshmen to seniors, (4) with an age range of 18 to 40 years old and (5) and know either English or know one other language in addition to English.

Any prospective participant who did not meet these specific requirements was automatically excluded from the study, upholding the integrity and relevance of our research outcomes.

**Figure 3.2**

*Prolific Pre-Screeners and Eligible Participants*

![Prolific Pre-Screeners and Eligible Participants](image)

**Study sample:** With pre-screening the population pool in *Prolific* for the targeted sample of undergraduate US students across different levels of undergraduate education, the number of eligible participants was reduced from globally available 120,000 participants to 4500 eligible participants in the United States. Factoring in other factors such as age (18-40) also reduced the number to 2500 participants and finally, including those whose first language was English but also knew additional languages set the pool to 1800 active participants. The eligible participants in the pool of 1800 received the advertisement for the study in their Prolific account. Prolific
makes it possible for researchers to recruit from a larger pool (i.e., 1800) and fill in the spots for participants based on selection criteria until the cap of the sample size is reached. The assignment from the 1800 to the 120 participants was done randomly so that each participant in the pool had an equal chance of being included in the study.

Randomization. Prolific allows for randomization in three different ways including representative, balanced and standard sampling. Representative Sampling is the most expensive option and selects participants from Prolific user base in a way that accurately represents the population of interest. It requires a sample of at least 300 participants. Since the representative sample requires a sample of at least 300 participants, and standard sampling is using a convenience sampling logic, a balanced sample was selected. When balanced sample is selected, Prolific will balance the sample to 50% male and 50% female participants. Balanced sampling on gender factor was used because it allowed inclusion of equal number of male and female participants so that gender effect is controlled as a confounding factor. From the sample pool of 120 participants, 60 males and 60 females were randomly assigned through a stratified randomization method to the experimental conditions—technology and non-technology—ensuring a balanced representation in each group. Balanced samples are frequently utilized in research to ensure equal representation of various groups, thereby minimizing biases (Kirk, 1995). Equitable group sizes in experimental designs have been shown to mitigate potential influences of confounding variables, allowing for more accurate attributions of differences to the treatments under investigation (Maxwell, 2004). Moreover, enhanced statistical power is often achieved with balanced samples, increasing the likelihood of detecting genuine effects (Jaccard & Guilamo-Ramos, 2002). Therefore, the validity and generalizability of research findings are bolstered by employing balanced samples (Cook et al., 2002).
Participants

**Rationale for inclusion and exclusion.** This study focuses on undergraduate students aged 18-40 located in the United States. This group spans across freshmen to senior educational years, including both genders, and communicates in English and in an additional language. The primary reason behind this demographic delimitation was to secure a balanced and representative sample of the undergraduate student community. The age bracket of 18-40 encapsulates the typical undergraduate demographic, allowing for participants who are navigating similar stages of cognitive and educational growth (Arnett, 2000). The inclusion of both male and female
participants was to ensure the sample’s representativeness, reflecting gender diversity in higher education (Hyde, 2005). Furthermore, data concerning the race/ethnicity of the participants was procured to counteract biases linked to specific racial or ethnic affiliations. The inclusion of English and an additional language as a criterion was grounded in the aim to account for linguistic diversity, cultural backgrounds, and cognitive abilities associated with bilingualism (Bialystok, 2009). Engaging with multiple languages often demands enhanced cognitive flexibility (Costa & Sebastián-Gallés, 2014) and offers a unique perspective on cultural nuances (Pavlenko, 2005), which can be instrumental in understanding broader patterns in research findings. Moreover, bilingualism and multilingualism are increasingly common, particularly in diverse settings such as college campuses (Fishman, 2000), hence ensuring a wider and more inclusive scope of data.

Another pivotal aspect of our sample design was the inclusion of students from different educational years. This approach was adopted to encapsulate a wide spectrum of experiences, challenges, and knowledge bases within the undergraduate domain (Tinto, 1993). Freshmen, for instance, offer insights from a transitionary phase, adapting from high school to college dynamics (Upcraft, Gardner, & Barefoot, 2005). Sophomores and juniors, on the other hand, provide mid-course perspectives, often juggling core courses and extracurricular activities (Gahagan & Hunter, 2006). Seniors, in their conclusive year, bring forth a culmination of their undergraduate journey, often reflecting a blend of academic maturity and future aspirations (Gardner, 2009). By harnessing insights across these educational years, the study aims for a holistic understanding, spanning the entire range of the undergraduate experience.

**Data cleaning.** After recruiting 120 participants, 60 each for the technology and non-technology groups, they were randomly assigned to their respective groups, and data collection began. The platform, Prolific, allows researchers to check the collected data and remove any data
point that does not meet the study’s requirements. Upon checking, it was found that some participants did not complete the study as expected. Common issues included participants copying the original text as their summary, writing summaries that were too brief, or adding unrelated content in the summary section. There were also instances where the study timed out before participants could finish. Due to these reasons, the final count was adjusted to 47 participants in the technology group and 46 in the non-technology group.

Table 3.2 provides demographic details of the remaining participants. The mean age was 23.50 years with a standard deviation of 4.8. For the technology group, the average age was 25.36 (SD=5.51), and for the non-technology group, it was 23.76 (SD=4.84). In terms of gender distribution, males made up 51.6% of the participants, with 31.2% in the technology group and 20.40% in the non-technology group. Females represented 48.4%, distributed as 26.90% in the technology group and 21.50% in the non-technology group.

In the category of race and ethnicity, the majority of participants identified as White, accounting for 73.1% of the total. In the technology group, 33.3% identified as White, whereas 39.8% did in the non-technology group. The non-technology group had a higher representation of Black and Mixed individuals at 6.5% and 7.5% respectively, compared to 3.2% and 4.3% in the technology group. Additionally, the technology group had 5.4% Asian participants, in contrast to 2.2% in the non-technology group. The non-technology group also included 3.2% of participants identifying as ‘Other,’ a category with a count of 0 in the technology group.

When it comes to language proficiency, monolingual participants accounted for 76.67% overall, with 79.17% in the technology group and 73.81% in the non-technology group. Bilingual participants made up 23.33%, with the technology group having 20.83% and the non-technology group having 26.19%.
Regarding academic progression, the distribution across both groups was as follows: 1st-year students accounted for 7.78% overall, with the technology group having 2.22% and the non-technology group 5.56%. The technology group had a larger representation of 2nd and 3rd-year students at 12.22% and 15.56% respectively, compared to 7.78% and 11.11% in the non-technology group. Fourth-year students constituted 45.56% overall, with 20% from the technology group and 25.56% from the non-technology group.

Table 3.2
Demographics of the Study Sample

<table>
<thead>
<tr>
<th>Category</th>
<th>Details</th>
<th>Technology (N=47)</th>
<th>Non-technology (N=46)</th>
<th>Total (93)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Average age</td>
<td>25.36 (5.51)</td>
<td>23.76 (4.84)</td>
<td>23.50 (4.8)</td>
</tr>
<tr>
<td>Gender</td>
<td>Male</td>
<td>29 (31.2%)</td>
<td>19 (20.40%)</td>
<td>48 (51.6%)</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>25 (26.90%)</td>
<td>20 (21.50%)</td>
<td>45 (48.4%)</td>
</tr>
<tr>
<td>Race/Ethnicity</td>
<td>White</td>
<td>31 (33.3%)</td>
<td>37 (39.8%)</td>
<td>68 (73.1%)</td>
</tr>
<tr>
<td></td>
<td>Black</td>
<td>3 (3.2%)</td>
<td>6 (6.5%)</td>
<td>9 (9.7%)</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>4 (4.3%)</td>
<td>7 (7.5%)</td>
<td>11 (14.17%)</td>
</tr>
<tr>
<td></td>
<td>Asian</td>
<td>5 (5.4%)</td>
<td>2 (2.2%)</td>
<td>7 (7.5%)</td>
</tr>
<tr>
<td></td>
<td>Other</td>
<td>0 (0%)</td>
<td>3 (3.2%)</td>
<td>3 (3.2%)</td>
</tr>
<tr>
<td>Language</td>
<td>Monolingual</td>
<td>38 (79.17%)</td>
<td>31 (73.81%)</td>
<td>69 (76.67%)</td>
</tr>
<tr>
<td></td>
<td>Bilingual</td>
<td>10 (20.83%)</td>
<td>11 (26.19%)</td>
<td>21 (23.33%)</td>
</tr>
<tr>
<td>Education Level</td>
<td>1&lt;sup&gt;st&lt;/sup&gt; Year</td>
<td>2 (2.22%)</td>
<td>5 (5.56%)</td>
<td>7 (7.78%)</td>
</tr>
<tr>
<td></td>
<td>2&lt;sup&gt;nd&lt;/sup&gt; Year</td>
<td>11 (12.22%)</td>
<td>7 (7.78%)</td>
<td>18 (20%)</td>
</tr>
<tr>
<td></td>
<td>3&lt;sup&gt;rd&lt;/sup&gt; Year</td>
<td>14 (15.56%)</td>
<td>10 (11.11%)</td>
<td>24 (26.67%)</td>
</tr>
<tr>
<td></td>
<td>4&lt;sup&gt;th&lt;/sup&gt; Year</td>
<td>18 (20%)</td>
<td>23 (25.56%)</td>
<td>41 (45.56%)</td>
</tr>
</tbody>
</table>

In summary, it is acknowledged that there are differences between the Technology and Non-technology groups in racial composition and language proficiency, although both groups show similarities in academic progression, age, and gender distribution which are not the focus of the current study. The results of the study will be discussed with these differences in mind.

Reading Passage and Evaluation Task

The design of the study, as depicted in Figure 3.1, involved both technology and non-technology groups engaging in evaluation of their own and their peer’s summary. Participants
were given a passage about Brain-Computer Interactions (see Appendix B) to read and were asked to write a summary ranging from 200 to 350 words. Those in the technology group used the SMART feedback for evaluation of summary in both self-assessment and peer-assessment. Meanwhile, the non-technology group also conducted self-assessment and peer-assessment but used an expert summary model to evaluate the summary. During the evaluation activities, which included both self-assessment and peer-assessment, both the technology group and the non-technology group applied the same rubric (Gao et al., 2019). The evaluation activities are described with further details.

**Reading passage.** The passage for the study (Figure 3.4) offers a comprehensive exploration of the field of Brain-Computer Interaction (BCI), starting with a comparison between the functionalities of the human brain and a computer. It details the implications of the potential integration of human and machine intelligence, with applications in detecting cognitive illnesses, controlling complex machinery, and navigating virtual worlds. This passage further discusses the transition of BCI from academic research to industry, the roles of different sensor systems in interpreting brain activity, and the pros and cons of invasive versus non-invasive methods. The text also addresses current limitations in real-world uptake, issues with machine learning accuracy, and problems with sensor systems. Furthermore, it highlights the role of wearable computers in BCI research, and the prospect of highly immersive experiences offered by these technologies. Finally, the passage describes ongoing research on next-generation direct-sense BCI (DS-BCI) technologies, focusing on speech and vision-based systems that decode brain signals linked to natural senses, offering an innovative and more intuitive way to interact with BCIs.

Given the diversity of academic backgrounds among the undergraduate participants, it was important to choose a reading passage that would resonate with them, irrespective of their
field of study. The selected reading passage, pertaining to the popular science theme of Brain-Computer Interactions, was deemed suitable for several reasons. First, it presented a topic that was both intriguing and universally applicable, ensuring its relevance to students from various disciplines. Second, the complexity of the content was carefully calibrated to match the reading capabilities of college students, aligning with the findings of Cervetti et al. (2009) which highlighted the importance of selecting appropriately challenging texts. Moreover, the richness of key concepts embedded within the passage was important. This not only provided a fertile ground for summary writing but also ensured that learners had a plethora of essential points to incorporate, facilitating the drawing of connections and fostering deeper comprehension. The decision to adopt a popular science text for this study is supported in empirical research which underscores the efficacy of such texts in engaging diverse student populations. Not only do these texts encapsulate complex scientific concepts in an accessible manner, but they also bridge the gap between academic rigor and general interest. For instance, Graesser et al. (1994) underscored the value of popular science texts in fostering deep comprehension, attributing this to their balanced mix of factual information with narrative elements. Similarly, Kintsch (1998) emphasized their potential in enhancing students’ ability to derive inferences, given the intricate interplay of facts and explanatory details in such texts. Furthermore, a study by McNamara et al. (2001) on text-based learning highlighted the role of popular science texts in bolstering cognitive engagement and facilitating the retention of key concepts. These findings were echoed by Snow (2002), who pointed out that the narrative style common to popular science texts aids in capturing student interest while ensuring the learning outcomes are achieved. The choice of a reading passage on Brain-Computer Interactions (BCI) was further underlined by its inherent interdisciplinarity, appealing to students from varied academic backgrounds.
Figure 3.4 Reading Passage for Summarization Task

**Summarization.** In the study, participants were tasked with summarizing a passage by writing a summary of 200–350 words. For those in the non-technology group, a text box in
Qualtrics was provided (Figure 3.5). This allowed them to view the passage and write their summary simultaneously. Participants in the technology group, on the other hand, read the passage in the reading space of the SMART platform and used SMART’s integrated text box to write their summaries as shown in Figure 3.6.
Instructions: Please use the provided space to write a summary of the passage. Your summary should be between 200-350 words. Your summary should not exceed 350 words.

Figure 3.5 Summarization Text Box

Figure 3.6 Summarization Text Box in SMART
Instructions: After finishing the summary writing task, you will now proceed to the self-Evaluation activity. First, you will compare your summary with the expert model. The source text is also provided if needed. Let's begin.

Note: You can find the summary that you wrote in the previous section here. You can also see an expert model. An expert model of a summary is a reference summary of the same passage you summarized in the previous section.

Your Summary

We are already familiar with the concept of wearable computers, through VR headsets, AR glasses and other trendy tech. As well as their entertainment value, wearable computers also have many promising practical applications. Wearable computers can offer highly immersive experiences for entertainment, health monitoring and research purposes, among many others. Their research applications are most exciting for us at present. Wearable computers have revolutionized the practicalities of BCI research. Until recently, BCI research has relied upon static and simple stimuli - presenting an object to a subject in a lab environment, for instance - which does not bear much resemblance to everyday life. By using wearable computers, researchers can design, simulate, and finely control experiments to examine human brain dynamics inside and outside the laboratory. VR and AR can now create sophisticated scenarios like real life; by monitoring a subject's brain activity when encountering these scenarios, results are far more meaningful in terms of relating findings to the real world.

Expert Model

Recent advancements in technology have allowed for the potential of linking brains and computers to become a practical reality. Brain-computer interfaces (BCIs) provide a channel for humans to interact with external artificial devices through their brain activity. While some measurement techniques for brain activity are invasive, non-invasive sensors are currently the only practical solution for investigating cognitive processes in the human brain. Wearable computers have revolutionized the practicalities of BCI research and allow for more sophisticated scenarios to be created, providing more meaningful results. Researchers are working on a next-generation solution called direct-sense BCI (DS-BCI), which aims to seamlessly decode brain signals linked to natural senses without additional stimulus. Direct-speech BCI translates 'silent speech' from neural signals into system commands and could be an important assistive tool for people who cannot speak naturally. Direct-sight BCI detects what objects are in a person's mind based on their EEG signals as they look around an environment, making it more innovative than current BCI methods. However, there are practical limitations to BCIs, including the need for improvement in the sensors that pick up signals from the brain and the fact that most BCIs only work when the user is stationary, limiting their use in many real-world applications. The accuracy of machine learning also provides another limitation, although this is a rapidly advancing field. The response feedback produced by the computer is far slower than our brains, creating a delay, and interacting with the real world via a computer is unnatural to humans. Overall, the possibilities for the future of BCIs are exciting to researchers, as the combination of human and machine intelligence has a staggering array of potential real-world applications, from the detection of cognitive illnesses or emotional states to precise control of sophisticated machines, to the intuitive navigation of immersive virtual worlds.

Instructions: In this section you are given a rubric to evaluate your own performance. Please read the rubric carefully and select option 0-5 that best describes your summary.

Figure 3.7 Feedback for Non-Technology Group
Feedback. Based on their assigned group, the non-technology participants completed the evaluation task via Qualtrics, where they read the passage, crafted their summary, and compared their summaries with an expert version during both self and peer-assessment processes (see Figure 3.7). Conversely, participants in the technology group engaged with the SMART system (Kim et al., 2019) where they read the passage, draft and submitted their summaries, and utilized feedback from SMART for self-assessment and peer-assessment (refer to Figure 3.8). The feedback provided by SMART includes a progress bar reflecting the summary’s quality, a message center that provides quantitative reports on the summary’s key concepts and relations, comparison between participant’s and expert’s concept maps of the summary, and feedback on missed concepts and connections.

![Figure 3.8 Feedback in Technology Setting](image)

Finally, using feedback information, either from the expert summary or from SMART (Kim et al., 2019), the participants in each group were required to use a rubric (Gao et al., 2019)
(Appendix D) to evaluate their own and their peer’s summary. A snapshot of the Rubric is given in Figure 3.9.

![Image of Evaluation Rubric]

**Figure 3.9 Evaluation Rubric**

**Instruments**

In addition to the evaluation task, the data collection was conducted through multiple instruments to measure the participants’ exiting reading comprehension ability, vocabulary knowledge, prior knowledge, and rating confidence. Each of these measures is described as follows.

**Reading comprehension test.** The Gates-MacGinitie Reading Comprehension Test (MacGinitie et al., 2000) was utilized to measure participants’ reading comprehension abilities. This selection was influenced by its well-documented reliability and validity in the literature (Allen et al., 2016; Calloway et al., 2022; Connor et al., 2015; MacGinitie et al., 2000), making it a
reliable and valid tool for reading comprehension assessment. The test evaluates various reading text types, encompassing narrative, expository, and informational genres. Specifically tailored for adult learners, it provides standardized scoring and interpretation. Furthermore, the outcomes of this test align well with findings from other studies, fostering uniformity in reading comprehension research.

Structured with passages extracted from diverse published sources, the test concludes each segment with a set of multiple-choice questions, cumulatively presenting 48 questions. These are designed to probe a range of reading competencies. An illustrative test item can be seen in Figure 3.10. The electronic administration of the Gates-MacGinitie test facilitates rapid and unbiased grading. Additionally, its concise format ensures completion within an approximate duration of 35 minutes.

