DALARNA UNIVERSITY

Degree Thesis

Master's level

Assessment equivalence

Assessing EFL student writing in a Swedish context

Likvärdig bedömning

Bedömning av skrivförmågan hos elever med engelska som främmande språk ur ett svenskt perspektiv

Author: Fredrik Mattsson
School: Högskolan Dalarna
Supervisor: Jonathan White
Examiner: Miguel Garcia-Yeste
Subject/main field of study: English
Course code: AEN25J
Credits: 30 credits
Date of examination: 2023-06-03

Abstract:

The purpose of this study is to examine the validity and reliability of summative assessment of EFL student writing in a Swedish context. Three teachers have assessed the same four student essays from the *English 6* course in Swedish upper secondary school. In addition to grading each essay, the teachers have indicated the extent of conformity to the grading criteria in terms of *flow, structure, cohesion, adaptation to purpose, clarity,* and *variation*. The analyzed data show a variation in assessment criteria interpretation, affecting assessment validity and reliability, and questioning the assessment equivalence of the Swedish criterion-referenced grading system.

Keywords:

Assessment, summative assessment, validity, reliability, criterion-referenced grading system, equivalence, assessment criteria.

Abstract:

Syftet med denna studie är att undersöka validiteten och reliabiliteten hos summativa bedömningar av studentuppsatser ur ett svenskt perspektiv. Tre lärare har bedömt samma fyra studentuppsatser från *engelska 6*. Förutom att betygsätta varje uppsats har lärarna angett graden av överensstämmelse med betygskriterierna: flöde, struktur, sammanhållning, anpassning till syfte, tydlighet och variation. De analyserade data visar en variation i tolkning av betygskriterier, vilket påverkar bedömningens validitet och reliabilitet och ifrågasätter bedömningslikvärdigheten i det svenska mål-relaterade betygssystemet.

Nyckelord:

Bedömning, summativ bedömning, validitet, reliabilitet, mål-relaterat betygsystem, likvärdighet, bedömningskriterier.

Table of contents

1. INTRODUCTION 1

2. BACKGROUND 2

2.1. TERMINOLOGY 2
2.2. BRIEF HISTORICAL OVERVIEW 3
2.2.1. The criterion-referenced system 3
2.3. ASSESSMENT IS SWEDISH SECONDARY SCHOOL 3
2.3.1. Assessment equivalence 3
2.3.2. The National Tests 4
2.4. ASSESSING UPPER SECONDARY ENGLISH 5
2.4.1. Knowledge requirements 5

3. PREVIOUS RESEARCH 5

3.1. ASSESSMENT VALIDITY AND RELIABILITY 6
3.2. CRITERION-REFERENCED ASSESSMENTS 6
3.3. FORMATIVE AND/OR SUMMATIVE ASSESSMENT 7
3.4. ASSESSMENT EQUIVALENCE 8
3.4.1. The National tests 8
3.4.2. Peer-assessment 9
3.5. ACCOUNTABILITY 9

4. THEORETICAL PERSPECTIVE 10

5. METHOD 12

5.1. MATERIAL AND METHOD 12
5.1.1. The students' essays 13
5.1.2. The questionnaire 13
5.1.3. Selection of informants 14
5.1.4. The pilot study 14
5.2. METHOD OF ANALYSIS 14
5.3. VALIDITY AND RELIABILITY 15
5.4. ETHICAL CONSIDERATIONS 16

6. RESULTS 16

6.1. THE QUANTITATIVE DATA 16
6.1. THE GRADES 17
6.2. GRADES AND CRITERIA 17
6.2.1. Grade parameters 18
6.2.2. Interrater Consistency 19
6.2.3. Intrarater reliability 19
6.2.4. Grade and confidence 20
6.3. THE QUALITATIVE DATA 20
6.3.1. Language and structure 21
6.2.2 Content 21
6.4. DATA CONFORMITY 22
6.5. DATA VALIDITY AND RELIABILITY 22

7. DISCUSSION 22

7.1. ASSESSMENT VALIDITY AND RELIABILITY 23
7.1.1. Assessment Validity 23
7.1.2. Assessment reliability 23
7.2. ASSESSMENT GUIDELINES 23
7.3. ASSESSMENT EQUIVALENCE 24

7.5 LIMITATIONS OF THE STUDY 25
7.6 FUTURE RESEARCH 25
8. CONCLUSION 26
REFERENCES 28
APPENDIX 1. LETTER OF INFORMED CONSENT 33
APPENDIX 2. THE QUESTIONNAIRE 35
APPENDIX 3. THE STUDENTS' ESSAYS 38

1. Introduction

Assessment is an integral part of education. Be it summative or formative, assessments provide crucial information to both students and teachers. Assessments are instrumental in tracking student progression in learning as well as providing a motivation for learning, and high grades are essential for the transition into higher education and working life. Therefore, it is imperative that assessments provide information which is *valid, reliable* and *comparable* (Black & Wiliam, 2018; Skolverket, 2020, 2022a).

A criterion-referenced grading system is more vulnerable to *grade inflation* than a norm-referenced system (Lok et al., 2016), which is also the case for the Swedish system (Utbildningsdepartementet, 2023). In a system where grades are high stakes instruments, grade inflation leads to negative consequences, not only for the individual who receives the grades, but also for the educational system and society at large (Wikström, 2005). It is known that there are both internal and external pressures for high grading, due to the grades' function as a quality indicator for schools as well as a selection instrument for students on the academic track (Wikström & Wikström, 2005). A lack of defined standards or problems with interpreting standards might be a factor that contributes to such a process and the potential for making subjective interpretations of grading criteria has been criticized for being a weakness in the Swedish system (Andersson, 2002; Tholin, 2003, cited in Wikström, 2005).