Regarding its psychometric properties, MacGinitie et al. (2000) have documented an internal consistency range of 0.91-0.93 and an alternate-forms reliability of 0.80-0.87. In the context of this study, the Gates-MacGinitie Reading Comprehension Test demonstrated Cronbach’s alpha of $\alpha = 0.85$. This value underscores a high level of internal consistency, suggesting that the items on the test reliably measure a singular underlying construct, in this case, reading comprehension. See Appendix A for the full version of the test.
In later life, John Quincy Adams recalled an incident typical of his mother Abigail's bravery and resourcefulness. In 1775 British troops from Boston were advancing on Braintree, searching for rebel arsenals. All day neighbors traveled the road in front of the Adams' farmhouse, retreating from the expected attack. Abigail was alone in her home with her children. When rebel troops arrived, they advised Abigail to flee. Instead, she stayed, handing over all her precious pewter to the rebels, helping them melt down the metal for bullets. The rebel soldiers departed, and Abigail remained, expecting the worst but refusing to give in to the panic that possessed some of the neighbors. "Do you wonder," wrote her son, "that a boy of seven who witnessed this scene is a patriot?"

The neighbors who passed the Adams' house were trying to

- Defend their homes.
- Avoid being hurt.
- Join one of the armies.
- Get to Boston.

The passage suggests that the rebels had little

- Ammunition.
- Concern for Abigail.
- Knowledge of the countryside.
- Warning that the British were advancing.

What demonstrated Abigail's resourcefulness was the way she

- Fooled the British troops.
- Sent messages to the rebel troops.
- Learned where the British troops had come from.
- Provided what was needed from what she had available.

John Quincy Adams believed that this experience was a source of his

- Resourcefulness.
- Interest in military history.
- Courage.
- Love of country.

---

**Figure 3.10 Gates-MacGinitie Reading Test Item**

**Vocabulary knowledge test.** To assess the vocabulary knowledge of participants, this study employed the Vocabulary Size Test (monolingual, 20000, version A, Nation, 1990). This test, specifically designed for native English speakers, is structured with 100 short questions.
Figure 3.11 Sample Item of Vocabulary Size Test

Each question has stems that range from 4 to 10 words, and participants are tasked with selecting the word that best mirrors the meaning of the stem word. The decision to adopt the Vocabulary Size Test was grounded in multiple considerations. First, its design tailors it well to
native English speakers, as underscored by Nation (1990). Moreover, the test presents an extensive assessment of vocabulary by covering a variety of topics. This ensures a deep and holistic understanding of participants’ vocabulary knowledge (Nation, 1990). Finally, the relevance of the Vocabulary Size Test to reading comprehension played a central role in its selection. Numerous studies have pinpointed significant correlations between this test and reading comprehension capabilities (Ha, 2021; Ibrahim et al., 2016; Laufer & Ravenhorst-Kalovski, 2010; Laufer, 1992; Liu, 2016; Zhang & Bin Anual, 2008). This consistent trend in previous research further solidifies the test’s appropriateness for our study’s objectives. In essence, the Vocabulary Size Test, with its robust design, proven reliability, and relevance to reading comprehension, emerged as an ideal choice for this study. Another crucial facet that influenced the selection of this test is its reliability. Historically, the Vocabulary Size Test has demonstrated robust reliability indicators. As detailed by Nation (1990), the test boasts a Cronbach’s alpha of 0.95 and a Rasch reliability of 0.97. This study’s reliability analysis aligns with this precedent, yielding Cronbach’s alpha of $\alpha = 0.85$. This result indicates a strong internal consistency. A snapshot of the test is given in Figure 3.11 and a full version is available in Appendix D.

**Prior knowledge.** To assess learners’ prior knowledge of the reading content, this study implemented a Likert scale question, prompting participants to rate their familiarity with the topic with response options ranging from *Yes, Rather Yes, Rather No* and *No*.

“Do you have prior knowledge of the topic of Human-Brain Interaction (HBI), a field in which computer technology is integrated with the human brain?”

a. Yes  
b. Rather Yes  
c. Rather No  
d. No
A response of "Yes" carries a numerical value of 4, indicating a strong positive response. "Rather Yes" is assigned a numerical value of 3, representing a moderate positive response. Conversely, "Rather No" is associated with a numerical value of 2, signifying a moderate negative response. Finally, a response of "No" corresponds to a numerical value of 1, indicating a strong negative response. These numerical values are employed to gauge the strength of respondents’ attitudes or levels of prior knowledge, enabling quantitative analysis and comparison of survey responses.

The significance of gauging topic familiarity is highlighted by Lucassen et al. (2013) and Bråten et al. (2018), pointing to its influence on learners’ evaluation strategies. Such scales are particularly valuable as they capture the gradations in familiarity levels, which can profoundly influence cognitive processing (Kintsch, 1998; McNamara & Kintsch, 1996). Barzilai et al (2020) further expound on the precision of Likert scales, noting their capability to discern subtle variations in an individual’s prior knowledge about a specific topic. This is reinforced by studies like Tabari and Wang (2022) which have showcased the efficacy of single-item Likert scale measures for gauging topic familiarity, underscoring their ability to reliably reflect a participant’s self-perception of knowledge. In a similar vein, findings by Barzilai et al. (2020) emphasize the richness of data derived from Likert-scale questions when compared to binary or open-ended questions. Beyond these qualitative benefits, Likert scales are also preferred for their analytical convenience. Their repeatability and standardized format facilitate cross-study or cross-population comparisons, a pivotal aspect in academic research aiming to build upon established findings (Sullivan & Artino, 2013).

**Summary evaluation score.** The participants evaluated the quality of their own and their peers’ summaries, using a rubric suggested by Gao et al. (2019). The rubric has been
successfully used in previous studies involving summary rating (Davies et al., 2021; Gao & Passonneau, 2021; Kim et al., 2022; Zhang et al., 2021). Gao et al. (2019) rubric provides a structured framework for evaluating student summaries across four distinct dimensions: content quality, content coverage, content coherence, and the strength of the argument presented.

<table>
<thead>
<tr>
<th>Points</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>CONTENT: quality</strong></td>
<td>Most of the ideas in the summary and argument are either not central to the topic, nor expressed clearly, or are vague or repetitive.</td>
<td>Many of the ideas in the summary and argument relate to the topic, but only a few are central to the topic, which may be due to vagueness, repetition, lack of clarity, or failure to express central ideas.</td>
<td>About half the ideas in the summary and argument are expressed clearly and are central to the topic, but about half the ideas do not meet the combined criteria of clarity and centrality, which may be due to vagueness, repetition, lack of clarity, or failure to identify central ideas.</td>
<td>About half the ideas in the summary and argument are expressed clearly and are central to the topic, and there is little to no vagueness or repetition. However, about half the ideas are either not central to the topic or unclear.</td>
<td>Most of the ideas in the summary and argument are related to the topic, most of them are central to the topic, and all or nearly all are expressed clearly, with little or no vagueness or repetition.</td>
<td>All or nearly all the ideas in the summary and argument are related to the topic, and all or nearly all are expressed clearly, with little or no vagueness or repetition.</td>
</tr>
<tr>
<td><strong>CONTENT: coverage</strong></td>
<td>Most of the central ideas from the article(s) are not expressed clearly in the summary and argument, and ideas from the article(s) are not included where that is unclear, vague, or repetitive.</td>
<td>Some of the central ideas from the article(s) are expressed clearly in the summary and argument, but many of the central ideas from the article(s) are missing, or are expressed in a way that is unclear, vague, or repetitive. Most ideas from the article(s) that are expressed clearly in the summary and argument are not central to the topic.</td>
<td>Many of the central ideas from the article(s) are expressed clearly in the summary and argument, but many of the central ideas from the article(s) are missing, or are expressed in a way that is unclear, vague, or repetitive. Many ideas from the article(s) that are expressed clearly in the summary and argument are not central to the topic.</td>
<td>Most of the central ideas from the article(s) are expressed clearly in the summary and argument. The remaining ideas from the article(s) are expressed clearly in the summary and argument are not central to the topic, and are expressed clearly, with little vagueness or repetition.</td>
<td>Most of the central ideas from the article(s) that are expressed in the summary and argument are not central to the topic, and very few are expressed in a way that is unclear, vague or repetitive.</td>
<td>All or nearly all the central idea(s) from the article are expressed clearly in the summary and argument. Very few of the ideas from the article(s) that are expressed are expressed clearly in the summary and argument are not central to the topic, and very few are expressed in a way that is unclear, vague or repetitive.</td>
</tr>
<tr>
<td><strong>CONTENT: coherence</strong></td>
<td>The ideas expressed in the summary and argument are not easy to follow, and do not relate well to one another.</td>
<td>Some of the ideas expressed in the summary and argument relate well to one another, but not all of the ideas do not relate well to one another, and are not easy to follow.</td>
<td>Many of the ideas expressed in the summary and argument relate well to one another, making it fairly easy to follow much of the discussion. But many of the ideas expressed in the summary and argument do not relate well to one another, so it is difficult to form a coherent understanding of.</td>
<td>Most of the ideas expressed in the summary and argument relate well to one another, and the discussion as a whole is fairly easy to follow. A few ideas seem out of place or less well integrated into the overall organization.</td>
<td>Most of the ideas expressed in the summary and argument relate well to one another, and the discussion as a whole is fairly easy to follow. Ideas flow well from one to the next, and the overall organization is very coherent.</td>
<td>All or nearly all the ideas expressed in the summary and argument relate well to one another, making it easy to follow the discussion as a whole. Ideas flow well from one to the next, and the overall organization is very coherent.</td>
</tr>
</tbody>
</table>

Figure 3.12 Summary Evaluation Rubric

The rubric consolidates individual scores from multiple aspects into a final aggregate score. The quality of content is evaluated on the clarity and relevance of ideas, ranging from 0 where most ideas lack detail and are vague, to 5 where all ideas are relevant, detailed, and clearly
expressed. Coverage is assessed based on how comprehensively the central ideas are presented and developed, with a score of 0 indicating many central ideas are missing, and a score of 5 indicating a thorough and detailed expression of ideas. Coherence pertains to the logical flow and connection of ideas, with a score of 0 for ideas that do not relate well to one another, scaling to 5 for ideas that are articulated in a very coherent manner. The strength of the argument is judged on the clarity and support of the claim, starting at 0 for a summary that does not respond in a relevant way to the issue, up to 5 for an argument that is exceptionally well-crafted and free of inconsistencies. The final score of the summary is an aggregate of these aspects, providing a comprehensive evaluation of the summary’s overall performance. Figure 3.12 indicates a snapshot of the rubric. For a full version of the rubric see Appendix C.

**Rating confidence.** Research indicates that learners’ confidence plays a significant role in both self-assessment (Hosein and Harle, 2018; Jones et al., 2020; Khan et al., 2001; Yuen-Reed and Reed, 2015) and peer assessment (Andrade, 2019; Falchikov, 1995; Logan, 2009; Theising and Sheehan, 2014). However, the relationship between confidence and assessment isn’t always straightforward. Overconfidence, where individuals consistently overestimate their abilities, can lead to significant miscalibrations in self-assessments. Dunning and Kruger’s work (1999) highlights this, where they found that individuals with lower skills in a particular domain were not only poor performers but also unaware of their lack of skill, leading to inflated self-assessments. Similarly, in case of peer-assessment, an individual’s confidence can influence the manner and accuracy of these evaluations. Studies suggest that individuals with higher confidence may be more assertive in their judgments, trusting their evaluations more compared to those with lower confidence (Falchikov & Boud, 1989). Their heightened self-efficacy may lead them to believe that they have a better understanding of the criteria and standards set for
assessment. Due to the effects of confidence on learners’ self-assessment and peer assessment, this study uses a single question item to measure participant’s confidence in their ratings of their own summary and that of their peers.

At this stage, you need to rate how confident you feel about your self-assessment. Use the following rating scale to present your level of confidence.

How confident do you consider yourself in your grading?

a. Confident
b. Rather Confident
c. Rather Not Confident
d. Not Confident

The single question survey using a Likert Scale for rating confidence, quantifies the respondents’ levels of agreement using an ordinal approach. The scale includes four options: "Confident," indicating a strong positive response with a numerical value of 4; "Rather Confident," indicating a moderate positive response with a numerical value of 3; "Rather Not Confident," indicating a moderate negative response with a numerical value of 2; and "Not Confident," indicating a strong negative response with a numerical value of 1. The score from this survey would reflect the overall confidence of the respondents, with higher scores demonstrating a stronger agreement or confidence in the subject of the question. This method allows for a nuanced view of respondents’ attitudes, capturing varying degrees of agreement or disagreement.

**Evaluative judgement quality.** In order to measure participants’ evaluative judgment quality, the participants’ evaluation scores were compared with those by experts. To do so, two experts used the same rubric (Gao et al., 2019) to evaluate the participants’ summaries. The raters for this study are two PhD candidates from the department of Learning Technologies. Both have a background in reading, writing, and reviewing scientific research articles. In addition to
their research, they teach undergraduate courses in their field. Their role as PhD candidates ensures that they have undergone rigorous training in reading, analyzing, and summarizing scientific and academic literature. Furthermore, their experience in teaching undergraduate courses has equipped them with the ability to evaluate students’ work critically. Given their familiarity with the standards and conventions of academic writing, combined with their continuous engagement with research in their domain, they possess the necessary expertise to evaluate summaries effectively. During the rater training process, the researcher provided a comprehensive overview of the rubric’s band scores and the associated scale. To ensure consistency and accuracy in ratings, both raters were trained using 5% of the summaries. This hands-on approach allowed them to familiarize themselves with the expected standards and the details of the rating process. Subsequent discussions were facilitated between the raters to address and resolve any discrepancies or differences in their evaluations. The inter-rater reliability measures affirmed the efficacy of this training approach. For the pilot phase (5% of the cases), the Interclass Correlation Coefficient (ICC) was recorded at .71, falling within the acceptable agreement range. In the main phase of the study, the interrater reliability further improved, registering an ICC of .81, which is indicative of a high level of agreement between the raters.

The participants’ evaluative judgment quality was computed based on the numerical similarity formula. This generates similarity measures, ranging 0 (completely dissimilar) to 1 (completely similar). The numerical similarity formula compares two numerical measures from a student’s evaluation and the experts’ evaluation.

$$s = 1 - \frac{|v_1 - v_2|}{\max(v_1, v_2)}$$

(1)

where $v_1$ is the index value of a student’s score, and $v_2$ is the average score of the expert raters.
The numeric similarity is an objective and precise measure that allows for consistent calculation of similarity between two numerical values. In addition to the facility of the use, the formula has been successfully used in previous studies on quality of learner summaries (Kim, 2021, 2015; Kim & McCarthy, 2020, 2021).

**Procedures**

Figure 3.13 outlines the research procedure adopted in the study. Participants, regardless of their feedback setting either technology-enabled (SMART) or non-technology-based (Qualtrics) followed a consistent sequence of activities. Initially, during the self-assessment phase, participants signed the consent form, followed by addressing a question about their prior knowledge on the reading passage’s topic. Subsequently, they engaged in reading the passage and produced a summary. The next step directed learners to self-assess their own summary using Gao et al (2019) rubric. This self-assessment was guided by feedback either from the technology-enabled SMART system for the technology group or a benchmark expert summary for the non-technology group. After assessing their own summary, participants measured their confidence level regarding their evaluation through a rating confidence question. The concluding part of the self-assessment phase had participants attempt two assessments focusing on reading comprehension and vocabulary knowledge.

For the peer-assessment phase, participants in each were paired, and each individual assessed their counterpart’s summary, using feedback from either SMART or an expert summary. Peer-assessment evaluation utilized the same rubric used in the self-assessment phase. Upon completion, they recorded their confidence in their peer-assessment by responding to the rating confidence question.
It’s important to highlight that the same participants engaged in both self-assessment and peer-assessment within their designated setting. What differentiated each group was the presence or absence of technology-enhanced feedback. Furthermore, during the peer-assessment phase in each setting, participants solely read their peer’s summary, evaluated it using the Gao et al. (2019) rubric, and then responded to the rating confidence question. Notably, during the peer-
assessment phase, participants did not engage in answering the prior knowledge question, reading, and summarizing, or taking the reading and vocabulary tests. These activities were exclusive to the self-assessment phase.

**Data Analysis**

This study focuses on three different dependent variables (DVs), including evaluation scores (RQs 1 and 2 below), evaluation judgment quality (RQs 3 and 4) and rating confidence (RQs 5 and 6). One of the dependent variables (rating confidence) is ordinal, while evaluation scores and evaluation judgment quality are measured on interval scale. Each variable was measured twice, with participants undergoing both the self-evaluation and peer-evaluation conditions.

The data analysis included a random-effect variable (ID/students) and several fixed-effect variables (Assessment type, group/setting, prior knowledge, reading comprehension and vocabulary knowledge). Accordingly, we used a linear mixed-effects model for the interval dependent variables (evaluation scores and evaluation judgment quality) and a mixed-effects ordinal logistic model for the ordinal dependent variables (rating confidence). Table 3.3 provides a detailed description of all the variables used in this study.

**Table 3.3**

*Variables of the Study*

<table>
<thead>
<tr>
<th>Variables</th>
<th>Scale</th>
<th>Variable Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evaluation Score</td>
<td>Interval</td>
<td>Response</td>
</tr>
<tr>
<td>Evaluation Judgment Quality</td>
<td>Interval</td>
<td>Response</td>
</tr>
<tr>
<td>Rating Confidence</td>
<td>Ordinal</td>
<td>Response</td>
</tr>
<tr>
<td>Prior Knowledge</td>
<td>Ordinal</td>
<td>Fixed Effect</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>Interval</td>
<td>Fixed Effect</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>Interval</td>
<td>Fixed Effect</td>
</tr>
<tr>
<td>Group</td>
<td>Nominal</td>
<td>Fixed Effect</td>
</tr>
<tr>
<td>Assessment Type</td>
<td>Nominal</td>
<td>Fixed Effect</td>
</tr>
<tr>
<td>Students (ID)</td>
<td>Nominal</td>
<td>Random Effect</td>
</tr>
</tbody>
</table>
**Linear mixed-effects model.** In order to investigate the research questions related to evaluation score (RQs 1 and 2) and evaluative judgment quality (RQs 3 and 4), the linear mixed-effects model (LMM), also known as a multilevel model, is apt for data that has an interval-dependent variable (Pinheiro & Bates, 2000). The LMM can handle both fixed and random effects.