Swedish schools are mandated by law to provide all students with an equal education, but the terminology *likvärdig undervisning* is not clearly defined (SFS 2010:800). As a result, studies report large discrepancies in equivalence in the Swedish criterion-referenced grading system (Gustafsson et al., 2014; Erickson, 2018; Utbildningsdepartementet, 2023).

The aim of this study is to investigate aspects of equivalence in assessments of student essays, and potential causes for a lack of validity and reliability in the assessment of EFL student writing in a Swedish context. In doing so, the criteria specified in both curriculum and syllabus, as well as the aims for the National tests will be analyzed and problematized. The basis for such an analysis is the following assumption: *with clear criteria and knowledge requirements comes assessment equivalence, regardless of the assessor*.

The following research questions will be addressed:

- How valid and reliable are the assessments of student writing in the Swedish upper secondary EFL context?
- Does the Swedish criterion-referenced grading system provide clear guidelines for assessment?
- Does the Swedish criterion-referenced grade system ensure assessment equivalence?

Much of the related previous research on assessment equivalence in Sweden focus on the National tests, which differ from 'regular' essay writing in that the tests are constructed externally. The National tests are also accompanied by extensive instructions for assessment and use a grading scale not normally used in the Swedish criterion-referenced system. The aim of this study is to investigate issues of assessment equivalence in "regular" essay writing and compare the findings with the previous research. Additionally, by relating the assessments to the explicit assessment criteria for the task, this study aims to highlight potential causes for lack of equivalence related to the criterion-referenced system. In doing so, additional factors affecting the assessments can be made clear. Furthermore, this study hopes to highlight

differences in how teachers balance the demands on research basis and proven experience when grading.

2. Background

In this section the specific terminology used will be explained and defined. After that follows a brief historical overview of the Swedish educational system and the criterion-referenced system. After which follows a section on assessment in the Swedish upper secondary school, which includes information regarding assessment equivalence and the role and importance of the National tests. This section ends with a section on the assessment of upper secondary English, which includes details about the knowledge requirements.

2.1. Terminology

There are two main fields in assessment: formative assessments (*FA*) and summative assessments (*SA*). Where formative assessments aim at helping students identify their strengths and weaknesses, summative assessments evaluate the achieved student learning in relation to the criteria or goals. Grading is the process of evaluating the knowledge shown by a student in relation to the criteria (Skolverket, 2022a). Legally secure grading requires grades which rest on a combination of research basis and proven experience (Skolverket, 2022a, p. 8). Proven experience must be documented, communicated, and reviewed and tested based on ethical principles in a collegial context (RFR, 2012/2013). Grading is a form of summative assessment, and for the purpose of this thesis the terms *assessment* and *grading* both refer to summative assessment.

The term *high-stakes* is used for systems and tests which have significant consequences for individuals, and this undoubtedly includes the grades and the National tests in the Swedish educational system (Gustafsson et al., 2014).

One definition of validity is *adequately capturing the essence of that which is being assessed, the construct* (Lundgren et al., 2020). Consequently, validity refers to the degree to which a method assesses what it claims or intends to assess, and the different types of validity include content validity, criterion validity, and construct validity. Performance based assessments are typically viewed as providing more valid data than traditional examinations (Erickson, 2018).

Reliability is often discussed in terms of consistency. *Interrater reliability* refers to the consistency between different assessors, whereas *intrarater reliability* refers to assessment inconsistencies with one single assessor (Gustafsson & Erickson, 2013). Reliability thus refers to the replicability of results, and reliable assessments should not vary between students and assessors/teachers.

Assessment equivalence can be defined as a given grade corresponding to the same knowledge or ability regardless of where or by whom it is given (Lundahl, 2014). Assessment equivalence thus requires assessments which are of high validity and high reliability.

The analysis of the data consists to a large extent of describing variations of data. Throughout the analysis the symbol, (Δ), is used to describe *variation*. In the analysis of the data, the term grade step is used to variation (Δ) of grades. The definition of one grade step corresponds to the distance between for instance E – D, or C – B, (Δ1). Consequently, a variation (Δ) of two grade steps equals the distance between F – D, and E – C, (Δ2).

2.2. Brief historical overview

Until 1994 Sweden had a norm-referenced grading system, containing detailed and specified goals. With the introduction of the new curriculum, Lpo94 in 1994, a criterion-referenced grading system was introduced, in which the previous criteria had been replaced by goals to be achieved, but without a specification of content and method (Jönsson & Thornberg, 2014). Lacking in clarity, the new criteria would require interpretation at both school level and teacher level (Skolverket, 2022b). The responsibility for providing education remained at state level, with terms and conditions regulated by the Education Act (SFS 2010:800) and the curriculum and syllabus specified by the National Agency for Education. The implementation, however, was largely transferred to municipality organizers and the organizers of independent schools. The educational system rapidly moved from centralization to decentralization, and many responsibilities were delegated from the national government to municipalities and individual schools (Wikström, 2005). As a result of these changes, Sweden has been described as having one of the most "decentralized and deregulated school systems in the world" (Gustafsson & Erickson, 2013, p. 12).