1. What is the effect of self-assessment compared to peer-assessment on the evaluation score, while accounting for the individual differences among participants?

2. What is the effect of technology versus non-technology settings on the evaluation score, while accounting for the individual differences among participants?

3. What is the impact of self-assessment versus peer-assessment on evaluative judgment quality, considering individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

4. What is the effect of technology versus non-technology settings on evaluative judgment quality, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

Linear Mixed Models (LMMs) are utilized when data is collected from hierarchical or nested data structures, often characterized by random effects in addition to fixed effects (Bates et al., 2015). LMMs account for both fixed factors, which affect the entire population, and random factors, which vary at different levels within the data (Pinheiro & Bates, 2000).

In R, LMMs are primarily implemented using the *lme4* package, which provides functions for fitting and analyzing mixed models. The principal functions of the *lme4* enable the specification of both fixed and random effects (Bates et al., 2015). For hypothesis testing and obtaining *p*-values, which are not provided by *lme4* due to the complexity of mixed model
distributions, the *lmerTest* (Kuznetsova et al., 2017) package is often employed. This package extends *lme4* by adding Satterthwaite’s degrees of freedom method, allowing for the computation of *p*-values for fixed effects in an LMM framework. Mathematically, the model can be represented as follows:

\[ Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_k X_{kij} + u_j + \epsilon_{ij} \]  

(2)

Where \( Y_{ij} \) is the value for dependent variable for the \( i \)-th observation in the \( j \)-th group. \( \beta_0 \) is the overall intercept—the expected mean value of the dependent variable when all predictors are at their reference levels. \( \beta_1, \beta_2, \ldots, \beta_k \) is the fixed effects coefficients for each independent variable \( X_1, X_2, \ldots, X_k \), representing the executed change in the dependent variable per unit change in the predictor variable. \( X_{1ij} + X_{2ij} + \cdots + X_{kij} \) is the observed values of the independent variables (fixed effects) for each observation. \( u_j \) is the random effect for the \( j \)-th group/cluster, capturing the deviation of the group’s intercept from the overall intercept (\( \beta_0 \)). \( \epsilon_{ij} \) is the residual error term for the \( i \)-th observation in the \( j \)-th group/cluster, representing the variability in the dependent variable not explained by the model.

**Mixed-effects ordinal logistic model.** For the ordinal dependent variables (i.e., rating confidence), we utilized mixed-effect ordinal logistic regression. The use of mixed-effects ordinal logistic regression is recommended, particularly for studies involving hierarchical or clustered ordinal data (Agresti, 2010; Gelman & Hill, 2006). It provides a methodological framework that accounts for both fixed effects, such as experimental conditions or treatment groups, and random effects, like individual differences that may not be directly measured but are inherent within the data clusters (Rabe-Hesketh & Skrondal, 2012). The implementation of this analysis in R is facilitated by the ‘ordina’ package (Christensen, 2019), which is utilized for the modeling. Additionally, the ‘MASS’ package (Venables & Ripley, 2002) is employed to check the assumptions of
the ordinal logistic regression model, offering complementary functions for the process. The research question(s) and its related regression equations are given below:

5. What is the effect of self-assessment versus peer assessment on rating confidence, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

6. What is the effect of technology versus non-technology settings on rating confidence, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

In order to answer the two research questions with DVs of evaluation score and rating confidence, a mixed-effects ordinal logistic regression model can be structured in the form of a cumulative logit model, which is used for ordinal dependent variables. The equation for this model is:

\[ \log \left( \frac{P(Y_{ij} \leq k)}{1-P(Y_{ij} \leq k)} \right) = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_p X_{p ij} + u_j \]  

where \( Y_{ij} \) represents the ordinal dependent variable for the \( i \)-th observation in the \( j \)-th cluster (ID/Student). \( k \) denotes the cut-off points on the ordinal scale (for example, for the rating confidence which has levels 1 to 4, \( k \) ranges from 1-4). \( P(Y_{ij} \leq k) \) is the probability that the \( i \)-th observation in the \( j \)-th cluster is at or below the \( k \)-th category of the dependent variable. \( \beta_0 \) is the intercept term, which varies across the levels of the ordinal outcome (not fixed for all levels). \( \beta_1 X_{1ij} + \beta_2 X_{2ij} + \cdots + \beta_p \) are the fixed effects coefficients for the independent variables \( X_1, X_2, \ldots, X_p \) (such as Assessment Type, Group, Prior Knowledge, Reading Comprehension, and Vocabulary Knowledge). \( u_j \) is the random effect for the \( j \)-th cluster (ID/Student), accounting for the correlation of measurement within the same cluster.
4 RESULTS

Descriptive Statistics

Table 4.1 summarizes descriptive statistics for participant performance in self-assessment and peer-assessment in technology and non-technology settings. The table includes data on prior knowledge, evaluation scores, rating confidence, reading comprehension, and vocabulary knowledge. For prior knowledge, the median remained at (Median = 1) for both the technology self-assessment and peer-assessment, indicating a uniform level of self-perceived knowledge amongst participants when engaged with technology. The range of prior knowledge, however, expanded from (Range = 2) in the technology self-assessment to (Range = 3) in the non-technology self-assessment. Evaluation scores demonstrated different magnitude across settings, with a mean of 2.38 (SD = 0.77) for the technology self-assessment setting and a mean of 3.24 (SD = .92) for the non-technology self-assessment setting. A similar pattern is also observed for evaluation score in technology (M = 2.87, SD = 0.97) versus non-technology peer-assessment (M=3.35, SD = 1.04). Mean Evaluative Judgment Quality scores in the technology group for self-assessment yielded a mean of 0.69 (SD = 0.20), slightly lower than peer assessment at 0.75 (SD = 0.17). Conversely, in the non-technology group, self-assessment had a higher mean of 0.76 (SD = 0.17) compared to peer assessment at 0.72 (SD = 0.21). In parallel, rating confidence across all settings consistently reported a median of (Median = 3), but with a narrowed range in non-technology assessments from (Range = 3) to (Range = 2). Reading comprehension and vocabulary knowledge presented subtle differences. Reading comprehension scored a mean of 36.48 (SD = 10.61) for the technology group and a slightly higher mean of 36.98 (SD = 8.26) for the non-technology group, indicating a marginal difference. Vocabulary knowledge displayed a slight decrease in mean scores from the technology to non-technology settings, moving from 74.08 (SD = 12.26) to 72.60 (SD = 13.20), with an increase in standard deviation denoting a broader spread of
scores in the non-technology setting. It must be mentioned that overall, with the exception of the evaluative judgment quality scores, within a group comparison shows that peer assessment scores are either equal or higher than self-assessment scores.

Table 4.1 Descriptive Statistics

<table>
<thead>
<tr>
<th>Measure</th>
<th>Tech</th>
<th>Non-Tech</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Self</td>
<td>Peer</td>
</tr>
<tr>
<td>Prior Knowledge*</td>
<td>1 (2)</td>
<td></td>
</tr>
<tr>
<td>Evaluation Score</td>
<td>2.38 (0.77)</td>
<td>2.87 (0.97)</td>
</tr>
<tr>
<td>Evaluative Judgment Quality</td>
<td>0.69 (0.20)</td>
<td>0.75(0.17)</td>
</tr>
<tr>
<td>Rating Confidence*</td>
<td>3 (3)</td>
<td>3 (3)</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>36.48 (10.61)</td>
<td></td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>74.08 (12.26)</td>
<td></td>
</tr>
</tbody>
</table>

* ordinal measures reporting median (range)

**Evaluation Score**

In addressing the first set of research questions, the results section provides a detailed analysis of the influence of assessment types and group/settings on the dependent variable, ‘evaluation score’. The findings, derived from a robust statistical analysis, reveal the extent to which ‘self-assessment’ and ‘peer-assessment’ differentially impact evaluation scores. Furthermore, the analysis compares ‘technology’ versus ‘non-technology’ settings, offering insights into the role of technology-enabled feedback in influencing evaluation scores. The following research questions guide this section’s results.

1. What is the effect of self-assessment compared to peer-assessment on the evaluation score, while accounting for the individual differences among participants?

2. What is the effect of technology versus non-technology settings on the evaluation score, while accounting for the individual differences among participants?
Figure 4.1 presents the evaluation score of the participants across assessment type and group factors.

From the visual representation, it is evident that there are differences in evaluation scores between the non-technology and technology groups, as well as between self and peer assessments. Participants in the non-technology group rated their peers slightly higher than themselves, with mean scores of 3.34 (SD = 1.04) for peer assessments versus 3.23 (SD = 0.92) for self-assessment. In contrast, the technology group showed a more pronounced difference, with peer
assessments receiving a notably higher mean evaluation score 2.87 (SD = 0.97) compared to mean score for self-assessments of 2.40 (SD = 0.78).

Assumptions of the Linear Mixed-Effects Model

To investigate the research questions, we are utilizing a linear mixed-effects model, which necessitates verifying several underlying assumptions to ensure the validity of our analysis. These assumptions include the linearity of the relationship between predictors and the outcome, homoscedasticity (constant variance of residuals across the range of predicted values), normality of the residuals, independence of residuals (lack of autocorrelation), absence of multicollinearity among predictors, and appropriate distribution of the random effects. Ensuring these conditions are met is crucial for the reliability and accuracy of the model’s results.

Linearity and Homoscedasticity. The relationship between the predictors and the response variable is assumed to be linear. This implies that changes in the predictor variables lead to proportional changes in the expected value of the response variable. In addition, for the assumption of homoscedasticity, the variance of the residuals should be constant across all levels of the predictor variables (Berry and Feldman, 1985).
Heteroscedasticity, or non-constant variance, can lead to inefficient estimates and affect the generalizability of the model results (Tabachnick & Fidell, 1996). The assumptions were checked by plotting the residuals of a model against the fitted values in Figure 4.2. The residual plot provided does not demonstrate a systematic pattern or curvature, implying that the linearity assumption is likely intact (Cohen & Cohen, 1983). Simultaneously, the spread of residuals
displays a consistent pattern across the range of fitted values, which tends to support the homoscedasticity assumption.

**Figure 4.3 Normality of Residuals**

**Normality of residuals.** The residuals (differences between observed and predicted values) are assumed to be normally distributed. This assumption is crucial for the validity of hypothesis tests and confidence intervals (Tabachnick & Fidell, 2000). In assessing the normality of the residuals for the linear mixed-effects model, a quantile-quantile (Q-Q) plot was examined. As shown in Figure 4.3, the Q-Q plot suggests that while the residuals of the model are not
perfectly normal, the deviations may not be severe enough to undermine the overall conclusions drawn from the model.

**Independence of residuals.** The residuals should be independent of each other. This is particularly important in time-series or spatial data, where autocorrelation might violate this assumption. An Autocorrelation Function (ACF) plot for the residuals of the linear mixed-effects model provides a visualization to assess the independence of residuals, which is a critical assumption for linear models (Nobre & Singer, 2007).

![ACF for Independence of Residuals](image)

Figure 4.4 ACF for Independence of Residuals

The residuals from the linear mixed-effects model were assessed for independence using the autocorrelation function (ACF). The analysis revealed that autocorrelation coefficients across
different lags remained within the expected confidence intervals, indicating no significant autocorrelation. This lack of significant autocorrelation suggested that the residuals were, in fact, independent. The absence of any discernible pattern or correlation over time confirmed the model’s compliance with the independence assumption, an essential criterion for the validity of a linear mixed-effects model’s inferences (Figure 4.4).

**Multicollinearity.** While mixed-effects models can handle some level of multicollinearity, excessive correlation between predictor variables can inflate the standard errors of the estimates and make it difficult to assess the effect of individual predictors (Morzuch & Ruark, 1991). The Variance Inflation Factor (VIF) analysis was conducted on the fixed effects within the linear mixed-effects model (Table 4.2). The resulting VIF values indicated minimal concerns regarding multicollinearity for reading comprehension (VIF = 1.536) and vocabulary knowledge (VIF = 1.497), suggesting a mild degree of multicollinearity. Despite these modest elevations, the VIF values did not approach the threshold commonly associated with severe multicollinearity, which typically is a VIF greater than 5 or 10. Consequently, the model’s estimates for these variables are unlikely to be substantially inflated due to multicollinearity.

Table 4.2 Variance Inflation Factor Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Comprehension</td>
<td>1.536087</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>1.496806</td>
</tr>
</tbody>
</table>

**Random effects distribution.** The distribution of random effects was evaluated for normality using the Shapiro-Wilk test. The test result ($W = 0.98, p = .254$) delivered a p-value larger than designated .05. The non-significant p-value ($p > .05$) suggests that the random effects are normally distributed within the linear mixed-effects model.
Results for Linear mixed-effects model analysis for Evaluation Score

In order to answer the research question on the effect of self-assessment versus peer-assessment and technology versus non-technology a linear mixed effect model was developed. The linear mixed-effect model included Assessment Type (self and peer assessment), Group (non-technology vs. technology group), prior knowledge, reading comprehension, and vocabulary knowledge as fixed effects, and a random intercept for ID to account for within-subject variability. The response variable was evaluation score.

The fixed effect of Assessment Type was significant, with an increase in the evaluation score by an estimate of 0.286 (SE = 0.123, t = 2.330, \( p = .022 \)) when comparing peer versus self-assessment. This suggests that assessments categorized as "peer" are associated with a higher evaluation score compared to "self" assessments. The Group effect was also significant; participants in the non-technology group scored higher by an estimate of 0.657 (SE = 0.157, \( t = 4.171, \ p < .001 \)) compared to the technology group. This indicates a notable difference in evaluation scores between the two groups, with non-technology group members receiving higher scores.

The effect of prior knowledge was not significant (Estimate = -0.017, SE = 0.299, \( t = -0.057, \ p = .955 \)). Similarly, the effects of reading comprehension (Estimate = -0.002, SE = 0.123, \( t = -0.019, \ p = .985 \)) and vocabulary knowledge (Estimate = 0.079, SE = 0.121, \( t = 0.654, \ p = .515 \)) on evaluation scores were not significant. The random effects of the model indicated variability in the intercepts across subjects, with a variance of 0.184 and a standard deviation of 0.429. The residual error of the model had a variance of 0.693 and a standard deviation of 0.832.

These results suggest that while Assessment Type and Group have a significant impact on evaluation scores, individual differences measured by prior knowledge, reading comprehension, and vocabulary knowledge do not significantly affect the scores. Additionally, there is
substantial variability in the baseline scores across individuals that is not explained by the fixed
effects in the model.

Table 4.3 Fixed Effects Estimates for Linear Mixed Effects Model

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>2.558</td>
<td>0.175</td>
<td>14.609</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Assessment Type (Peer)</td>
<td>0.286</td>
<td>0.123</td>
<td>2.330</td>
<td>0.022</td>
</tr>
<tr>
<td>Group</td>
<td>0.657</td>
<td>0.157</td>
<td>4.171</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Prior Knowledge</td>
<td>-0.017</td>
<td>0.299</td>
<td>-0.057</td>
<td>0.955</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>-0.002</td>
<td>0.123</td>
<td>-0.019</td>
<td>0.985</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>0.079</td>
<td>0.121</td>
<td>0.654</td>
<td>0.515</td>
</tr>
</tbody>
</table>

Table 4.4 Random Effects Variance for Linear Mixed Effects Model

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>Variance</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID (Intercept)</td>
<td>0.184</td>
<td>0.429</td>
</tr>
<tr>
<td>Residual</td>
<td>0.693</td>
<td>0.832</td>
</tr>
</tbody>
</table>

Finally, regarding the variance explained by the model, the linear mixed-effects model
for evaluation score showed that the fixed effects collectively explained approximately 15.47%
($R^2_m$) of the variance in the evaluation scores. This indicates that the combination of assessment
type, group, prior knowledge, reading, and vocabulary accounted for about 15.47% of the variab-
ility observed in the evaluation scores across different individuals. Furthermore, considering
both fixed and random effects, the entire model explained approximately 33.19% ($R^2_c$) of the
variance in the evaluation scores. This suggests that when incorporating both fixed and random
effects, including individual variability (random effects) in addition to the fixed effects, the
model could account for about 33.19% of the observed variability in the evaluation scores.

**Evaluative Judgment Quality**

For evaluative judgment quality, the impact of self-assessment versus peer-assessment on the
quality of evaluative judgment is investigated, particularly considering individual differences
such as reading comprehension ability, vocabulary knowledge, and prior knowledge levels (research question 3). Additionally, the effect of technology versus non-technology settings on evaluative judgment quality is analyzed, again considering the same individual differences (research question 4). The following research questions are articulated:

3. What is the impact of self-assessment versus peer-assessment on evaluative judgment quality, considering individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

4. What is the effect of technology versus non-technology settings on evaluative judgment quality, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

The evaluative judgment quality of the participants is quantified using a numerical similarity formula, which compares the student’s evaluation with that of experts. This formula yields a similarity measure ranging from 0 (completely dissimilar) to 1 (completely similar), providing a quantitative assessment of the participants’ evaluative skills in relation to expert judgment. Figure 4.4 indicates the descriptive statistics for numeric similarity by group and across both self-assessment and peer-assessment.

Figure 4.5 offers a visual comparison across non-technology and technology groups for self and peer-assessment conditions. Within the non-technology group, the results show a tendency for participants to assess their own performance in closer alignment with expert judgments compared to assessments of their peers. Conversely, within the technology group, the data suggests a different trend: participants tend to assign evaluations to their peers’ summaries that more closely align with expert judgments. In contrast, their assessments of their own performance tend to deviate more from expert evaluations. The descriptive statistics supports the visual inspection
in that the non-technology group, individuals rated the quality of their own summary closer to an expert judgment’s, with a higher mean evaluative judgment quality of 0.76 (SD = 0.17), compared to how they judge their peers’ work, which has a mean quality score of 0.72 (SD = 0.21). This indicates not only a higher self-assessment but also a tighter clustering around that higher mean, suggesting more agreement with expert judgment when judging their own work.

Figure 4.5 Evaluative Judgment Quality

In contrast, within the technology group, peer assessments are viewed as higher in quality, with a mean of 0.75 (SD = 0.17), than self-assessments, which have a lower mean of 0.69 (SD = 0.20). Additionally, the self-assessments in the technology group show a wider spread of scores (as indicated by a higher SD) than in the non-technology group, implying a greater diversity in self-perception of quality.
Assumptions of Linear Mixed Effect Model

To investigate the research questions, we are utilizing a linear mixed-effects model, which necessitates verifying several underlying assumptions to ensure the validity of our analysis. These assumptions include the linearity of the relationship between predictors and the outcome, homoscedasticity (constant variance of residuals across the range of predicted values), normality of the residuals, independence of residuals (lack of autocorrelation), absence of multicollinearity among predictors, and appropriate distribution of the random effects. Ensuring these conditions are met is crucial for the reliability and accuracy of the model’s results.