2.2.1. The criterion-referenced system

The criterion-referenced grading system, which replaced the norm-referenced grading system, was in part designed for monitoring the *quality* and *equality* of educational achievement (Gustafsson & Erickson, 2013, p. 70). In a norm-referencing system, a predetermined percentage of students would obtain a certain grade, regardless of achievement. In a criterion-referenced system, students are judged against predetermined standards or criteria, without regard to other students' performance, making it theoretically possible for every student to obtain the highest, or the lowest grade (Lok et al., 2016). The central concepts of such a system are goals and criteria where the goals state knowledge requirements and the criteria specify the extent of goal fulfilment. Although the criteria form the basis for grading, it has been argued that criteria can never be clear and explicit enough (Lok et al., 2016), causing The National Agency for Education to acknowledge that the criteria formulations cannot be sufficiently precise to solely guarantee grade equivalence (Skolverket, 2022a). A large part of the responsibility for achieving equivalent grading was thus placed on the teachers and the schools themselves, for example by organizing professional discussions about how the criteria should be interpreted (Gustafsson et al., 2014).

2.3. Assessment is Swedish Upper Secondary School

In Sweden, all grading is carried out by qualified teachers, and grading decisions are primarily based on classroom assessments. Grading should be a comprehensive assessment, based on all information available in relation to the knowledge requirements. This grading model relies on the teachers' ability to make accurate judgements and requires possessing the necessary assessment skills in combination with an expert knowledge of the grading criteria for the specific subject and course (Wikström, 2005).

2.3.1. Assessment equivalence

The Education Act dictates that everyone should have equal access to education, regardless of geographical location and social or financial circumstances (SFS 2010:800). Education should moreover be equal throughout every stage regardless of geographical location (Gustafsson et al., 2014, p. 22). However, the definition of equality is vague both in terms of definition and responsibility, and there is no specification of material or method to be taught or used for achieving educational equality. Equality, in this sense, is equated with students having the *same*

opportunity of reaching the next level, throughout both teaching and assessment (Skolverket, 2022a). The Education Act clarifies that equality does not mean uniformity, meaning that even though the teaching is and should be different, students should be given the same opportunity to progress and learn. In addition, the Education Act dictates that equality of education does not equate with the practical teaching being the same, nor that resources will be equally distributed (SFS 2010:800).

For assessments to be reasonably equivalent they should be based on as secure a foundation as possible, and a grade is no more reliable than the assessments on which they rest (Skolverket, 2012a). Assessment and grading should be directly linked to the governing documents, *syllabus* and *curriculum,* and be based on a research foundation (*vetenskaplig grund*) as well as proven experience (*beprövad erfarenhet)* (Skolverket, 2022c). The process of linking teaching and assessment to the learning criteria is referred to as *constructive alignment*. Clarification has been needed regarding the terminology of *proven experience*, seeing as it was open to interpretation and new to the teacher community. It has been established that it needs to be *documented, communicated, and shared*. Furthermore, it must be reviewed and tested based on ethical principles in a collegial context (RFR, 2012/2013). Consequently, proven experience is not equated with the tacit knowledge of a professional experienced teacher. The Department for Education states the following definition of equivalent grading:

> the grade in a certain subject corresponds to similar knowledge, regardless of which teacher the student had in the subject or which school the student attends (Utbildningsdepartementet, 2023).

2.3.2. The National Tests

The National tests are a set of *proficiency tests* with the primary function of acting as support in the grade calibration process (Wikström, 2005; NAFS, 2022), which were introduced as a means of a nation-wide evaluation of quality and equality of education (Gustafsson & Erickson, 2013). The tests are designed to have a high degree of authenticity, which is widely perceived as contributing to high test validity.

The National tests guarantee the testing of all students using the same materials and with similar contextual factors. They come with clear guidelines for combining partial test results to determine a test grade. However, there are no specific guidelines for how the test grade should be related to semester, final or course grades, more than that the exam result should be advisory, and be taken into *particular consideration* rather than governing the final grade (Skolverket, 2020). The fact that test results must be taken into *particular consideration* is explained as the results of the National tests *occupying a special position* in grading, which thus have greater significance than other assessments. No percentage or discrepancy between the grade of the National test and the final grade has been specified, only that these test results are of greater importance (Skolverket, 2020). The *reliability* of teacher assessment is thus reliant upon consensus in the interpretation of criteria, aided by additional instructions and examples unique for the National tests.

Assessment reliability is dependent on assessor consensus and uniformity which is traditionally low in the assessment of student essays (Gustafsson & Erickson, 2013). For purposes of evaluation of educational quality and equality, the Swedish school's inspectorate (SSI) conducts re-assessments of the National tests. The results are made public as well as presented to the Ministry of Education, to form the base for educational reform.

2.4. Assessing Upper secondary English

The curriculum states that all students should be granted the opportunity to develop an *all-round communicative competence* (Skolverket, 2022c). This competence entails understanding spoken and written language, expressing oneself and interacting with others in speaking and writing, as well as the ability to adapt one's language to fit different situations, purposes, and recipients (Skolverket, 2022c). The grade represents to what extent the individual student has met the national criteria.

2.4.1. Knowledge requirements

The knowledge requirements are presented as the *core content* of each subject, and the differentiation into separate grades is formulated by value words describing the particular qualities of skills and abilities required (Skolverket, 2020).

In the Swedish criterion-referenced grading system, only the knowledge requirements for A, C and E are specified. For B and D, respectively, the requirements for the lower grade must be met and to a large extent the requirements for the higher grade (Skolverket, 2020). To understand the intended meaning and minimize individual interpretation of these value words teachers are referred to a separate publication, namely *Kommentarmaterial till ämnesplanerna i moderna språk och engelska* (Skolverket, 2022b).

Table 1 below shows the requirements for oral and written productions equivalent to the grade C. Items in bold refer to the desired ability or skill: **varied, clearly**, and item in italics specifies the degree of criteria fulfilment. *Clarity* refers to linguistic precision and the term *relatively clearly* indicates that the student can use simple grammatical structures and patterns, convey a content and for the most part make themselves understood in oral and written presentations (Skolverket, 2022b). Accordingly, the progression of knowledge refers to an increased linguistic accuracy and variety with the addition of *structure*.

Table 1. *Progression of knowledge (Skolverket, 2022c)*

Year/Course	**Progression**
Year 9	relatively **varied**, relatively **clear** and relatively **coherently.**
English 5	*relatively* **varied**, **clear**, **coherently**, and *relatively* **structured.**
English 6	**varied, clear** and **structured.**
English 7	**Varied, balanced, clear,** and **structured.**

3. Previous research

Much of the research on summative assessment outside Sweden does not primarily focus on the final grades. Instead, international research on grading often focuses teachers' assessment practices, although grading reliability is also investigated (Lundahl et al., 2015). In Sweden there has been an increase in research on grading, which is explained by growing focus on grades and grade equivalence (Lundahl et al., 2015). Lundahl et al. claim the origin for this surge to be a quality review by The National Agency for Education (Skolverket, 2000), in which deficiencies in how teachers grade in relation to the governing documents were demonstrated. In this study, the function of grades as a selection instrument was questioned, as the grades do not appear to be fair or equivalent (Selghed, 2010).

3.1. Assessment validity and reliability

As previously mentioned, validity can be conditioned by sufficient assessment of the construct only as described in the criteria. The main threats to assessment validity are *construct underrepresentation*, where not enough data is measured, and *construct irrelevant variance*, where non-relevant data is measured (Erickson, 2018; Lundgren et al., 2020). In this section, international and Swedish research on factors affecting the validity and reliability will be briefly presented. The international results are included both as an introductory display of the field of research, and because the questions and findings are relevant to the study and hand.

Although test results remain the most decisive factor in teachers' grading practices, teachers can be affected by aspects of student behavior, such as motivation, attitude, classroom behavior as well and students' attitudes to homework (Lundahl et al., 2015). Similar results are documented in a Swedish study by *Vetenskapsrådet* (2015), showing that teachers of mathematics mainly focus on study results, and primarily on test results when grading, while language teachers to a greater extent consider behavioral aspects (cited in Lundahl et al., 2015). Teachers also tend to be more generous in their grading if the student is part of a high-performing student group (Bonesrønning, 2004, cited in Lundahl, 2015). The physical grading process is examined in a study by Klein (2002), showing that the order in which teachers correct examinations affects grading, in that teachers tend to set higher grades the more examinations they correct.

International research on grading from a teacher's perspective (Black et al., 2010) has examined the practice of grading and grading in relation to various standards and national knowledge tests. Among the findings are reports of grading being increasingly regulated and that teachers need to relate to different types of standards in their grading. As a result, the test result has become an increasingly dominant aspect of assessment. The same tendencies are reported in Sweden, where studies (Forsberg & Lundahl, 2006; Lundahl et al., 2015) have examined how an increased focus on test results could affect the teaching. Forsberg and Lundahl (2006) fear that an increased focus on students' results, such as the PISA measurements and the National tests, could affect how knowledge is valued, leading to changes in how teachers teach and a devaluation of teacher professionalism. Another concern is that an increased focus on quantifiable knowledge will lead to surface learning and ultimately to an erosion of the concept of knowledge (Lundahl et al., 2015).

3.2. Criterion-referenced assessments

This section will present a selection of research related to criterion-references. To begin, research about general issues, from international and Swedish studies relevant to the present study will be presented. To conclude, Swedish research will be presented to further focus the issues at hand.

Bloxham (2016) points to several risks of assessment variation in a criterion-referenced grading system. Unreliability in marking is well documented; yet few studies have investigated assessors' detailed use of assessment criteria (Bloxham et al., 2016). Assessment criteria can themselves be a potential cause for variability in assessment, and thus pose a threat to assessment equivalence. In addition to interpreting criteria differently, studies show that assessors "may not agree with, ignore or choose not to adopt the criteria" (Bloxham et al., 2016, p. 3). Another reason for variation is that although assessors may agree on what a construct, such as *developing argument* means, they may have differing understandings of what constitutes *excellent*, *adequate*, and *weak* in relation to *developing argument*.

The arguments made by Bloxham (2016) are equally relevant to the Swedish criterion-referenced system, where common, not profession-specific words are used to indicate precise differentiation of knowledge. A distinction can be made between *discrete differences* and *continuous differences* (Bergqvist & Mårtensson, 2015). The discrete criteria clearly define the degree of understanding and knowledge required for different grades, whereas continuous criteria consist of a sliding scale. Bergqvist and Mårtensson (2015) use the term *the adjective trap*, when describing the continuous criteria of the Swedish curriculum. They argue that the continuous criteria with its sliding scale of criteria fulfilment, as demonstrated in Table 1., makes relating criteria to student production difficult to both teachers and students. Consequently, how criteria are expressed can necessitate interpretation, further hindering assessment equivalence (Bergqvist & Mårtensson, 2015; Gustafsson et al., 2014).

Following the introduction of the criterion-referenced grading system, the question of equivalence has been central, and several studies point to significant shortcomings (Skolinspektionen, 2010; Gustafsson & Erickson, 2013; Gustafsson et al., 2014; Utbildningsdepartementet, 2023). The main reason is related to the necessity for interpretation of criteria, as a shared understanding of criteria is essential for grade equivalence (Bergqvist & Mårtensson, 2015; Lundahl et al., 2015; Lok et al., 2016; Roth et al., 2016; Bloxham et al., 2016).