**Linearity of model predictions.** The assumption of linearity in linear mixed-effects models (LMMs) refers to the expectation that there is a linear relationship between the predictors and the outcome variable. This means that any change in a predictor variable is associated with a proportional change in the response variable. This assumption is foundational to the model’s validity and the accuracy of its predictions. To assess the linearity assumption in linear mixed-effects models, a plot of residuals versus fitted values is utilized. In this diagnostic plot, residuals (the differences between observed and predicted values by the model) are plotted on the y-axis, and the fitted values (the model’s predictions) are on the x-axis. Ideally, for the assumption to hold, the residuals should be randomly distributed around the horizontal axis at zero, indicating a consistent linear relationship across all levels of fitted values. The presence of patterns, such as curves or fan shapes, in this plot suggests systematic deviations from linearity. Figure 4.5 indicates that the residuals appear to be scattered without a clear pattern or systematic structure around the horizontal line, suggesting no obvious violations of the linearity assumption. The random scatter of data points suggests that the linear model’s predictions are consistent across the
range of fitted values, and there is no apparent trend, such as a curve or a funnel shape, which would indicate heteroscedasticity or non-linear relationships.

However, there seems to be a slight concentration of residuals above and below the zero line in certain sections of the plot (notably around the middle range of fitted values. This could suggest minor deviations from perfect linearity, but not necessarily to a degree that would undermine the validity of the model.

Figure 4.6 Residuals versus Fitted Values for Evaluative Judgment Quality
**Homoscedasticity.** For homoscedasticity we expect that the variance of the residuals (the differences between the observed values and the values predicted by the model) be constant across all levels of the predicted values.

![Residual Plot](image)

Figure 4.7 Normality of Residuals for Evaluative Judgment Quality

In the residual plot, we would expect to see the residuals scattered randomly around the horizontal axis (where the residual value is zero) without any discernible pattern. If the residuals fan out or form a funnel shape as the fitted values increase or decrease, it indicates heteroscedasticity, meaning that the variance of the residual’s changes with the level of the predictor variable, violating the homoscedasticity assumption. From the visual inspection of Figure 4.6, the spread of residuals appears to be consistent across the range of fitted values, without any clear pattern.
indicating increasing or decreasing variance. There doesn’t seem to be any signs of a funnel shape (wide at one end and narrow at the other) or other patterns that would suggest heteroscedasticity.

**Normality of residuals:** Normality of residuals in LME models ensures that the statistical inference about fixed effects is valid. This includes hypothesis tests on the coefficients of the fixed effects and the construction of confidence intervals. Since LME models are often used for data that include random effects due to repeated measures on subjects or hierarchical nesting structures, ensuring that the residuals of these models are normally distributed is key for the correct interpretation and validity of the model’s results. A Q-Q plot (Quantile-Quantile plot), where the quantiles of residuals are plotted against the expected quantiles of a normal distribution is used to check the assumptions of the model. If the residuals are normally distributed, the points should fall approximately along a straight line. Figure 4.7 suggests that the residuals from the Linear Mixed-Effects model are approximately normally distributed in the central portion of the data, as indicated by the alignment of points with the reference line. However, there are deviations from normality in the tails; specifically, the lower tail shows more extreme negative residuals, and the upper tail displays some extreme positive residuals than would be expected in a normal distribution. While there is an indication of normality in the central range of data, the deviations in the tails suggest that the residuals might not be perfectly normally distributed.

**Independence of residuals:** Figure 4.8, displays the autocorrelations of the residuals from a statistical model at various lags. The autocorrelation at lag 0 is naturally 1, as this represents the correlation of the series with itself. Crucially, the ACF values for subsequent lags fall within the confidence bounds, indicated by the blue dashed lines, which suggests that there is no significant autocorrelation at those lags. This lack of significant autocorrelation implies that the
residuals can be considered independent, a key assumption for many statistical modeling techniques. The independence of residuals is an important condition for the validity of standard errors and test statistics in regression models. Given that the ACF plot does not exhibit any systematic patterns outside the confidence bounds, it supports the conclusion that the model’s residuals do not violate the assumption of independence and therefore, the model is likely to produce reliable inferences.

Figure 4.8 Autocorrelations of Residuals
**Multicollinearity**: Variance Inflation Factors (VIF) is used to quantify the severity of multicollinearity. It provides a measure of how much the variance of an estimated regression coefficient increases due to collinearity. Generally, a VIF above 5 or 10 indicates a problematic level of collinearity. As shown in Table 4.5, the VIF values for all the variables are well below 5, suggesting that multicollinearity is not a concern for this set of predictors.

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Comprehension</td>
<td>2.656439</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>2.523609</td>
</tr>
</tbody>
</table>

**Random effects distribution.** The distribution of random effects was evaluated for normality using the Shapiro-Wilk test. The test did indicate a significant departure from normality \(W = 0.95, p = .002\), suggesting that the random effects are not normally distributed within the linear mixed-effects model. Although the assumption of the normality of the distribution was not upheld for the model, however, as discussed by West, Welch, and Galecki (2015), the inference regarding fixed effects remains generally reliable even when the normality assumption for random effects is not strictly met, especially in larger samples where the central limit theorem ensures the normality of fixed effect estimates. This perspective is reinforced by the detailed examination of mixed models by Pinheiro and Bates (2000), who highlight the model’s resilience to deviations from normality in random effects, suggesting that such deviations do not critically impair the model’s inferential capabilities regarding fixed effects. This robustness is particularly relevant for the estimates and standard errors of fixed effects, which are often the primary focus in mixed models’ applications. Therefore, while it is essential to consider all model assumptions, small deviations from normality in the random effects distribution may not substantially impact
the overall inference drawn from the model, thus allowing for a certain level of flexibility in practical research applications.

**Results for Linear Mixed Effect Model for Evaluative Judgment Quality**

In the examination of the impact of assessment type, group, and individual differences on evaluative judgment quality, a linear mixed-effects model was utilized. The model, fitted using Restricted Maximum Likelihood (REML) and employing Satterthwaite’s method for t-tests, analyzed the relationship of evaluative judgment quality with assessment type, group, reading, vocabulary, and prior knowledge, including a random intercept for ID to account for within-subject variability.

The result of the linear mixed-effects model (Table 4.6) did not indicate significant fixed effects for assessment type or group on evaluative judgment quality. Specifically, the model showed that assessment type (Estimate = 0.007, SE = 0.025, t = 0.308, p = 0.759) and being in the non-technology group (Estimate = 0.025, SE = 0.033, t = 0.774, p = 0.441) did not have statistically significant impacts on evaluative judgment quality. This suggests that neither assessment type nor group/settings (technology vs. non-technology) significantly alters the evaluative judgment quality, as measured by numeric similarity. Furthermore, the individual difference variables—reading comprehension, vocabulary knowledge, and prior knowledge—also did not significantly affect evaluative judgment quality. Reading (Estimate = -0.003, SE = 0.026, t = -0.126, p = 0.900) and vocabulary knowledge (Estimate = 0.022, SE = 0.025, t = 0.877, p = 0.383) had minimal and non-significant influences. Similarly, prior knowledge showed no significant effects on evaluative judgment quality (Estimate =0.014, SE = 0.063, t = 0.230, p = 0.818). While none of the variables—such as assessment type, group, reading comprehension, vocabulary knowledge, and prior knowledge—met statistical significance criteria, the intercept for the model was
statistically significant (Estimate = 0.733, SE = 0.036, t = 19.844, p < .001). This finding suggests a notable baseline level of evaluative judgment quality across participants, independent of the variables considered in the model.

Table 4.6 Fixed Effects

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Estimate</th>
<th>SE</th>
<th>t</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.733</td>
<td>0.036</td>
<td>19.844</td>
<td>&lt;.001</td>
</tr>
<tr>
<td>Assessment Type</td>
<td>0.007</td>
<td>0.025</td>
<td>0.308</td>
<td>0.759</td>
</tr>
<tr>
<td>Group</td>
<td>0.025</td>
<td>0.033</td>
<td>0.774</td>
<td>0.441</td>
</tr>
<tr>
<td>Prior Knowledge</td>
<td>0.014</td>
<td>0.063</td>
<td>0.23</td>
<td>0.818</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>-0.003</td>
<td>0.026</td>
<td>-0.126</td>
<td>0.900</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>0.022</td>
<td>0.025</td>
<td>0.877</td>
<td>0.383</td>
</tr>
</tbody>
</table>

Finally, the model’s random effects analysis revealed variability in the intercepts across subjects (Variance = 0.009, SD = 0.096), suggesting individual differences in baseline evaluative judgment quality. The residual error variance (0.029) and its standard deviation (0.170) indicate additional variability in evaluative judgment scores not accounted for by the fixed effects.

Table 4.7 Random Effect

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>Variance</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID (Intercept)</td>
<td>0.009</td>
<td>0.096</td>
</tr>
<tr>
<td>Residual</td>
<td>0.029</td>
<td>0.170</td>
</tr>
</tbody>
</table>

Finally, the analysis using the linear mixed-effects model (LME_EJQ) revealed that the fixed effects accounted for approximately 2.71% ($R^2_m$) of the variance in the outcome variable. This suggests that the combination of the predictors included in the model explained about 2.71% of the observed variability in the outcome variable. Moreover, considering both fixed and random effects, the entire model explained approximately 26.36% ($R^2_c$) of the variance in the outcome variable. This indicates that when incorporating both fixed and random effects,
including individual variability (random effects) along with the fixed effects, the model could account for approximately 26.36% of the observed variability in the outcome variable.

**Rating Confidence**

This final section examines research questions aimed at understanding the dynamics of rating confidence. The first question investigates the comparative influence of self-assessment versus peer assessment on the confidence with which individuals rate performance, integrating individual differences in reading comprehension, vocabulary, and levels of prior knowledge as related variables. The second question shifts the focus to the learning environment/group, evaluating how technology-enriched versus non-technology settings affect rating confidence, while again considering the same individual differences factors. The research questions guiding the section are as follows:

5. What is the effect of self-assessment versus peer assessment on rating confidence, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

6. What is the effect of technology versus non-technology settings on rating confidence, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?
Figure 4.9 Rating Confidence

Figure 4.9 shows that in the non-technology group, individuals exhibit a higher mean confidence when assessing their peers (M = 3.277, SD = 0.579) compared to self-assessments (M = 3.000, SD = 0.626). Similarly, in the technology group, the confidence in peer assessments (M = 2.822, SD = 0.747) is also higher than self-assessments (M = 2.622, SD = 0.777). Finally, the standard deviations across all categories demonstrate a relatively consistent spread of rating confidence, with the technology group displaying slightly greater variability.

Assumptions of Ordinal Mixed-Effect Logistic Regression Model

In conducting our analysis using an ordinal linear mixed-effects model to address the research questions, we checked for compliance with key assumptions integral to the model’s integrity and the credibility of our findings. These prerequisites encompass the proportional odds assumption, particularly relevant when the response variable is ordinal, ensuring that the relationship between each pair of outcome groups is consistent across the levels of the predictors. We also verified the random effects distribution, confirming that these effects are normally
distributed, which is fundamental for the model’s random components. Homoscedasticity was assessed to guarantee that the residuals have constant variance across the predicted values, thereby ensuring uniformity in the data’s spread. Lastly, we scrutinized the data for multicollinearity to ensure that the predictors are not highly correlated, as this could undermine the reliability of the parameter estimates.

**Proportional Odds Assumption.** The proportional odds assumption is a key consideration in ordinal logistic regression models, such as the Cumulative Link Mixed Model (CLMM). This assumption posits that the relationship between each pair of outcome categories is the same for all categories. In simpler terms, it means that the effect of the predictor variables on the odds of being in a higher category versus all lower categories is consistent across all thresholds of the ordinal response variable.

In this analysis, the proportional odds assumption was examined using a likelihood ratio test, which compared a restricted model (assuming proportional odds) against an unrestricted model (allowing varying effects across outcome categories). The restricted model, CLMM model, included fixed effects for assessment type, group, reading, vocabulary, and prior knowledge, along with random intercepts for individual differences. The unrestricted model, on the other hand, allowed these effects to vary across the different levels of the response variable. The non-significant result of the likelihood ratio test ($\chi^2 = 3.48, df = 5, p = .626$) suggests that there is no substantial difference between the two models, indicating that the proportional odds assumption likely holds for this dataset. This means that the simpler model with the proportional odds assumption is adequate for describing the relationship between the predictors and the ordinal outcome. Table 4.8 summarizes the result of the model comparison.
Table 4.8 Likelihood Ratio Test Between CLMMs

<table>
<thead>
<tr>
<th>Model</th>
<th>$\chi^2$ (df)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CLMM vs. Unrestricted CLMM</td>
<td>3.48 (5)</td>
<td>0.626</td>
</tr>
</tbody>
</table>

**Random Effects Distribution.** The assumption of normally distributed random effects in mixed-effect models, including ordinal mixed-effect models, is crucial for several reasons. It affects the interpretation of the model, the estimation of the parameters, and the validity of any conclusions drawn from the model. The nature of the distribution can be checked in several ways including visual inspections and using Shapiro-Wilk test of normality of distribution. The Shapiro-Wilk test was conducted to assess the normality of the random intercepts in the mixed-effects model. The test yielded a W statistic of 0.98646 ($p = 0.4632$), indicating that the distribution of random intercepts closely approximates a normal distribution. This result suggests that there is no significant deviation from normality in the distribution of the random intercepts. Therefore, the assumption of normality for the random intercepts in this mixed-effects model is not violated.

**Homoscedasticity.** The variance of the residuals should be consistent across different levels of the predictor variables. In mixed models, this also extends to the random effects, where the variance should be consistent across groups or clusters. In assessing homoscedasticity, or the assumption of equal variances, one would expect to see the residuals spread randomly around the horizontal line at zero, with no systematic pattern in the spread as the fitted values increase or decrease.
Figure 4.9 indicates that the variance of the residuals remains consistent across the range of fitted values. There is no obvious pattern, such as a funnel shape where the spread of residuals increases with the fitted values (indicative of heteroscedasticity), or any other pattern suggesting non-constant variance. This visual assessment suggests that the assumption of homoscedasticity may not be violated in the model from which these residuals were derived.

**Multicollinearity.** The Variance Inflation Factor (VIF) is used to detect the presence of multicollinearity among predictor variables in regression models. Multicollinearity occurs when two or more predictors are correlated with each other, which can inflate the variance of the estimated regression coefficients and may lead to less reliable statistical inferences. Developing a correlation matrix shows that a substantial correlation exists between reading and vocabulary ($r = 0.77072$), indicating a strong linear relationship between the two predictor variables. Despite this
strong correlation, the VIF values for both reading and vocabulary do not exceed 5 or 10 indicators of problematic level of multicollinearity. With VIF values below these conventional thresholds, there isn’t a strong indication of multicollinearity adversely impacting the model estimates.

Table 4.9 Variance Inflation Factor Results

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading Comprehension</td>
<td>2.46311</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>2.46311</td>
</tr>
</tbody>
</table>

**Other Assumptions of Ordinal Mixed-Effect Models**

In addition to the standard assumptions that underpin mixed models, two further considerations are crucial for robust analysis: the nature of missing data and the choice of link function. Mixed models have the capability to manage missing data, provided the missingness occurs at random (MAR). This implies that the probability of a data point being missing is not related to the missing data itself or the outcome of the study (Yang & Maxwell, 2014). If data is missing not at random (MNAR), it can introduce bias into the results, potentially compromising the validity of the model’s inferences. For the current model, no missing data was reported (Yang & Maxwell, 2014).

Another critical assumption pertains to the selection of an appropriate link function. Selecting an appropriate link function is a fundamental consideration in ordinal mixed models and serves as a crucial assumption for the analysis of data with ordinal outcomes (Lindstrom & Bates, 1990). The link function is the component of the model that relates the linear predictor to the mean of the response variable. In the context of ordinal outcomes, where the response variable is categorical with an inherent order but no consistent scale, the link function is essential for mapping predictors to an ordered probability scale. Common choices for the link function include the logit and probit functions. We chose the logit link, because it is frequently used in
logistic regression models and is often preferred for its interpretability, especially when the outcome odds can be assumed to increase linearly with the predictor variables (Menard, 2010).

**Results for Ordinal Mixed-Effect Logistic Regression Model for Rating Confidence**

**Fixed Effects**

The output of the model indicates that key predictors in the ordinal mixed effect logistic regression model show varying levels of influence on rating confidence. Assessment type was a significant predictor, with peer assessment associated with an increase in the log-odds of higher rating confidence compared to self-assessment ($\beta = .920, SE = 0.33, z = 2.791, p = .005$). Additionally, participants in the non-technology group exhibited higher rating confidence compared to those in technology group ($\beta = 1.659, SE = 0.511, z = 3.246, p = .001$). Also, reading ($\beta = 0.311, SE = 0.371, z = 0.840, p = 0.401$), vocabulary ($\beta = 0.131, SE = 0.368, z = 0.357, p = 0.720$), and prior knowledge ($\beta = 0.282, SE = 0.875, z = 0.328, p = 0.742$), were not found to significantly predict rating confidence.

**Table 4.10 Fixed Effects**

<table>
<thead>
<tr>
<th>Fixed Effects</th>
<th>Estimate</th>
<th>SE</th>
<th>z</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assessment Type (Peer)</td>
<td>0.920</td>
<td>0.329</td>
<td>2.791</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Group (Non-Technology)</td>
<td>1.659</td>
<td>0.511</td>
<td>3.246</td>
<td>&lt; .001</td>
</tr>
<tr>
<td>Prior Knowledge</td>
<td>0.282</td>
<td>0.875</td>
<td>0.328</td>
<td>0.742</td>
</tr>
<tr>
<td>Reading Comprehension</td>
<td>0.311</td>
<td>0.371</td>
<td>0.840</td>
<td>0.401</td>
</tr>
<tr>
<td>Vocabulary Knowledge</td>
<td>0.131</td>
<td>0.368</td>
<td>0.357</td>
<td>0.720</td>
</tr>
</tbody>
</table>

Finally, The beta coefficient values in the represent the log odds of moving from one rating category to the next. Specifically, for Assessment Type, a beta value of 0.92 translates to an odds ratio of approximately 2.51, indicating that transitioning from "self" to "peer" assessments increases the odds of a higher rating by about 2.51 times. Similarly, for Group, a beta value of 1.66 corresponds to an odds ratio of roughly 5.25, signifying a substantial increase in the odds of
higher ratings when moving from a "technology" to a "non-technology" group setting. However, other predictors like reading, vocabulary, and prior knowledge did not demonstrate significant coefficients.