The main findings in teachers' assessments of complex tasks are accredited to *low interrater reliability*, even when having access to detailed assessment instructions (Jönsson & Thornberg, 2014). The main problem is that teachers lack a common interpretation of the governing documents and how student performance should be assessed. True assessment equivalence dictates that equivalent performances should be assessed equally and requires teacher consensus on the governing documents and a shared understanding of criteria. Otherwise, assessment equality risks being mere *surface veneer*, or having the appearance of equivalence (Black & Wiliam, 2018; Jönsson & Thornberg, 2014).

3.3. Formative and/or summative assessment

Much of the recent research on assessment focuses solely on the benefits of formative assessment as a tool for learning. The seminal works of Hattie and Timperley (2007) and Wiliam (2011) are now mandatory readings in teacher training programs and are referred to by the National Agency for Education, arguably at the expense of summative assessment. Seeing that formative assessment focuses on the process, *student learning*, and not of the product of learning, *the grade*, it can be perceived by policy makers and teachers alike as a superior form of assessment. In doing so, one risks losing focus of the summative aspects of assessments and the necessity for *validity, reliability,* and *comparability* in all assessments (Black & Wiliam, 2018).

When advocating formative assessment, its potential to impact student learning is often highlighted (Hattie & Timperley, 2007; Lau, 2014; Wiliam, 2019). However, the positive effects of well-designed summative assessment have also been documented (Bennett, 2011; Lau, 2014). Actually, there is research claiming that the dichotomy of summative and formative assessment is a widespread myth in assessment research (Black & Wiliam, 2018; Lau, 2014). Studies (Taras, 2005; Lau, 2014; Black & Wiliam, 2018) demonstrate that the perceived dichotomy was in fact, created by researchers and further reenforced by policy makers. Black and Wiliam (2018) argue that summative assessments have been given a more prominent role due to increasing demands for accountability, causing teachers to *teach to the test*. According

to Lau (2014), summative and formative assessments "need to work in harmony and should not be seen as contrary to each other" (p. 509).

When summative and formative assessment occur simultaneously, students often focus on the summative, the grade, at the expense of the formative part of assessments (Jönsson, 2020). However, such a statement is not to be taken as proof of the detriment of summative assessments on student learning and motivation; rather there are numerous studies showing the benefits of summative assessment on student learning (Bennett, 2011; Lau, 2014). Biggs (1996) used the term *backwash* to describe the idea of students (only) focusing on what they will be tested on. Backwash is often used when describing the negative aspects of summative assessments as it is believed to only lead to surface learning. However, Biggs (1996) argued that it is only natural that students focus on the grade, given that good grades hold the key to a better future. Instead of stigmatizing and abandoning summative assessments he proposed capitalizing on this student motivation by better alignment of assessment and criteria, *constructive alignment*. The studies by Biggs (1996), Taras (2005), Bennett (2011), and Lau (2015) demonstrate the effective use and positive benefits of the combined use of FA and SA.

3.4. Assessment equivalence

Grades are debated and linked to a discourse about the decay of the school system, questioning their equivalence (Selghed, 2010; Lundahl et al., 2015). Gustavsson (2012) claims that no grades can be truly equal, referring to the re-assessments of the National tests conducted by the SSI, which made it clear that teachers assess the same performance differently.

Since the late 1980s, the Swedish educational system has been subject to several changes, not least concerning the systems for knowledge assessment at both student and system level and research shows that the Swedish school's knowledge assessment system does not meet the quality requirements necessary (Gustafsson et al., 2014; Gustafsson & Erickson, 2013; Gustavsson et al., 2012; Jönsson & Thornberg, 2014; Utbildningsdepartementet, 2023). The criterion-referenced grades suffer from a lack of equivalence between teachers and schools, and in particular secondary school grades and some primary school grades are affected by inflation (Gustafsson et al., 2014). Selghed (2004) argues that few teachers understand the criterion-referenced grading system and the view of knowledge that should be the basis for the grading, resulting in teachers weighing in other factors in their assessment and comparing the students with each other (cited in Lundahl et al., 2015, p. 43).

3.4.1. The National tests

Several studies (Gustavsson et al., 2012; Gustafsson et al., 2014; Utbildningsdepartementet, 2023) have reported issues of equivalence related to the National tests. The re-assessments conducted by the SSI (Skolinspektionen, 2010) concluded that there were large variations between assessments done by teachers and the re-assessments *i.e., low interrater consistency*. In most of the cases, the grades given by teachers were higher than those of the external assessors (Skolinspektionen, 2010).

One method of increasing the reliability of complex assessments is by increasing the number of independent assessors, arguing that the resulting summary assessment would have good reliability. Such a system is, however, resource-intensive, especially finding multiple independent assessors from different schools. Beyond increasing reliability, such a practice could be valuable in further developing teacher assessment skills by providing teachers with

direct experience of how the variation in students' abilities is portrayed at the national level (Gustafsson et al., 2014).

A second problem emphasized by the SSI was that the tests and the assessment instructions leave wide margins for interpretation, not supporting comparable assessment (Skolinspektionen, 2010). This applied in particular to essay writing, a task which by design is less reliable (Gustafsson & Erickson, 2013). For instance, the results of the 2010 and 2011 re-assessments found substantial differences between the original marking and the remarking for certain subtests, particularly for the essay parts. In most cases the differences were negative, meaning that the SSI assigned a lower mark than the teacher (Skolinspektionen, 2010; Gustafsson & Erickson, 2013).