**Random Effect**

The analysis revealed significant individual variability in rating confidence, as evidenced by the random effects for the ID grouping factor. The variance of these random intercepts was 2.537, with a corresponding standard deviation of 1.593, indicating notable differences in baseline rating confidence across individuals.

Table 4.11 Random Effects

<table>
<thead>
<tr>
<th>Random Effects</th>
<th>Variance</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID (Intercept)</td>
<td>2.537</td>
<td>1.593</td>
</tr>
</tbody>
</table>

**Threshold Coefficients**

The model’s threshold coefficients, which delineate the boundaries between ordinal categories of rating confidence, were significant, indicating well-defined separations between these categories. Specifically, the estimates for thresholds 1|2, 2|3, and 3|4 were -3.8736 (SE = 0.8094, z = -4.786), -0.4899 (SE = 0.5301, z = -0.924), and 3.5030 (SE = 0.6555, z = 5.344), respectively.

Table 4.12 Model Threshold

<table>
<thead>
<tr>
<th>Threshold</th>
<th>Estimate</th>
<th>Std. Error</th>
<th>z value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>-3.8736</td>
<td>0.8094</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>-0.4899</td>
<td>0.5301</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>3.503</td>
<td>0.6555</td>
</tr>
</tbody>
</table>
Model Fit

The model’s fit was indicated by an Akaike Information Criterion (AIC) of 379.82 and a log-likelihood of -178.91. The model’s convergence was successfully achieved (max.grad = 1.20e-04, cond.H = 9.6e+01), suggesting the estimation process was stable and reliable.
5 DISCUSSION

Research Questions 1 & 2: Evaluation Score

To interpret the findings, it is essential to revisit the assessment types and the contextual differences between technology and non-technology settings. In both self-assessment and peer-assessment conditions, learners utilized the same rubric (Gao et al., 2019) to evaluate summaries derived from a similar reading passage, whether their own or their peers’. In the non-technology group, participants had access to an expert summary of the reading passage. This expert summary offered a concise, easily understandable example for learners to benchmark their own summaries against. While serving as a valuable reference point, the expert summary did not offer any additional tools or guidance to aid participants in making evaluative judgments. For the technology group, in contrast to the expert summary, SMART feedback provides tailored evaluative judgment information for individual learner summaries. This feedback mechanism incorporates various modalities, including the learner’s knowledge structure, expert knowledge structure, key concepts and their relations, a list of missed key concepts, and a progress bar. By utilizing these diverse elements, SMART feedback delivers detailed and targeted assessments of learners’ summaries, focusing on specific areas of strength and improvement all without offering an expert summary for comparison. The following research questions were articulated for evaluation score as a dependent variable.

1. What is the effect of self-assessment compared to peer assessment on the evaluation score while accounting for the individual differences among participants?

2. What is the effect of technology versus non-technology settings on the evaluation score, while accounting for the individual differences among participants?
This research question set aimed to investigate the effect of type of assessment and group/setting on evaluation scores, after controlling for individual differences. For these research questions, the anticipated hypothesis, premised on works by Panadero et al. (2016) and Shore et al. (1992), posited that technology-supported evaluation whether self or peer would yield higher scores. However, the findings contradicted this hypothesis. Using R’s lme4 package, a linear mixed-effects model was fitted to evaluate the effect of assessment type, group, prior knowledge, reading comprehension, and vocabulary knowledge as fixed effects, and a random intercept for ID to account for within-subject variability, with evaluation score as the response variable. The results indicated significant effects for assessment type and group. Peer assessment was associated with an increase in evaluation scores, and participants in the non-technology group tended to give higher scores to self and peer than their counterparts in the technology group. The effects of prior knowledge, reading comprehension, and vocabulary knowledge were not statistically significant.

The findings of the current research questions suggest that participants in the non-technology setting tended to assign higher grades to both their own and their peers’ summaries. This inclination could be attributed to the influence of social-relational factors. The role of social-relational factors in influencing grading during peer assessment is supported by various studies in literature. Falchikov and Goldfinch (2000) observed that students, influenced by leniency and social dynamics, tended to grade their peers higher than themselves during peer assessments. Additionally, empathy and compassion towards peers, particularly in high-pressure academic environments, may contribute to a leniency effect in grading, as suggested by Hanrahan & Isaacs (2001). Similarly, Strijbos and Sluijsmans (2010) acknowledged the learning benefits of peer assessment but also recognized biases such as overrating peer performance. Studies by Lin et al. (2001) and
Chang et al. (2012) found that learners engaged in peer assessment were more likely to assign higher grades compared to self-assessment scenarios. Moreover, the anonymity of both raters and those being rated played a significant role in peer assessment, leading to more liberal judgments and a tendency to overrate peer performance, as indicated by Bloom & Hautaluoma (1987).

Conversely, in the technology group, where SMART feedback provided comprehensive performance evaluation information, providing criterion-based evaluation, participants may have felt more constrained in the range of grades they could assign and thus adopted a more conservative evaluation approach. The phenomenon where raters become more conservative in their scoring due to the performance evaluation information provided by technology-enabled feedback systems like SMART (Kim et al., 2019) can be explained by several factors. Firstly, when individuals receive comprehensive and detailed feedback on their performance, they might become more aware of the specific criteria and standards against which their work is being evaluated (Hattie & Timperley, 2007). This increased awareness might lead to a greater sense of accountability and scrutiny in the evaluation process, prompting participants to adopt a more cautious and conservative approach in assigning grades. Moreover, technology-enabled feedback systems often provide objective and data-driven assessments, which may influence participants to adhere more closely to predetermined standards and benchmarks (Sadler, 2010). Black and William (2009) emphasize how clear guidelines and criteria in feedback can act as crucial anchors for raters. Such well-defined benchmarks minimize the variability and subjectivity inherent in evaluative judgments, promoting consistency and fairness in the assessment process. By offering a concrete standard against which to measure student work, these guidelines encourage raters to
adopt a more conservative and reflective grading approach, ensuring that evaluations are more closely aligned with established educational objectives and learning outcomes.

Sadler (1989) aligns with Black and William’s perspective on the potential role of explicit assessment criteria, arguing that such clarity is fundamental for fostering accurate evaluative judgments. He posits that well-defined criteria demystify quality standards for learners, thereby reducing subjectivity and enhancing the consistency of assessments. Sadler’s work underscores the necessity of transparent criteria in feedback to facilitate targeted improvements and learner self-regulation. Similarly, Gregory et al. (2020) explores the impact of providing clear, structured feedback in educational settings, highlighting its potential to guide student self-assessment and reflection. Their research suggests that when feedback is precise and aligned with well-defined criteria, it can significantly influence how students perceive and evaluate their own work, leading to more consistent and objective self-assessments. This aligns with the idea that clear guidelines in feedback can reduce subjectivity and promote a more conservative grading approach by anchoring student evaluations to specific standards.

In summary, the findings show that learners grade themselves and their peers across technology-based and non-technology settings differently. However, without a standard measure of quality, higher grade does not necessarily mean better evaluations. While we observe different grading patterns across settings, we cannot say they lead to better evaluations. To understand the quality of these evaluations, it’s crucial to compare student ratings with those from experts, as investigated in research questions 3 and 4.

**Research Questions 3 & 4: Evaluative Judgment Quality**

For evaluative judgment quality, the following research questions investigated the effects of self-assessment versus peer-assessment and technology versus non-technology settings on
evaluative judgment quality, taking into account individual differences such as reading comprehension, vocabulary knowledge, and prior knowledge. The following research questions were articulated for that purpose:

3. What is the impact of self-assessment versus peer-assessment on evaluative judgment quality, considering individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

4. What is the effect of technology versus non-technology settings on evaluative judgment quality, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

The evaluative judgment quality of the participants was quantified using a numerical similarity formula, which compares the participants’ evaluation with that of experts’ evaluation. This formula yields a similarity measure ranging from 0 (completely dissimilar) to 1 (completely similar), providing a quantitative assessment of the participants’ evaluative skills in relation to expert judgment.

For the research question set we hypothesized that controlling for individual differences among learners, the type of assessment would significantly influence the quality of evaluative judgment in both technology-related and non-technology-related settings. It was hypothesized that when learners engage in peer assessment, both with and without technology, they will exhibit higher quality evaluative judgment than when they engage in self-assessment, regardless of the technological context. Our hypothesis was guided by several studies (Cowan, 2010; Gielen, et al., 2011; Xiao & Gu, 2021) that indicate peer assessment encourages critical thinking and self-reflection, enhancing the depth and quality of evaluative judgment. Additionally, Nicol and Macfarlane-Dick (2006) suggest that peer assessment provides opportunities for learners to
evaluate others’ work from multiple perspectives, contributing to a more comprehensive assessment of the content. However, in self-assessment, learners might be more prone to overlooking gaps in their understanding, leading to potentially lower quality evaluative judgment outcomes.

Furthermore, we hypothesize that participants receiving technology-supported feedback will demonstrate higher evaluation quality compared to those relying on non-technology feedback, as the former provides more comprehensive and tailored evaluation information. This hypothesis is grounded in the idea that technology-supported feedback systems, such as SMART (Kim et al., 2019), offer more detailed and personalized evaluation information compared to traditional feedback. Research by Hattie and Timperley (2007) suggests that feedback is most effective when it provides specific information about the task, clarifies goals, and offers actionable strategies for improvement. Technology-enabled feedback systems often excel in providing detailed and individualized feedback (Winstone & Carless, 2019), allowing learners to identify areas of strength and weakness in their work more effectively. Moreover, technology-supported feedback systems have been shown to promote deeper engagement with the evaluation process by encouraging learners to reflect on their own work and actively seek out feedback (Carless and Boud, 2018). This increased engagement may lead to greater learning gains and improved evaluation quality (Nicol and Macfarlane-Dick, 2006). Conversely, traditional exemplars, although valuable as reference points, may lack the specificity and customization necessary to guide learners in evaluating their own work and that of their peers effectively.

Descriptive statistics of the evaluative judgment quality scores show a mixed result. Within the non-technology group, the results show a tendency for participants to assess their own performance in closer alignment with expert judgments compared to their assessments of peers.
Conversely, within the technology group, the data suggests a different trend: participants tend to assign evaluations to their peers’ summaries that more closely align with expert judgments.

The linear mixed effect model of evaluative judgment quality revealed that neither assessment type nor the group/setting had a significant effect on the evaluative judgment quality. Individual difference variables such as reading comprehension, vocabulary knowledge, and prior knowledge were also found to have no significant impact on the quality of evaluative judgments. While none of the variables—such as assessment type, group, reading comprehension, vocabulary, and prior knowledge—met statistical significance criteria, the intercept for the model was statistically significant. This finding suggests a notable baseline level of evaluative judgment quality across participants, independent of the variables considered in the model.

While the effect of the assessment type, group and individual differences factors were not significant, several accounts for the observed descriptive findings are possible. In the non-technology group, participants were presented with an exemplar, a common best practice aimed at providing learners with a benchmark against which to compare their own work (Handley & Williams, 2011; Hendry et al., 2011). This exemplar might have served as a model of what constituted a well-crafted summary, offering participants a reference point for evaluating the quality of their own summaries as well as those of their peers (Bouwer et al., 2018; Carless et al., 2018; Dixon et al., 2020).

The higher evaluative judgment quality observed for self-assessments compared to peer assessments may be attributed to the self-referencing effect (Rogers, Kuiper, & Kirker, 1977). The phenomenon suggests that individuals tend to process information more deeply and recall it better when it is related to themselves. Consequently, individuals may engage in more critical
and quality-focused self-assessment, as they are inherently more familiar with their own work (Dunning, Heath, & Suls, 2004).

Furthermore, while the exemplar provided a best practice model, it lacked specific evaluative judgment references tailored to the summary being assessed (Sadler, 1983). Without clear evaluative criteria, participants may have relied more heavily on subjective judgments and personal interpretation when comparing peers’ summaries to the exemplar (Taras, 2001). This subjective evaluation process could introduce inconsistencies in the assessment, as participants may interpret the exemplar differently and apply varying standards (Hattie & Timperley, 2007) as observed in the ratings assigned to peer summaries (Figure, 4.5).

In contrast, participants in the technology group, facilitated by the comprehensive evaluation information provided by SMART feedback, were presented with direct evaluations of their summaries. However, interpreting multimodal feedback, particularly when driven by technology-enabled systems’ evaluation algorithms, may require a higher learning demand. This observation is supported by several previous studies. Wulf and Shea (2002) discuss how augmented feedback can facilitate learning but also emphasize the importance of information structure or how this feedback is presented to and integrated by learner. Similarly, Raaijmakers et al. (2019) explores the cognitive load imposed by different types of feedback and suggested that overly complex feedback can hinder rather than help performance. Likewise, Martin (2020) suggests that while multimodal feedback can enhance learning opportunities, students may struggle with the varying formats and the depth of feedback provided across different mediums. This complexity requires students to adapt to different modes of communication, which can potentially hinder their ability to effectively utilize feedback. The findings in self-assessment and peer-assessment align with the observations in the literature. For the technology group, in the initial self-evaluation task,
participants might not have fully grasped the SMART feedback’s significance or how to effectively apply the provided information. This could stem from unfamiliarity with the system’s feedback mechanisms or the complexity of the feedback itself, requiring a learning period to interpret and utilize the feedback effectively. As participants become more acquainted with the feedback system and its features, their ability to understand and leverage the information for self-evaluation is likely to improve.

In the context of evaluative judgment, the act of critically evaluating one’s own work before assessing others can lead to a deeper understanding and a more accurate evaluation of peer performance. This process, often referred to as the learning effect, suggests that individuals who engage in self-reflection and self-assessment are likely to apply the same critical lens when evaluating their peers’ work. This heightened awareness and understanding of the assessment criteria can contribute to more objective and fair peer evaluations. The observation that the sequence of self-assessment and peer-assessment enhances the accuracy of evaluative judgments due to a learning effect is corroborated in the literature. Cho and MacArthur (2010) and Topping (2009) highlight that multiple rounds of evaluations refine evaluative skills, and lead to more precise assessments. This is further supported by Van Zundert et al. (2010) and Gielen et al. (2010), who argue that such sequential evaluation processes foster critical thinking and a deeper understanding of quality standards among learners (Kaufman & Schunn, 2011), and significantly contribute to learners’ ability to critically evaluate work.

Furthermore, the non-significant findings observed in the research questions on evaluative judgment quality can be attributed to several factors. Firstly, evaluative judgment is a complex construct that involves the ability to make decisions about the quality of work of oneself and others (Lin-Siegler et al., 2015; Tai et al., 2018). It is influenced by numerous cognitive and
affective factors, and its development might require more than just the context of assessment or the medium of feedback. The non-significant findings might indicate that evaluative judgment is more deeply rooted in individual cognitive and affective processes (Sadler, 1998) than previously thought, and it might require a long-term development because it is a higher-order socio-affective and cognitive capability (Ecclestone, 2001; O’Donovan et al. 2004).

Second, the impact of technology-supported feedback on learning is not uniform and can vary significantly based on how learners interact with and interpret the feedback provided. Wisniewski et al. (2020) point out that the relationship between feedback and learning outcomes is complex and influenced by the individual’s engagement with the feedback content. When learners encounter feedback, their perceptions, and the way they implement the suggestions can greatly affect the feedback’s effectiveness. Some learners might find the feedback insightful and use it to make substantial improvements, while others might misunderstand or disregard the feedback, leading to minimal changes in their learning process. Narciss (2013) expands on this by discussing the concept of feedback in learning environments, emphasizing that the way feedback is designed and delivered can influence its effectiveness. Feedback that is not tailored to the learner’s current level of understanding or that fails to provide clear, actionable steps for improvement might not lead to the desired learning outcomes. Additionally, the learner’s ability to self-regulate and actively engage with the feedback process plays a crucial role in determining the effectiveness of the feedback. Vandewaetere and Clarebout (2014) further elaborate on the importance of learner characteristics in the feedback process. They argue that individual differences such as prior knowledge, motivation, and learning strategies can influence how feedback is perceived and utilized.
Third, as acknowledged earlier, the participants in the current study were not homogeneous in terms of their background. Therefore, the heterogeneity in the participant group regarding their prior experience with assessment, familiarity with technology, and engagement with the task could contribute to the non-significant findings (Hattie, 2009; Tomlinson, 2017) because such individual differences can have a profound impact on how learners respond to educational interventions (Eccles & Wigfield, 2020). In this case, the variability within the participant group might have obscured potential effects of the assessment type and group.

In conclusion, the investigation into the impact of self-assessment versus peer-assessment and technology versus non-technology settings on evaluative judgment quality, while considering individual differences has yielded non-significant results yet insightful trends. The hypothesis that peer assessment and technology-supported feedback would enhance evaluative judgment quality more effectively than self-assessment and non-technology counterparts was not supported by the data. However, these results do not diminish the complexity and importance of developing high-quality evaluative judgment. The lack of significant differences highlights the intricate nature of evaluative judgment as a cognitive skill deeply rooted in individual cognitive and affective processes, which may not be easily influenced by short-term, one-time interventions. This complexity is further compounded by the variation in learners’ engagement and perception of feedback, combined with diverse backgrounds, experiences, and levels of task engagement with the evaluation information and the evaluation task in general.

Research Questions 5 & 6: Rating Confidence

For rating confidence, a research question explored the influence of self-assessment compared to peer-assessment on the level of confidence individuals hold in their ratings, also considering individual differences such as reading comprehension skills, vocabulary knowledge, and
prior knowledge (research question 5). Another research question focused on the effect of the learning environment—whether it is technology or non-technology—on rating confidence, with the same individual differences in mind (research question 6). The two research questions guiding this investigation into rating confidence are articulated as follows:

1. What is the effect of self-assessment versus peer-assessment on rating confidence, while accounting for individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

2. What is the impact of technology-enriched versus non-technology environments on rating confidence, while considering individual differences in reading comprehension ability, vocabulary knowledge, and prior knowledge levels?