At present, the Education Act dictates that the grade of the National tests should be taken into *particular consideration* when assigning the final grade. However, there are no regulations specifying the degree of variation between the grade of the National test and the final grade. Even though the SSI has reported large deviations in equivalence (Skolinspektionen, 2010), The National agency for Education still claim that the National tests provide support for equivalent and accurate assessments and grading of students' knowledge (Skolverket, 2020). In an attempt to increase grade equivalence The National Audit Office has proposed legislative measures forcing a closer correspondence between the grade of the National test and the final grade (Utbildningsdepartementet, 2023). Formalizing acceptable measures of deviation at group or school level, could support teachers in their assessment and grading, and thus, increase the degree of grade equivalence (Utbildningsdepartementet, 2023).

3.4.2. Peer-assessment

Peer-assessment is presented as a means of achieving a higher degree of reliability, which would, inevitably, be a step toward grade equivalence. One proposed method of increasing the reliability of complex assessments is by increasing the number of independent assessors, arguing that the resulting summary assessment (of all assessors) would have good reliability. Beyond increasing reliability, such a practice could be valuable in further developing teacher assessment skills by providing teachers direct experience of how the variation in students' abilities is portrayed at the national level (Gustafsson et al., 2014). Black and Wiliam (2018) caution that an increased dependency on external assessors for high-stakes assessment could counteract the desired learning outcomes, as focus shifts from *learning* to *test scores*.

There is a strong belief in the ability of peer-assessment to contribute to greater equivalence in the teachers' assessment and grading, not least by the teachers assessing students' performances from the National tests together (Jönsson & Thornberg, 2014; Skolverket, 2020). However, having conducted a review of the research on the effects of peer-assessment, Jönsson and Thornberg (2014) found no clear evidence of peer-assessments resulting in a higher degree of assessment equivalence. Although peer-assessments may increase the degree of criteria consensus on the local level, it has limited power to affect the national level (Jönsson & Thornberg, 2014).

3.5. Accountability

As mentioned, Black and Wiliam (2018) argue that summative assessments have been given a more prominent role due to increasing demands of accountability. Given the standardized nature of the Swedish National tests, there is reason to believe that these tests are *high stakes*

for both students and teacher, thus increasing the pressure to perform. Interviews suggest that teachers, also, feel pressured that their students succeed on the National tests (Roth et al., 2016).

As previously stated, assessments cannot by themselves be formative or summative. Assessments, when used for purposes of accountability, tend to shift focus away from student learning (Black & Wiliam, 2018). Given the evidence of the impact of formative assessment practices on student achievement on standardized tests (Black & Wiliam, 2018), more effort could be put into assisting student learning.

As a result of the curriculum reform in 1994, the political significance of grades has shifted towards being used as a measure of school quality. The *good school*, and by extension the *good teacher*, is thereby defined based on students' goal fulfillment. In this sense, the function of grades has shifted from measuring student knowledge towards measuring school quality in relation to goal fulfillment (Lundahl et al., 2015). With an increased focus on high stakes testing with standardized assessment and grading, the teachers' role as a professional assessor is diminished (Lundahl et al., 2015).

4. Theoretical perspective

There are many theoretical approaches available for the study of assessment. For the study of the purpose of assessment, one must first decide on what to focus on: *the student, the learning process, the institution,* and/or *external stakeholders*. It would be possible to assume a more traditional view of knowledge as being quantifiable and examine aspects of reliability in high stakes assessments connected to accountability and psychometrics. The 'traditional' definitions of validity and reliability are connected with psychometrics, a branch of psychology dealing with the design and interpretation of qualitative tests for measurements of intelligence, aptitude and personality traits, which are often perceived as fixed (Gipps & Farajnezhad, 2022). Lau (2014) also demonstrates the link between assessment and trait theory, in which qualities such as capacities and intelligence are measured in relation to others. According to such a view, education and assessment primarily serves as preparation and selection for future education and employment. Key elements in this 'traditional' view of assessment are standardization and reliability, often at the expense of validity, giving such assessment an *illusionary veneer of 'objectivity'* to the idea of assessment. In a criterion-referenced system, when assessments are made against standards, the key feature is *validity*, ensuring that assessments are aligned with the criteria (Lau, 2014). Such a view also points to areas of tension within the Swedish curriculum. The action-oriented views of the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2020) adopted by Swedish policy makers strongly emphasizes the importance of knowledge through mediation and the necessity for assessment authenticity: what students can do is far more important that what students know (Skolverket, 2022c).

From the perspective of the assessor, the *process* of assessment can be observed as a cognitive act, wherein the professional interprets the student performance in relation to their personal subject knowledge and understanding of the criteria. Li (2020) argues that teacher cognition, or teacher *beliefs*, are affected by their learning and teaching experience, the teaching practice, and the workplace culture.

> What teachers think, believe, and perceive could strongly influence the way teachers plan their lessons, the activities, and tasks they design, the evaluation of learning, and all kinds of decisions they make in the teaching process (Li, 2012; Pajares, 1992, cited in Li, 2020).

In this sense, teacher cognition is relevant in relation to assessment as it can give an insight into teachers' decisions about what knowledge is regarded as relevant and what goals should be accomplished (Speer, 2005, p. 365, cited in Li, 2020). Teacher cognition refers to "what teachers know, belief, and think" and how this affects their pedagogical decisions (Borg, 2003, p. 81). By observing what teachers do and say in the classrooms a link can be made between the personal beliefs and the teaching practice (Li, 2020). Li (2020) discusses teacher cognition from different theoretical perspectives. According to her, the cognitive perspective views teacher cognition as static, or fixed mental entities that exist in teachers' minds and which guides their action and decision-making. Teachers' beliefs are "considered as a person's static traits that remain constant across situations" (Li, 2020). Unlike the views of the Sociocultural Perspective, the role of context is as explanation for teachers acting contrary to their beliefs. From a cognitive perspective, there is overwhelming research evidence demonstrating the influence of teachers' beliefs on their classroom practices (Li, 2020).

Li (2020) also demonstrates how the role of context and culture is highlighted in the sociocultural view of second language acquisition, where the construction of meaning is mediated through language and the sociocultural context. Lantolf and Thorne (2006) claim that the Sociocultural theory "offers a framework through which cognition can be systematically investigated without isolating it from social context" (2006, p. 1, cited in Li, 2020). Borg (2003) also emphasizes the importance of context, claiming that "the study of cognition and practice without an awareness of the context in which these occur will inevitably provide partial, if not flawed, characterisations of teachers and teaching" (p. 106). Li (2020) concludes that teacher cognition from a Sociocultural perspective is fluid and interactive understanding, situated in a given context, and that cognition is *social*, and not considered a state at a cognitive level.

The assessment process can also be studied through a sociocultural lens, examining the effects on the assessment of mediating tools and contextual factors, such as colleagues, peer-rating, the professional community in which the assessment takes place. Bloxham et al. (2011) emphasize that all assessment cultures are socially constructed, and that the ways in which assessments are conducted as well as what is to be rewarded is something that is under constant renegotiation within every learning environment. Grading criteria get their collective meaning only within in a certain sociocultural context and within the same context or assessment culture, written grading criteria can be interpreted differently by different teachers, which is why an ongoing discussion of criteria interpretation and implementation is vital (Bergqvist & Mårtensson, 2015).

Black and Wiliam (2018) argue that assessment cannot be understood without a consideration of the wider context within which that assessment takes place. "Teachers and schools are constrained by the cultural traditions, the political and public expectations of education, and the norms of the various institutions within which they operate" (Black & Wiliam, 2018, p. 570). The use of a criterion-referenced system by design acknowledging the need for a Sociocultural perspective of knowledge as being mediated in a social context (Lundgren et al., 2020). Accepting that the essence and definition of a criteria can never be precisely captured eliminating the need for interpretation, its meaning needs to be operationalized through mediation and co-construction (Lok et al., 2016). Accordingly, a contextual foundation, such as a *community of practice,* can be regarded as a prerequisite for criterion-referencing. Though *individual* in terms of performance, assessments are *collective* due to the necessity of a common understanding (Allal, 2013).

Allal (2013) proposes a view of the assessment process to be a combination of individual psychological processes and shared social practices to be joined in a reflexive, mutually constitutive relationship. She argues that although teacher assessments are generally carried out individually, they are shaped by the collective practices of a professional community. Adapting such a perspective allows an acknowledgment of teacher professionalism, in that it requires the combination of knowledge of the student, subject knowledge together with institutional norms and regulations, both formal and tacit (Allal, 2013). Assessment skills are often described as developed with experience, where teachers have an almost intuitive understanding of the quality of student work (Gustavsson et al., 2012). Such a view implies that assessment skills to a large part consist of *tacit knowledge*. However, studies show that the assessments of novice assessors do not differ from those of more experienced assessors (Bloxham et al., 2016; Roth et al., 2016). Allal (2013) stresses the importance of social moderation when constructing shared assessment practices as a means of overcoming the tensions between the official standards and the tacit standards developed through classroom experience (*Swedish: beprövad erfarenhet)* (Allal, 2013).

For this study, the main theoretical focus lies on *teacher* cognition, examining the process of assessment by looking at teacher beliefs, here specifically "*what makes an essay good?"* as well as teachers' perceptions of the meanings of the value words of the knowledge requirements. The notion of assessments being both a cognitive act *and* a socially situated practice as proposed by Allal (2013), cannot be examined in this study. However, this study acknowledges that assessors form part of local assessment cultures, which is why potential results of intrarater discrepancies would be worth investigating from the perspective of Allal (2013). Finally, demands conflicting due to a difference of theoretical approaches, such as curricular demands and accountability, which could affect the assessments, will be addressed.

5. Method

In this section, the selected methods for the study are presented. The selection of student essays and the design of the questionnaire are explained together with the selection of participants and the methods of analysis. Finally, validity and reliability are discussed along with the ethical considerations.

5.1. Material and method

For data collection, four authentic student essays were chosen from a corpus provided by a former teacher training supervisor. The number was limited to four, as to not be too time consuming to the external assessors. The essays selected are of various lengths and on different topics, displaying a variety of linguistic features, qualities and 'errors'. Thus, the external assessors would be faced with a variety of features when grading, such as linguistic variation, content and structure, content versus 'language mistakes'. Two essays were marked B, and the remaining two C, and D. As there are no specific criteria for the grades B and D in the Swedish curriculum, this choice was believed to further fucus the assessments with the criteria, as the assessors would be compelled to further justify not assessing A, C, or E. In doing so, a clear connection with the criteria would be needed, fitting with the purpose of this study.

The student essays were originally written in *DigiExam*, and have, for the purpose of this study, been copied to a Word document as well as into a PDF-file. They have been assigned a uniform font for both the body and headline of the text. Spellings and paragraphs are true to the original. Also, the writers' names have been removed to preserve anonymity.

5.1.1. The students' essays

All essays selected for this study have been anonymized, minimizing the risk of peer-assessors having a personal relationship with the student writers (Gustavsson et al., 2012). When the SSI conducted re-assessments (re-grading) of the National tests in English they concluded that the *interrater reliability* was low, but also that there was no way of knowing which marking was correct, the original marking or that of the peer-raters (Gustafsson & Erickson, 2013). This study replicates the conditions of the SSI, in that there is a pre-existing assessment containing a summative and a formative assessment, though on a much smaller scale in terms of samples and co-raters. This study does not claim to be representative of the population (teachers grading trends) at large due to the above-mentioned factors.