We hypothesized that after accounting for individual differences among learners, there will be a significant effect for the type of assessment (self vs. peer) and the presence of technology (technology vs. non-technology) on rating confidence. Learners participating in self-assessment, especially with technology support, will demonstrate higher rating confidence compared to those involved in peer assessment, regardless of the technological context. Research by Panadero et al. (2016) suggests that learners tend to exhibit higher levels of confidence in self-assessment compared to peer assessment. This could be attributed to learners’ familiarity with their own work and a potential overestimation of their own abilities. Furthermore, McCarthy (2017) highlights the potential of technology to provide immediate and consistent feedback, which can contribute to higher rating confidence among learners participating in technology-supported self-assessment scenarios. The results of the study, however, did not support the hypotheses.

The descriptive statistics indicated that participants in the non-technology context displayed higher confidence during peer evaluations as opposed to evaluating their own work. This
trend was also evident in the technology group, where peer assessment confidence exceeded that of self-assessment, with a notable difference between the two. The spread of confidence ratings across all conditions was fairly consistent, although those in the technology setting showed a broader range of confidence levels.

The findings on rating confidence revealed that the type of assessment was a significant predictor. Specifically, peer assessments were associated with an increase in the likelihood of higher rating confidence compared to self-assessments. Additionally, being in the non-technology group was linked with higher rating confidence as opposed to being in the technology group. However, individual differences such as reading comprehension, vocabulary, and prior knowledge did not significantly influence rating confidence, indicating that these cognitive and knowledge-based factors might not be as crucial in determining how confident individuals are in their ratings. Moreover, the analysis showed a significant individual variability in rating confidence. This suggests that regardless of the group or type of assessment, there are notable individual differences in baseline confidence levels when it comes to rating confidence.

The model’s threshold coefficients, which define the ordinal categories of rating confidence, were found to be significant. This implies clear distinctions between different levels of rating confidence, reinforcing the model’s capacity to differentiate between varying degrees of confidence. Finally, the model fit indices, including the Akaike Information Criterion (AIC) and the log-likelihood, along with the successful convergence of the model, indicate that the model estimation was stable and reliable, providing a trustworthy analysis of the factors affecting rating confidence.

The significant difference in rating confidence observed between the self-assessment and peer-assessment in both non-technology and technology groups raises several arguments about
the psychological and pedagogical underpinnings of assessment practices in these settings. In case of higher rating confidence in peer-assessment it might be argued that the higher rating could be due to learning effect (Becker & Rosen, 1992; Sun et al., 2017). The learning effect refers to the improvement in performance or knowledge retention through repeated exposure or practice (Emeny et al., 2021; Vlach & Sandhofer, 2012). It could be that the sequence of self-assessment followed by peer assessment likely engendered a learning effect, thereby enhancing participants’ confidence in their evaluative judgments. This phenomenon aligns with findings by Roediger and Karpicke (2006), who found that retrieval practices, such as tests, can enhance long-term memory and learning more effectively than additional study. This suggests that the act of self-assessment, serving as a form of retrieval, might solidify participants’ understanding and judgment criteria, subsequently applied in peer assessment. Butler and Winne (1995) also highlight the role of self-monitoring in assessments, suggesting that self-assessment can improve metacognitive awareness, further supporting the learning effect argument. The learning effect is further supported by the work of Larsen et al. (2009), who emphasize the durability of learning through testing, and Bangert-Drowns et al. (1991), who highlight the positive effects of frequent testing on students’ academic performance, underpinning the argument that the practice of self-assessment can enhance metacognitive awareness and rating confidence.

In non-technology settings, the reliance on exemplars without explicit evaluative criteria may have led to overconfidence due to the subjective nature of the judgments made. Tversky and Kahneman (1974), suggests that people rely on instant exemplars that come to mind when assessing a given subject, perception, procedure, or outcome. Without explicit criteria to guide the evaluation process, individuals may overestimate their abilities, or the quality of their work based on readily available examples or exemplars, leading to overconfidence. Nelson and Narens
(1990) believe that effective self-monitoring allows individuals to evaluate their cognitive processes against some standard, while control mechanisms enable them to regulate these processes to improve learning and performance. In the absence of clear evaluative criteria, individuals may rely on flawed internal standards, leading to inaccurate self-assessments and overconfidence.

Regarding the findings in the technology group, the introduction of detailed feedback in technology settings, through SMART, appears to induce a cognitive conflict that may have affected confidence levels. Detailed, objective feedback challenges learners’ self-efficacy, as highlighted by Kluger and DeNisi (1996) and Narciss (2008), leading to a more cautious and conservative evaluative approach. The impact of detailed feedback is further explored by Shute (2008), who discusses its role in enhancing learning and motivation, and by Yeager et al. (2014), who emphasize the importance of feedback for growth mindset development, suggesting that while such feedback promotes learning, it might also introduce more certainty in evaluations, thus affecting confidence. The criterion-referenced nature of SMART evaluation demonstrated by explicit criteria via technology-based feedback introduces a level of conformity that may lead to assurance and adjustment in evaluative judgments. The assurance is created through exposure to the criterion-referenced nature of SMART evaluative feedback. The concept of criterion-referenced evaluation focuses on assessing students’ performance against specific criteria or standards, rather than comparing their performance to that of their peers, as in norm-referenced assessment (Burton, 2006; Lok et al., 2016; Prince, 2016). Exposure to criterion-referenced evaluation with explicit evaluation criteria can significantly influence learners’ confidence (Dun et al., 2002; Pui et al., 2021), as it provides a clear framework for understanding expectations and gauging personal progress. This type of assessment is considered to offer higher reliability, validity, and transparency, fostering a more objective and constructive learning environment where
students can focus on meeting set standards rather than competing with one another (Bond, 1996; Burton, 2006; O’Donovan et al., 2001). The explicitness of criteria can also introduce a level of conformity and comparison, which might lead to adjustment of learners’ confidence.

In summary, the investigation of rating confidence in relation to assessment type and technology provides intriguing insights into the dynamics of evaluative processes in educational settings. Despite initial hypotheses suggesting higher confidence in self-assessments, particularly with technology support, the findings reveal a complex picture where peer assessments, regardless of technological context, tended to boost rating confidence more significantly. This counterintuitive result underscores the multifaceted nature of assessment practices and their psychological underpinnings. The learning effect, attributed to the practice of self-assessment followed by peer evaluation, might play an essential role in enhancing rating confidence, suggesting that sequential engagement in assessment activities fosters a deeper understanding and more accurate self-reflection. Additionally, the study’s findings highlight the importance of clear evaluative criteria in adjusting learner confidence. Criterion-referenced evaluations, especially those facilitated by technology, offer a structured framework that aids in objective self-assessment, potentially adjusting confidence levels stemming from subjective judgment.

**Conclusions**

*Evaluation Score*

The research into the effects of self-assessment versus peer-assessment and the influence of technology versus non-technology learning environments on evaluation scores revealed interesting findings. Contrary to the initial hypothesis informed by Panadero et al. (2016) and Shore et al. (1992) research, which posited the superiority of technology-supported evaluations in enhancing evaluation scores, the empirical findings of this study presented an intriguing counter-
narrative. Peer assessments, irrespective of the technological context, were consistently associated with higher evaluation scores. This unexpected outcome prompts a reevaluation of the value assigned to technology in educational assessments and underscores the role of peer-assessment dynamics in the evaluative process. Furthermore, the lack of significant influence of individual differences, such as reading comprehension, vocabulary knowledge, and prior knowledge, on evaluation scores in this study suggests a minimal impact of these cognitive factors in the context of the assessment type and learning environments in this study. This revelation calls for further research on individual cognitive abilities in predicting assessment outcomes and highlights the need for a broader understanding of the factors that contribute to evaluation scores in educational settings.

**Evaluative Judgment Quality**

The investigation of evaluative judgment quality sought to unravel how self-assessment compares to peer-assessment and how technology-enhanced environments stack against non-technology settings in aligning participants’ evaluations with those of experts. Grounded in the anticipation that peer assessments and technology-supported feedback might foster higher quality evaluative judgments, the study ventured into this investigation expecting to find significant differences attributable to the type of assessment and the learning environment. However, the results did not corroborate these expectations. The analysis revealed that neither the type of assessment nor the presence of technology significantly impacted the quality of evaluative judgments. This outcome suggests that evaluative judgment quality may be deeply rooted in more complex cognitive and affective processes that are not readily influenced by external factors such as assessment modality or technological interventions. The lack of significant results could suggest that the capacity for evaluative judgment may be more intricately connected to
deeper cognitive and emotional aspects of individuals (Sadler, 1998) than was initially believed. This might imply that the development of evaluative judgment, as a complex blend of social, emotional, and cognitive skills (Ecclestone, 2001; O’Donovan et al., 2004), unfolds over a longer period.

**Rating Confidence**

Finally, the investigation into rating confidence brought to light unexpected findings that contradicted the established hypotheses, particularly those influenced by the research of Panadero et al. (2016) and McCarthy (2017). Despite the anticipation that self-assessments, especially those augmented by technology, would engender higher rating confidence, the study unveiled a discernible preference for peer-assessment in boosting confidence levels across both technology and non-technology environments. These findings challenge the commonly held belief in the confidence-boosting potential of immediate, technology-supported feedback. The preference for peer-assessment in enhancing rating confidence underscores the complex interplay of social dynamics, cognitive processes, and the learning environment in shaping rating confidence. It invites a reconsideration of the factors that foster confidence in assessment settings and highlights the influential role of peer interactions in evaluation.

In summary, this dissertation contributes to a deeper understanding of assessment practices in educational contexts by challenging conventional wisdom about the role of technology, the nature of self versus peer-assessment, and the determinants of evaluative outcomes such as evaluation scores, evaluative judgment quality, and rating confidence. The findings underscore the multifaceted interplay between assessment types, learning environments, and individual differences, suggesting that these elements might combine in intricate ways to influence evaluative judgment. The absence of significant differences in the findings of the
current study highlights the need for ongoing research to dissect the complexities of evaluative practices and their implications for learners.

**Implications**

The implications of the findings from the research on evaluation scores and the interplay between self-assessment and peer-assessment, as well as technology-enhanced versus non-technology settings, present several key considerations for educators and curriculum designers. Firstly, the variability in evaluation scores associated with peer-assessment, especially in non-technology contexts, indicates that both the evaluation task and evaluation context influence how learners evaluate their own and their peer performance. This suggests a need for educators to carefully consider the methodology, contexts, design, usability, and various affordances of learning environments in which assessment is employed. Moreover, the lack of significant impact from individual differences such as reading comprehension, vocabulary knowledge, and prior knowledge on evaluation scores underscores the importance of focusing on a variety of other relevant learner characteristics that might impact learners’ evaluative judgment.

Furthermore, the findings relating to evaluative judgment quality, where neither the type of assessment nor the technological context significantly influenced outcomes, highlight the complexity of evaluative judgment as a skill. This suggests that developing high-quality evaluative judgment may require more comprehensive interventions than simply altering the mode of assessment or introducing technology. Educators might need to integrate explicit instruction on evaluative criteria and judgment strategies within their curricula to enhance this critical skill.

Regarding rating confidence, the unexpected higher confidence associated with peer-assessment, irrespective of the technological context, points to the potential benefits of peer-assessment in enhancing self-efficacy among learners. This may be attributed to the learning effect and
the opportunity for self-reflection and metacognitive awareness that peer-assessment provides. Consequently, educators might consider incorporating structured peer-assessment activities that encourage critical reflection and constructive feedback among learners.

Overall, these findings advocate for investigating different approaches to assessment in educational settings, emphasizing the role of peer interactions, clear evaluative criteria, and opportunities for reflective practice in enhancing evaluative judgments and learner confidence. As educational paradigms continue to evolve, particularly with the increasing integration of technology, these insights contribute to a deeper understanding of the factors that influence assessment outcomes and learner development. Future research should continue to explore these dynamics, particularly the long-term development of evaluative judgment and the optimal integration of technology in assessment practices, to further refine and enhance student’s evaluative judgment.

Suggestions for Further Research

The current research, while providing valuable insights into the effects of assessment types and technology on learner evaluative judgment, is subject to certain limitations that present avenues for future research. The use of crowdsourcing platforms for participant recruitment and data collection, although time and resource efficient, restricts the diversity of the participant to the participant pool within the platform. To address this limitation and validate the findings, future studies could employ lab-based studies, allowing for a more controlled environment to replicate the study and compare the results with the current findings. Another limitation in the design of the study relates to the limited number of evaluation attempts, which involved only two assessment sessions: one for each condition of self-assessment and peer-assessment. As such, the current design does not capture longitudinal effects, leaving unexplored how sustained engagement with different evaluative information might influence evaluation score, evaluative
judgment quality, and rating confidence over time. Longitudinal studies employing several evaluative sessions could provide deeper insights into the evolving impact of feedback on evaluative judgment.

Furthermore, the aggregation of undergraduate participants into a single group in this study, regardless of their educational progression, introduces another potential confound. Learners at different stages of their undergraduate education may respond differently to evaluative information due to their varying levels of experience with evaluative information and feedback. Future research should consider stratifying participants by educational level in undergraduate studies to discern more detailed effects of feedback and assessment types across different stages of academic development. Another limitation relates to the focus of the present study on a single technology-enabled feedback system. Focusing on the SMART system (Kim et al., 2019) per se, limits the generalizability of the findings to other technology-based feedback tools. Future studies could explore comparative effects of a range of similar technology-enabled feedback systems in a single study to compare their impacts on learners’ evaluative judgment and how they stack up against traditional and non-technological feedback systems.

Building upon the previous limitation, it is important to note that the use of expert exemplars as the sole form of non-technology feedback in the study might not fully represent the spectrum of non-technology feedback methods. Future research could incorporate various types of non-technology feedback to provide a broader understanding of its effects on learner outcomes. Similarly, the design of the current study combined multiple feedback sources (student knowledge structure, expert knowledge structure, list of key concepts, progress bar, etc.) from SMART (Kim et al., 2019) in a single intervention and precluded the isolation of the effects of
the individual feedback components. Future research should consider separating these components to examine their distinct impacts on learners’ evaluative judgment.

Furthermore, while the current study focused on self-assessment and peer-assessment, there exists a range of other assessment strategies, such as portfolios and work analysis assessment techniques, that warrant further exploration. Investigating these strategies could reveal their potential for fostering deeper engagement and enhancing learners’ evaluative judgment quality and rating confidence. Also, the lack of significant findings related to individual difference factors such as reading comprehension, vocabulary, and prior knowledge in the current study suggests the need to consider other factors such as writing proficiency, self-efficacy and motivation in future research designs. These factors could provide additional insights into the variables influencing learners’ evaluative judgment.

Lastly, the operationalization of rating confidence through a single-question survey and evaluative judgment through a rubric-based task may not capture the full complexity of these constructs. Future studies could employ different measures, including qualitative approaches like interviews and surveys, to elicit a more comprehensive understanding of learners’ rating confidence and evaluative judgments. Similarly, expanding the conceptualization of prior knowledge (McCarthy & Goldman, 2019) beyond a single-question perception to include various subconstructs could enrich the understanding of its role in evaluative processes.

In summary, this study has illuminated some aspects of how assessment types and feedback mechanisms influence learner evaluative judgment, yet it also underscores the need for further research. The limitations encountered due to the methodological choices and the specificity of feedback mechanisms highlight the complexity of educational assessment and the multifaceted nature of evaluative judgment. Future investigations should strive to diversify study settings,
incorporate a broader range of feedback methods, and delve into the longitudinal impacts of assessment practices. Additionally, acknowledging and examining the role of individual learner differences and expanding the scope to include various assessment strategies and measures can provide new insights.
REFERENCES

ACT. (2006). Reading between the lines: What the ACT reveals about college readiness in reading. ACT, Inc.


Bråten, I., Strømsø, H. I., & Samuelstuen, M. S. (2018). Direct and indirect relationships between individual differences in syntopical reasoning skills, prior topic knowledge, and comprehension when learning from multiple texts. Contemporary Educational Psychology, 53, 154-166.


Heller, R., & Greenleaf, C. L. (2007). *Literacy instruction in the content areas: Getting to the core of middle and high school improvement*. The Alliance for Excellent Education.


Nicol, D. (2013). Resituating feedback from the reactive to the proactive. In D. Boud & E. Molloy (Eds.), *Feedback in higher and professional education: Understanding it and doing it well* (pp. 34-49). Routledge.


APPENDICES

Appendix A: Gates-MacGinitie Reading Comprehension Test

In this section you will read a series of passages. Each of the passages in the test tells what you need to know to answer the questions about it. If you are not sure which answer is right, read the passage again. If you still are not sure, do your best to mark the one you think is right and go on.

Please press the next button to proceed to the practice questions.

Practice Questions

Read the passage below and then answer the following questions. When you have answered the questions, press the next button.

Sometimes - not very often - we get two full moons in one month. That second full moon is called a "blue moon". No one knows why. Now we say, "once in a blue moon" to mean "once in a long time."

To be a "blue moon," the moon must be

○ dark.
○ long.
○ blue.
○ full.

What is it that no one knows?

○ What the name is.
○ Who uses the name.
○ Where the name comes from.
○ What the name means.

Thank you for completing the practice questions. Below are the answers to the two practice questions.
The first question stated: “To be a ‘blue moon,’ the moon must be…”

This was an unfinished sentence. From the options presented, you were asked to decide which answer finished the sentence correctly. The correct answer was “full.”

"To be a ‘blue moon,’ the moon must be full.”

The second question asked: “What is it that no one knows?”

This was a question that asked you to pull information from the passage in order to answer the question.

The correct answer was: “Where the name comes from.”

Next you will complete the timed section of the reading test. You will have 20 minutes to answer as many questions as you can. The survey will automatically progress once the time limit has expired. Press the next button to proceed.

A crowd of people surged into the Eighth Avenue express at 59th Street. By elbowing other passengers in the back, by pushing and heaving, they forced their bodies into the coaches, making room for themselves where no room had existed before. As the train gathered speed for the long run to 125th Street, the passengers settled down into small private worlds, thus creating the illusion of space between them and their fellow passengers. The worlds were built up behind newspapers and magazines, behind closed eyes or while staring at the varicolored show cards that bordered the coaches.

Why was it difficult to get on the train?

- The train didn’t stop long enough.
- There was a barrier in the way.
- The train was already full.
- The people were reading newspapers.
The newspapers helped the passengers

- pass the time.
- forget where they were going.
- sleep.
- feel that they were by themselves.

Staring at the show cards served the same purpose as

- finding a seat.
- getting on the train.
- shutting one’s eyes.
- staring at other passengers.

In this passage, the word **run** means

- trip.
- race.
- string of good luck.
- series of performances.

In later life, John Quincy Adams recalled an incident typical of his mother Abigail’s bravery and resourcefulness. In 1775 British troops from Boston were advancing on Braintree, searching for rebel arsenals. All day neighbors traveled the road in front of the Adams’ farmhouse, retreating from the expected attack. Abigail was alone in her home with her children. When rebel troops arrived, they advised Abigail to flee. Instead, she stayed, handing over all her precious pewter to the rebels, helping them melt down the metal for bullets. The rebel soldiers departed, and Abigail remained, expecting the worst but refusing to give in to the panic that possessed some of the neighbors. "Do you wonder," wrote her son, "that a boy of seven who witnessed this scene is a patriot?"
The neighbors who passed the Adams’ house were trying to

- defend their homes.
- avoid being hurt.
- join one of the armies.
- get to Boston.

The passage suggests that the rebels had little

- ammunition.
- concern for Abigail.
- knowledge of the countryside.
- warning that the British were advancing.

What demonstrated Abigail’s resourcefulness was the way she

- fooled the British troops.
- sent messages to the rebel troops.
- learned where the British troops had come from.
- provided what was needed from what she had available.
John Quincy Adams believed that this experience was a source of his resourcefulness.

- interest in military history.
- courage.
- love of country.

Fresco involves painting into wet lime plaster with pigment mixed into limewater. The layer of calcium carbonate formed by the limewater binds the pigments to the plaster wall, and the mutual wetness of the pigment and the surface causes the color to dye the wall. This makes for a highly permanent decoration, as long-lived as the building itself. Permanence is the main advantage of fresco and is, of course, its own recommendation. Michelangelo’s Creation of Adam, like all other works on the ceiling of the Sistine Chapel in the Vatican, is an example of fresco painting. Since plaster cannot be rewet, once it is dry, the fresco artist never applies more plaster to his surface that he knows he can finish in a single day. Consequently, we can find places in this fresco where plaster joints occur. There is a seam where Adam’s neck fits onto his body and another at the line between the torso and the legs. Adam is about twelve feet long, and it took Michelangelo three sessions to complete him.

In fresco painting, the pigment is first mixed into

- plaster.
- limewater.
- oil.
- the wet part of the wall.
Why are fresco paintings long lasting?

- The seams are strong.
- The pigment becomes part of the wall.
- The plaster is protected by the layer of pigment.
- The painting is protected by the layer of plaster.

About how long does the plaster stay wet enough to paint?

- Ten minutes.
- An hour.
- A day.
- A week.

A fresco artist must be careful to

- rewet the plaster as needed.
- apply the plaster in small enough areas.
- let the plaster dry before beginning to paint.
- let the painting dry before applying plaster.
**A seam in a fresco is a line**

- where the wall joins the ceiling.
- between the different colors.
- between areas painted at different times.
- where material has been added to strengthen the plaster.

**The example of the Creation of Adam shows how one can tell**

- where the artist applied plaster.
- how long ago the fresco was painted.
- how large the figures on a ceiling fresco are.
- how many sessions it took to do a fresco.

An author’s introduction to the story of his life:
I had planned to write chronologically, but then realized that, of course, I don’t think chronologically. Writing a memoir is like fishing. You cast your line and you pull on it when a fish strikes, but you never know what will be on the other end, for the ocean is deep and is filled with marvelous creatures that do not break the surface in expected order. Nor do they swim under the waves with the whales leading and the minnows at the end of long straight line. A memoir, like a fish, will not thrive under every discipline. Another way of putting this is that if you alphabetize
the Iliad, you will have approximately the Athens telephone book. When I think back, things
don’t line up, they stand out, so I will take them as they come, as once I took them as they came.

In this passage, the author explains why he

○ decided to write about himself.

○ waited so long to begin writing.

○ included details that seem unimportant.

○ changed his mind about how he would write.

What do the ocean creatures represent?

○ Events in the author’s life.

○ People the author has known.

○ All the words in the language.

○ The dangers of looking into one’s past.

In this passage, the word discipline means

○ punishment.

○ a field of study.

○ rules by which something is organized.

○ training that perfects mental or moral qualities.
The Athens telephone book is used as an example of something that is

- too long.
- impossible to read.
- orderly but boring.
- full of information.

When the author says "…as once I took them," he means that

- he was always eager to do things.
- he could stand up to any difficulty.
- he believed that he deserved what he got.
- he dealt with experiences as they happened.

All "symmetrical" organisms develop asymmetries. A fruit fly, no longer than the tip of a lead pencil, having developed while stuck to the inside of a glass culture vessel, has different numbers of sensory bristles on his left and right sides, some flies having more on the left, some more on the right. Moreover, this side-to-side variation is as large as the difference among different flies. But the genes on the left and right sides of a fly are the same, and it seems absurd to think that the temperature, humidity, or concentration of oxygen was different between left and right sides of the tiny developing insect. The variation between sides is a result of random events in the timing of division and movement of the individual cells that produce the bristles, so called-developmental noise.
Why does the author put *symmetrical* in quotation marks?

- It is a scientific term.
- It is a new word that the author made up.
- The author is referring to another author’s use of the term.
- The usual meaning of the word is not completely accurate in this context.

**In this passage, the **vessel** is**

- a boat.
- a container.
- a vein or artery.
- a window.

**The passage implies that differences such as that between right- and left-handed fingerprints could be a result of**

- differences in genes.
- differences between individuals.
- symmetry.
- unpredictable variations in the way cells divide.
How does the number of bristles on the right side affect the number of bristles on the left side?

- It has no effect.
- It makes the left side have fewer bristles.
- It makes the left side have an equal number of bristles.
- It makes the left side have more bristles.

Margaret had just gotten her first pair of sunglasses, perfect cat-eyes, and she was amazed at how much she could see. She lay in the scrub grass beneath a stand of cottonwoods, took them off, and watched the branches turn gauzy and familiar. Then she put the glasses back on, bracing a little for the barrage of detail. Thousands of leaves leaped out, trembling and hard edges. The narrow river, a few yards away, turned crunchy looking again. Bird sounds attached themselves to small shapes on high branches.

She didn’t know when her vision had started to go seriously bad. It had been so gradual, this nearsightedness, that she hadn’t noticed it for a while. At first, it seemed only that a luxurious vagueness had come into her life. Then it had begun to make her uneasy. But this sudden return of all the details was more than she really wanted. It was unnerving. It gave her the same feeling she got when someone explained how something scientific works - osmosis, say, or photosynthesis. The explanations crowded out her imagination and made her feel bleak with information.

What was Margaret not sure of?

- Why she had been feeling uneasy.
- When she started to need glasses.
- Whether the glasses were working properly.
- Why everything looked so different through glasses.
What had Margaret liked about not seeing well?

- She needed to imagine things.
- She didn’t have to work.
- She could get people to explain things.
- She wasn’t expected to understand science.

It seems to Margaret that, when she wore glasses, she had

- a feeling of luxury.
- a greater enjoyment of nature.
- too much information.
- a greater awareness of sounds.

The passage suggests that Margaret would have been happier with glasses that were

- weaker.
- smaller.
- like cat-eyes.
- more stylish.
In this passage, the word **bracing** means

- turning.
- pushing away.
- stimulating.
- getting ready.

A pulsar is thought to be a rapidly spinning neutron star. Such stars can arise from the gravitational collapse of a supernova’s core. It is in conserving angular momentum as it shrinks to a diameter of only several kilometers that the neutron star attains its high rotational velocity. If the neutron star continuously emits a beam of electromagnetic radiation from a spot in the magnetized plasma overlying its surface, the beam is swept around like the beacon of a lighthouse. Such a radio beam, striking the earth with each revolution of neutron star, can account for the observed radio-frequency pulsations.

**A supernova’s core becomes a neutron star because of**

- rotation.
- gravity.
- pulsation.
- magnetized plasma.

**A neutron star speeds up because it**

- gets smaller.
- has a radio frequency.
- is magnetized.
- emits a beam.
Pulsars are thought to send out a radio beam from

- their magnetic poles.
- explosions in their interior.
- one place near their surface.
- the place where the beam strikes the earth.

What does **like the beacon of a lighthouse** describe?

- Radiation sent out by a pulsar.
- The star from which a pulsar is formed.
- Signals scientists send out to detect pulsars.
- The path of an object caught in a pulsar’s gravity.

How often the beam from a pulsar strikes the earth depends on

- how far the pulsar is from the earth.
- how large the pulsar is.
- how fast the pulsar is spinning.
- how strong the pulsar’s magnetic field is.

It is customary to place the date for the beginnings of modern medicine somewhere in the mid-1930s, with the entry of the sulfonamides and penicillin into the pharmacopoeia, and it is usual to ascribe to these events the force of a revolution in medical practice. This is what things seemed like at the time. Medicine was upheaved, revolutionized indeed. Therapy had been discovered for great numbers of patients whose illnesses had previously been untreatable. Cures were now available. As we saw it then, it seemed a totally new world. Doctors could now cure disease, and this was astonishing, most of all to the doctors themselves.
During the 1930s, what did people believe had happened in the field of medicine?

- A destructive trend.
- A dramatic change.
- A return to old practices.
- A slowing down.

**Sulfonamides and penicillins made doctors feel**

- confused.
- like scientists.
- old-fashioned.
- more confident.

**In this passage, pharmacopoeia means**

- a medical research laboratory.
- medical school textbooks.
- a school for pharmacists.
- a stock of available medicines.
According to the passage, who was most amazed by sulfonamides and penicillin?

- Sick patients.
- Doctors.
- Patients who had recovered.
- Pharmacists.

Stephen’s mother and his brother and one of his cousins waited at the corner of quiet Foster Place while he and his father went up the steps and along the colonnade where the Highland sentry was parading. When they had passed into the great hall and stood at the counter Stephen drew forth his orders on the governor of the bank of Ireland for thirty and three pounds; and these sums, the moneys of his exhibition and essay prize, were paid over to him rapidly by the teller in notes and in coin respectively. He bestowed them in his pockets with feigned composure and suffered the friendly teller, to whom his father chatted, to take his hand across the broad counter and wish him a brilliant career in the afterlife.

The passage suggests that the building was

- hidden.
- crowded.
- impressive
- hard to get into.

What had Stephen done?

- He had won a prize.
- He had carried out orders.
- He had sold a painting.
- He had had a brilliant career.
Why did the teller give the notes to Stephen rapidly?

- To get rid of Stephen.
- To show that he was not impressed.
- Because he was being efficient.
- Because Stephen’s mother was waiting.

It was difficult for Stephen to

- act calmly.
- pass into the hall.
- give up the orders.
- leave his mother waiting.

The teller took Stephen’s hand to

- greet him.
- congratulate him.
- give him confidence.
- show him how to handle money.

The Museum that Alexander the Great set up in Alexandria was in effect the first university in the world. As its name implies, it was dedicated to the service of the Muses. It was, however, a religious body only in form, in order to meet the legal difficulties of the endowment in a world that had never foreseen such a thing as a secular intellectual process. It was essentially a college of learned men engaged chiefly in research and record, but also to a certain extent in teaching.
Why was the Museum set up as a religious body?

- So money could be given to it.
- So people could come worship there.
- So priests could work there.
- So religion could be taught.

The Museum was most like a

- temple.
- university.
- hospital.
- show.

Which answer best describes the Museum?

- Famed for its athletics.
- Ineffective.
- Pioneering.
- Entertaining.

Any list of mutualistic relationships would be heavily weighted toward the highly organized, impersonal world of the insects. The story of ants protecting and "milking" their cattle-like aphids, for example, is well known. Much less common is evidence of mutualism among warm blooded vertebrates, and mutualistic relationships that cross taxonomic class lines, say between birds and mammals, are especially rare.
The passage mentions the relation between ants and aphids as an example of

- crossing taxonomic class lines.
- insects.
- an impersonal world.
- mutualism.

In this passage, *class* means a

- style.
- school group.
- social group.
- category.

The passage characterizes insect societies as

- ordered.
- highly motivated.
- small in scale.
- weighted.
Appendix B: Reading Passage

Brain and Computer Interaction

Brains and computers share lots of similarities. Both use electrical signals transmitted through complex networks to trigger a particular event. The human brain remains far more sophisticated than even the most advanced computer, but the gap is narrowing, and, with the help of other innovative technologies, the potential of linking up brains with computers becomes closer to a practical reality. The combination of human and machine intelligence has a staggering array of potential real-world applications, from the detection of cognitive illnesses or emotional states to precise control of sophisticated machines, to the intuitive navigation of immersive virtual worlds.

Recent advances in AI, material sensor technologies and attention to user-friendliness in computer hardware have spurred BCI research, taking it from pure academia into the world of industry. BCIs provide a channel for humans to interact with external artificial devices through their brain activity. Usually, a machine learning algorithm decodes electrical signals in the brain to work out the user’s intentions and then transmits a ‘mental command’ to the device. There are a variety of different sensor systems to pick up and interpret brain activity, each with their own pros and cons. Some measure the brain’s electrical signals, some measure its magnetic activity and others measure changes in blood oxygen.
While some measurement techniques are non-invasive, others are invasive, meaning the sensors must be placed under the scalp. While invasive methods usually provide more precise measurements, they also have obvious disadvantages for the user. Understandably, lots of people do not want electrical devices under their skin, and there are risks involved with prolonged or repeated use, which makes invasive methods unlikely to be widely taken up in society. Research on invasive methods is mostly limited to animals now and, although latest findings indicate they might be safe for human use, it’s unlikely that human implants will be approved any time soon. For the foreseeable future, non-invasive sensors are the only practical solution for investigating cognitive processes in the human brain.

While BCIs have crossed the threshold from fiction to fact, uptake in the world outside the lab is slow. Interacting with the real world via a computer is unnatural to humans. Furthermore, the response feedback produced by the computer is far slower than our brains, creating a delay. The accuracy of machine learning provides another limitation, though this is a rapidly advancing field. Recent progress in deep learning provides substantial potential benefits for neural networks, computer vision and BCI techniques.

There are other practical drawbacks too. The sensors that pick up signals from the brain still need improvement. Most BCIs only work when the user is stationary, limiting their use in many real-world applications. Because non-invasive methods rely upon detecting quite subtle stimuli, any interference – such as the noise created by movement or physically interacting with an object – can affect their functioning. Staying motionless while ‘interacting’ with the world is unnatural to us and means such methods can only be used in controlled environments.

We are already familiar with the concept of wearable computers, through VR headsets, AR glasses and other trendy tech. As well as their entertainment value, wearable computers also have many promising practical applications. Wearable computers can offer highly immersive experiences for entertainment, health monitoring and research purposes, among many others. Their research applications are most exciting for us at present. Wearable computers have revolutionized the practicalities of BCI research. Until recently, BCI research has relied upon static and simple stimuli – presenting an object to a subject in a lab environment, for instance – which does not bear much resemblance to everyday life. By using wearable computers, researchers can design, simulate, and finely control experiments to examine human brain dynamics inside and outside the laboratory. VR and AR can now create sophisticated scenarios like real life; by monitoring a subject’s brain activity when encountering these scenarios, results are far more meaningful in terms of relating findings to the real world.

Based on such technological advancements and their own innovations, researchers are working on a next-generation solution called direct-sense BCI (DS-BCI). They are developing two systems within DS-BCI, based on speech and vision, to seamlessly decode the brain signals linked to our natural senses without additional stimulus. Such techniques are far closer aligned to our experiences of the real world than most BCI research. To make its measurements, the team is principally using non-invasive sensors that monitor electrical signals in the brain, the signal known as electroencephalography (EEG). Direct-speech BCI aims to translate ‘silent speech’ from neural signals into system commands. Currently, the EEG sensors can decode the intention conveyed when participants imagine themselves speaking. This approach provides a novel channel of interaction with BCIs for any user and could be an important assistive tool for people not able to speak naturally.
Direct-sight BCI detects what object is in a person’s mind based on their EEG signals as they look around an environment. This is more innovative than current BCI methods, which rely mostly on stimulus onset – in other words, displaying a specific series of objects to the subject. The ability to actively recognize objects is an essential skill in daily life, which is why this is an active research field. Overall, the possibilities for the future of BCIs are exciting to researchers. Building a system that translates user intentions into BCI instructions has shifted from a distant goal to a feasible possibility.

Source: https://futurumcareers.com/plugging-in-directly-linking-the-brain-to-a-computer

Copyright: All Futurum content is produced, created, and published under the Creative Commons license: Attribution-NonCommercial 4.0 International (CC BY-NC 4.0): https://creativecommons.org/licenses/by-nc/4.0/. This means that you are free to share (copy and redistribute the material in any medium or format) and adapt (remix, transform and build upon the material).

Appendix C: Rubric

<table>
<thead>
<tr>
<th>Points</th>
<th>CONTENT: quality</th>
<th>CONTENT: coverage</th>
<th>CONTENT: coherence</th>
<th>ARGUMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Most of the ideas in the summary and argument are either not central to the topic, not clearly expressed, vague or repetitive.</td>
<td>Most of the central ideas from the article(s) are not expressed clearly in the summary and argument, and ideas from the article(s) that are included are expressed in a way that is</td>
<td>The ideas expressed in the summary and argument are not easy to follow, and do not relate well to one another.</td>
<td>Essay responds to the topic in some way but does not state a claim on the issue.</td>
</tr>
<tr>
<td>Points</td>
<td>CONTENT: quality</td>
<td>CONTENT: coverage</td>
<td>CONTENT: coherence</td>
<td>ARGUMENT</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>1</td>
<td>Many of the ideas in the summary and argument relate only to the topic, but some of them may be central to the topic, which may be due to vagueness, repetition, lack of clarity, or failure to express central ideas.</td>
<td>Some of the central ideas from the article(s) are clearly expressed in the summary and argument, but many of the central ideas from the article(s) are missing, or are expressed in a way that is unclear, vague, or repetitive.</td>
<td>Some of the ideas expressed in the summary and argument relate well to one another, but most of the ideas do not relate well to one another and are not easy to follow.</td>
<td>Essay states a claim, but no evidence or reasons are given to support the claim, or the reasons given are unrelated to the claim, or inconsistent with one another, or are incoherent.</td>
</tr>
<tr>
<td>2</td>
<td>About half the ideas in the summary and argument are expressed clearly and relate to the topic, but about half the ideas do not meet the criteria of being clear and centrally, which may be due to vagueness, repetition, lack of clarity, or failure to identify central ideas.</td>
<td>Many of the central ideas from the article(s) are expressed clearly in the summary and argument, but many of the central ideas from the article(s) are missing, or are expressed in a way that is unclear, vague or repetitive.</td>
<td>Many of the ideas expressed in the summary and argument relate well to one another, making it fairly easy to follow much of the discussion. But many of the ideas expressed in the summary and argument do not relate well to one another, so it is difficult to form a coherent understanding of the discussion.</td>
<td>Essay states a clear claim and gives one or two reasons to support the claim, but the reasons are not explained or supported in any coherent way. The reasons may be of limited plausibility, and inconsistencies may be present.</td>
</tr>
<tr>
<td>3</td>
<td>About half the ideas in the summary and argument are expressed clearly and are central to the topic, and the ideas are related to the topic, though there is</td>
<td>Most of the central ideas from the article(s) are expressed clearly in the summary and argument. The remaining ideas from the article that are expressed in the</td>
<td>Most of the ideas expressed in the summary and argument relate well to one another, and the discussion as a whole is fairly easy to follow. A few ideas seem out of place or less well</td>
<td>Essay states a claim and gives some reasonable reasons to support the claim. The elaborated reasoning is generally plausible although not thoroughly convincing.</td>
</tr>
<tr>
<td>Points</td>
<td>CONTENT: quality</td>
<td>CONTENT: coverage</td>
<td>CONTENT: coherence</td>
<td>ARGUMENT</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------</td>
<td>------------------</td>
<td>-------------------</td>
<td>----------</td>
</tr>
<tr>
<td>4</td>
<td>little to no vagueness or repetition. However, about half the ideas are either not central to the topic or unclear.</td>
<td>summary and argument are either not central, not clear, vague or repetitive.</td>
<td>integrated with the overall organization.</td>
<td>There may be minor inconsistencies, irrelevant information, or problems with organization and clarity.</td>
</tr>
<tr>
<td>5</td>
<td>Most of the ideas in the summary and argument are expressed clearly and are central to the topic, and there is little to no vagueness or repetition. However, some ideas are either unclear, or not central to the topic, or some combination.</td>
<td>Most of the central ideas from the article(s) are expressed clearly in the summary and argument. Nearly all ideas from the article(s) expressed in the summary and argument are related to the topic, and are expressed clearly, with little vagueness or repetition.</td>
<td>Most of the ideas expressed in the summary and argument relate well to one another, and the discussion as a whole is fairly easy to follow. A few ideas seem out of place or less well integrated with the overall organization.</td>
<td>Essay states a clear claim and gives reasons to support the claim. The reasons are elaborated using evidence clearly. The reasoning is generally plausible although not perfectly convincing. There may be little inconsistencies that weaken the argument.</td>
</tr>
<tr>
<td>5</td>
<td>All or nearly all the ideas in the summary and argument are related to the topic, most of them are central to the topic, and all or nearly all are expressed clearly, with little or no vagueness or repetition.</td>
<td>All or nearly all of the central ideas from the article(s) are clearly expressed in the summary and argument. Very few of the ideas from the article(s) that are expressed in the summary and argument are not central to the topic, and very few are unclear, vague or repetitive.</td>
<td>All or nearly all of the ideas expressed in the summary and argument relate well to one another, making it easy to follow the discussion as a whole. The flow from one idea to the next and from one part of the argument to the next is very coherent, and the overall organization is very coherent.</td>
<td>Meets the criteria for the previous level. In addition, the essay is clear and well organized with undeniable and convincing statements. The writing is free of inconsistencies and irrelevant items that would weaken the argument.</td>
</tr>
</tbody>
</table>
Appendix D: Vocabulary Knowledge Test

1. see: They <saw it>.
   
a closed it tightly
b waited for it
c looked at it
d started it up

2. time: They have a lot of <time>.
   a money
   b food
   c hours
   d friends

3. period: It was a difficult <period>.
   a question
   b time
   c thing to do
   d book

4. figure: Is this the right <figure>?
   a answer
   b place
   c time
   d number

5. poor: We <are poor>.
   a have no money
   b feel happy
   c are very interested
   d do not like to work hard

6. microphone: Please use the <microphone>.
   a machine for making food hot
b machine that makes sounds louder

c machine that makes things look bigger

d small telephone that can be carried around

7. nil: His mark for that question was <nil>.

   a very bad

   b nothing

   c very good

   d in the middle

8. pub: They went to the <pub>.

   a place where people drink and talk

   b place that looks after money

   c large building with many shops

   d building for swimming

9. circle: Make a <circle>.

   a rough picture

   b space with nothing in it

   c round shape

   d large hole

10. dig: Our dog often <digs>.

    a solves problems with things

    b creates a hole in the ground

    c wants to sleep

    d enters the water
11. soldier: He is a <soldier>.
   a person in a business
   b person who studies
   c person who uses metal
   d person in the army

12. restore: It has been <restored>.
   a said again
   b given to a different person
   c made like new again
   d given a lower price

13. pro: He’s <a pro>.
   a someone who is employed to find out important secrets
   b a stupid person
   c someone who writes for a newspaper
   d someone who is paid for playing sport

14. compound: They made a new <compound>.
   a agreement
   b thing made of two or more parts
   c group of people forming a business
   d guess based on past experience

15. deficit: The company <had a large deficit>.
   a spent a lot more money than it earned
   b went down a lot in value
c had a plan for its spending that used a lot of money
d had a lot of money stored in the bank

16. strap: He broke the <strap>.
   a promise
   b top cover
   c shallow dish for food
   d strip of strong material

17. weep: He <wept>.
   a finished his course
   b cried
   c died
   d worried

18. haunt: The house is <haunted>.
   a full of decorations
   b rented
   c empty
   d full of ghosts

19. cube: I need one more <cube>.
   a sharp thing used for joining things
   b solid square block
   c tall cup with no saucer
   d piece of stiff paper folded in half

20. butler: They have a <butler>. 
21. nun: We saw a <nun>.  
   a long thin creature that lives in the earth  
   b terrible accident  
   c woman following a strict religious life  
   d unexplained bright light in the sky

22. olive: We bought <olives>.  
   a oily fruit  
   b scented flowers  
   c men’s swimming clothes  
   d tools for digging

23. shudder: The boy <shuddered>.  
   a spoke with a low voice  
   b almost fell  
   c shook  
   d called out loudly

24. threshold: They raised the <threshold>.  
   a flag  
   b point or line where something changes  
   c roof inside a building
25. demography: This book is about <demography>.
   a the study of patterns of land use
   b the study of the use of pictures to show facts about numbers
   c the study of the movement of water
   d the study of population

26. malign: His <malign> influence is still felt.
   a good
   b evil
   c very important
   d secret

27. strangle: He <strangled her>.
   a killed her by pressing her throat
   b gave her all the things she wanted
   c took her away by force
   d admired her greatly

28. dinosaur: The children were pretending to be <dinosaurs>.
   a robbers who work at sea
   b very small creatures with human form but with wings
   c large creatures with wings that breathe fire
   d animals that lived an extremely long time ago

29. jug: He was holding <a jug>.
   a a container for pouring liquids
b an informal discussion

c a soft cap

d a weapon that blows up

30. crab: Do you like <crabs>?

a very thin small cakes

b tight, hard collars

c sea creatures that always walk to one side

d large black insects that sing at night

31. quilt: They made a <quilt>.

a statement about who should get their property when they die

b firm agreement

c thick warm cover for a bed

d feather pen

32. tummy: Look at my <tummy>.

a fabric to cover the head

b stomach

c small soft animal

d finger used for gripping

33. eclipse: <There was an eclipse>.

a A strong wind blew all day

b I heard something hit the water

c A large number of people were killed

d The sun was hidden by the moon
34. excrete: This was <excreted> recently.
   a pushed or sent out
   b made clear
   c discovered by a science experiment
   d put on a list of illegal things

35. ubiquitous: Many unwanted plants <are ubiquitous>.
   a are difficult to get rid of
   b have long, strong roots
   c are found everywhere
   d die away in the winter

36. marrow: This is <the marrow>.
   a symbol that brings good luck to a team
   b soft centre of a bone
   c control for guiding a plane
   d increase in salary

37. cabaret: We saw the <cabaret>.
   a painting covering a whole wall
   b song and dance performance
   c small crawling creature
   d person who is half fish, half woman

38. cavalier: He treated her <in a cavalier manner>.
   a without care
   b with good manners
c awkwardly

d as a brother would

39. veer: The car <veered>.
   a moved shakily
   b changed course
   c made a very loud noise
   d slid without the wheels turning

40. yoghurt: This <yoghurt> is disgusting.
   a dark grey mud found at the bottom of rivers
   b unhealthy, open sore
   c thick, soured milk, often with sugar and flavouring
   d large purple fruit with soft flesh

41. octopus: They saw <an octopus>.
   a a large bird that hunts at night
   b a ship that can go under water
   c a machine that flies by means of turning blades
   d a sea creature with eight legs

42. monologue: Now he has a <monologue>.
   a single piece of glass to hold over his eye to help him to see
   b long turn at talking without being interrupted
   c position with all the power
   d picture made by joining letters together in interesting ways

43. candid: Please <be candid>.
a be careful
b show sympathy
c show fairness to both sides
d say what you really think

44. nozzle: Aim the <nozzle> toward it.
   a space that light passes through in a camera
   b dry patch of skin
   c pipe attachment that forces water
   d sharp part of a fork

45. psychosis: He has <a psychosis>.
   a an inability to move
   b an oddly coloured patch of skin
   c a body organ that processes sugar
   d a mental illness

46. ruck: He got hurt in the <ruck>.
   a region between the stomach and the top of the leg
   b noisy street fight
   c group of players gathered round the ball in some ball games
   d race across a field of snow

47. rouble: He had a lot of <roubles>.
   a very valuable red stones
   b distant members of his family
   c Russian money
d moral or other difficulties in the mind

48. canonical: These are <canonical examples>.
   a examples which break the usual rules
   b examples taken from a religious book
   c regular and widely accepted examples
   d examples discovered very recently

49. puree: This <puree> is bright green.
   a fruit or vegetables in liquid form
   b dress worn by women in India
   c skin of a fruit
   d very thin material for evening dresses

50. vial: Put it in a <vial>.
   a device which stores electricity
   b country residence
   c dramatic scene
   d small glass bottle

51. counterclaim: They made <a counterclaim>.
   a a demand response made by one side in a law case
   b a request for a shop to take back things with faults
   c an agreement between two companies to exchange work
   d a decorative cover for a bed, which is always on top

52. refectory: We met in the <refectory>.
   a room for eating
b office where legal papers can be signed  
c room for several people to sleep in  
d room with glass walls for growing plants  

53. trill: He practised the <trill>.  
   a repeated high musical sound  
   b type of stringed instrument  
   c way of throwing the ball  
   d dance step of turning round very fast on the toes  

54. talon: Just look at those <talons>!  
   a high points of mountains  
   b sharp hooks on the feet of a hunting bird  
   c heavy metal coats to protect against weapons  
   d people who make fools of themselves without realizing it  

55. plankton: We saw a lot of <plankton> here.  
   a poisonous plants that spread very quickly  
   b very small plants or animals found in water  
   c trees producing hard wood  
   d grey soil that often causes land to slip  

56. soliloquy: That was an excellent <soliloquy>!  
   a song for six people  
   b short clever saying with a deep meaning  
   c entertainment using lights and music  
   d speech in the theatre by a character who is alone
57. puma: They saw a <puma>.
   a small house made of mud bricks
   b tree from hot, dry countries
   c large wild cat
   d very strong wind that lifts anything in its path

58. augur: It <augured well>.
   a promised good things for the future
   b agreed with what was expected
   c had a colour that looked good with something else
   d rang with a clear, beautiful sound

59. emir: We saw the <emir>.
   a bird with two long curved tail feathers
   b woman who cares for other people’s children in eastern countries
   c Middle Eastern chief with power in his own land
   d house made from blocks of ice

60. didactic: The story <is very didactic>.
   a tries hard to teach something
   b is very difficult to believe
   c deals with exciting actions
   d is written with unclear meaning

61. cranny: Look what we found in the <cranny>!
   a sale of unwanted objects
   b narrow opening
c space for storing things under the roof of a house

d large wooden box

62. lectern: He stood at the <lectern>.

a desk made to hold a book at a good height for reading
b table or block used for church ceremonies
c place where you buy drinks
d very edge

63. azalea: This <azalea> is very pretty.

a small tree with many flowers growing in groups
b light natural fabric
c long piece of material worn in India
d sea shell shaped like a fan

64. marsupial: It is <a marsupial>.

a an animal with hard feet
b a plant that takes several years to grow
c a plant with flowers that turn to face the sun
d an animal with a pocket for babies

65. bawdy: It was very <bawdy>.

a unpredictable
b innocent
c rushed
d indecent

66. crowbar: He used a <crowbar>. 
a heavy iron pole with a curved end
b false name
c sharp tool for making holes in leather
d light metal walking stick

67. spangled: Her dress was <spangled>.
   a torn into thin strips
   b covered with small bright decorations
   c made with lots of folds of fabric
   d ruined by touching something very hot

68. aver: She <averred> that it was the truth.
   a refused to agree
   b declared
   c believed
   d warned

69. retro: It had <a retro look>.
   a a very fashionable look
   b the look of a piece of modern art
   c the look of something which has been used a lot before
   d the look of something from an earlier time

70. rascal: She is such <a rascal> sometimes.
   a an unbeliever
   b a dedicated student
   c a hard worker
d a bad girl

71. tweezers: They used <tweezers>.
   a small pieces of metal for holding papers together
   b small pieces of string for closing wounds
   c a tool with two blades for picking up or holding small objects
   d strong tool for cutting plants

72. bidet: They have a <bidet>.
   a low basin for washing the body after using the toilet
   b large fierce brown dog
   c small private swimming pool
   d man to help in the house

73. sloop: Whose <sloop> is that?
   a warm hat
   b light sailing boat
   c left over food
   d untidy work

74. swingeing: They got <swingeing fines>.
   a very large fines
   b very small fines
   c fines paid in small amounts at a time
   d fines that vary depending on income

75. cenotaph: We met at the <cenotaph>.
   a large and important church
b public square in the centre of a town

c memorial for people buried somewhere else

d underground train station

76. denouement: I was disappointed with the <denouement>

a ending of a story which solves the mystery

b amount of money paid for a piece of work

c small place to live which is part of a bigger building

d official report of the results of a political meeting

77. bittern: She saw a <bittern>.

a large bottle for storing liquid

b small green grass snake

c false picture caused by hot air

d water bird with long legs and a very loud call

78. reconnoitre: They have gone to <reconnoitre>.

a think again

b make an examination of a new place

c have a good time to mark a happy event

d complain formally

79. magnanimity: We will never forget her <magnanimity>.

a very offensive and unfriendly manners

b courage in times of trouble

c generosity

d completely sincere words
80. effete: He has become <effete>.
   a weak and soft
   b too fond of strong drink
   c unable to leave his bed
   d extremely easy to annoy

81. rollick: They were <rollicking>.
   a driving very fast
   b staying away from school without being permitted to
   c having fun in a noisy and spirited way
   d sliding on snow using round boards

82. gobbet: The cat left a <gobbet> behind.
   a strip of torn material
   b footprint
   c piece of solid waste from the body
   d lump of food returned from the stomach

83. rigmarole: I hate the <rigmarole>.
   a very fast and difficult dance for eight people
   b funny character in the theatre
   c form which must be completed each year for tax purposes
   d long, pointless and complicated set of actions

84. alimony: The article was about <alimony>.
   a feelings of bitterness and annoyance, expressed sharply
   b money for the care of children, paid regularly after a divorce
c giving praise for excellent ideas

d a metal which breaks easily and is bluish white

85. roughshod: He <rode roughshod>.

a travelled without good preparation

b made lots of mistakes

c did not consider other people’s feelings

d did not care about his own comfort

86. copra: They supply <copra>.

a a highly poisonous substance used to kill unwanted plants

b the dried meat from a large nut used to make oil

c an illegal substance which makes people feel good for a short time

d strong rope used on sailing ships

87. bier: She lay on the <bier>.

a folding garden chair

b grass next to a river

c place where boats can be tied up

d board on which a dead body is carried

88. torpid: He was <in a torpid state>.

a undecided

b filled with very strong feelings

c confused and anxious

d slow and sleepy

89. dachshund: She loves her <dachshund>. 
a warm fur hat

b thick floor rug with special patterns

c small dog with short legs and a long back

d old musical instrument with twelve strings

90. *cadenza*: What did you think of the *cadenza*?

a cake topped with cream and fruit

b large box hanging from a wire that carries people up a mountain

c slow formal dance from Italy

d passage in a piece of music that shows the player’s great skill

91. *obtrude*: These thoughts *obtruded* themselves.

a got themselves lost or forgotten

b did not agree with each other

c got mixed up with each other

d pushed themselves forward in the mind

92. *panzer*: They saw the *panzers* getting nearer.

a players in a marching band

b fighter planes

c large, slow windowless army cars

d policewomen

93. *cyborg*: She read about *a cyborg*.

a an integrated human-machine system

b a musical instrument with forty strings

c a small, newly invented object
d a warm wind in winter

94. zygote: It is <a zygote>.

a an early phase of sexual reproduction
b a lot of bother over nothing
c a small animal found in southern Africa
d a gun used to launch rockets

95. sylvan: The painting had a <sylvan> theme.

a lost love
b wandering
c forest
d casual folk

96. sagacious: She had many ideas that were <sagacious>.

a instinctively clever
b ridiculous and wild
c about abusing people and being abused
d rebellious and dividing

97. spatiotemporal: My theory is <spatiotemporal>.

a focussed on small details
b annoying to people
c objectionably modern
d oriented to time and space

98. casuist: Don’t <play the casuist> with me!

a focus only on self-pleasure
b act like a tough guy

c make judgments about my conduct of duty

d be stupid

99. cyberpunk: I like <cyberpunk>.

a medicine that does not use drugs

b one variety of science fiction

c the art and science of eating

d a society ruled by technical experts

100. pussyfoot: Let’s not <pussyfoot around>.

a criticize unreasonably.

b take care to avoid confrontation

c attack indirectly

d suddenly starts.