5.1.2. The questionnaire

A questionnaire was constructed in *Microsoft Forms*, chosen for efficiency in construction, distribution to informants and in collecting answers. No data has been stored on external platforms. The questionnaire, which was used for the collection of quantitative and qualitative data comprised of a total of 37 questions utilizing a selection of multiple-choice answers, according to a Likert scale ranging from *strongly disagree* to *strongly agree*, and in-depth explanation. Of the 37 questions, 36 were mandatory.

In the questionnaire the respondent teachers were asked to answer questions relating to their assessments, as well as motivating their marking in accordance with their summative assessments. The underlying design of the questionnaire was the examination of a summative assessment in relation to the explicit criteria of *structure, clarity, cohesion, variation, adaptation to purpose/recipient* and *flow*. Therefore, the respondents were asked to answer how each essay related to these criteria according to a Likert scale, the underlying theory being that this would show the relation of each text to theses explicit criteria as well as displaying tendencies of favor, in terms of criteria, from the respondents.

There are a total of nine questions concerning the assessment of each essay (see Table 2). The reason for this design choice is two-fold: firstly, informants are required to motivate their assessments directly correlating to each criterion, thus decreasing the risk of *factor-irrelevant variation,* and potentially increasing the validity of the assessments. Secondly, any individual criterion-preferences in the task of essay writing from individual assessors could be revealed. To conclude, two questions were asked pertaining to the *validity* and *reliability* of the assessments, followed by a non-mandatory option for further comments.

Table 2. *The questionnaire, Essay 1*

1. I assess Text 1
2. I am confident in my assessment of the text
3. This text meets the requirement of **struktur**
4. The text meets the requirement of **tydlighet**
5. The text meets the requirement of **sammanhängande text**
6. The text meets the requirement of **variation**
7. The text meets the requirement of **anpassning till syfte och mottagare**
8. The text meets the requirement of **flyt**
9. What were the most important features that decided the grade for this essay?

For question 1, the respondents were asked to assign a grade, F-A. For questions 2-9, the respondents were asked to respond according to a Likert scale ranging from strongly disagree to strongly agree. Question 9 is an open-ended question providing qualitative data.

5.1.3. Selection of informants

The informants for this study consist of the teacher who originally marked the essays, henceforth referred to as *Teacher*, and three external assessors, referred to as *A1, A2* and *A3*. Admittedly, this study would have benefitted from having a larger selection, as this would have increased the possibilities of generalizing based on the data received. Due to time constraints, however, the decision was made to proceed with the three external assessors available, rather than postponing the study.

The *Teacher* has previous knowledge of the students, having taught the group which forms the basis for the selection for English 5. The *Teacher* is a native speaker of English and a senior teacher (*förstelärare*). Having originally graded the students' essays, *the Teacher's* grades are used in comparison with those of the external assessors.

The external assessors, *A1-A3,* are experienced professionals, working at different schools and with no previous knowledge of the student group. As it is not the purpose of this study to relate the grading to years of experience, sociological factors or gender of students and assessors such data does not form part of the criteria for informant selection. Specifically for this study, the external assessors have assessed the students' essays, and answered the questionnaire relating their assessments to the specific criteria.

Before starting the study, the respondents were given background information regarding the study. Upon agreement of participation, they received a digital invitation to the questionnaire, followed by an email with the letter of informed consent (Appendix 1) and the student essays in two formats, Word*,* and PDF (Appendix 3). The questionnaire (Appendix 2) was completed digitally via Microsoft Forms, at a time convenient to the respondents, similar to the re-assessments of the National tests conducted by the SSI (Gustafsson & Erickson, 2013).

5.1.4. The pilot study

A pilot study was conducted to test the reliability and validity of the methods. Though, limited to only one participant, an English teacher student, the feedback from the pilot study proved valuable to the final study. The results of the pilot study prompted alterations of the questionnaire as comments suggested unclarity of purpose.

Firstly, in the pilot study, the grade criteria were included. However, due to changes in the curriculum for English, the definitions were not up to date. In the final study, the grade criteria were omitted. Secondly, in the questionnaire for pilot study, the possibility to mark the grade E, was missing. This was added in the final study. Thirdly, three questions were removed from the questionnaire, as their answers would have been speculative and based on personal belief, rather than research based. The first question asked if other teachers would assess similarly, and the following two questions pertained to the validity and reliability of the own assessment.

5.2. Method of analysis

In analyzing and contrasting the data, the grades were assigned a numerical value of 1-5 (E-A). The same was done in correlation with criteria where *strongly disagree* was assigned *1,* and *strongly agree 5.* In doing so an in-depth contrastive analysis would be made possible. For comparison, the original marks are included, to discuss any tendencies of overmarking by original teachers compared to re-assessments (Gustavsson et al., 2012; Gustafsson & Erickson, 2013). Following the arguments of Gustafsson and Erickson (2013), this paper makes no claim about the correctness of the markings. Still, this study hopes to address the reported 'trend' of

over-marking, and any potential implications with regards to *validity, reliability,* and *comparability.* To illustrate differences in the assigned numerical values, the symbol Δ is used, representing variation. For instance, the variation (Δ) between *strongly disagree* (1); and *strongly agree* (5) is described as Δ4.

The quantitative data have been analyzed by comparing the numerical values for grades and for the criteria with the original grades of the *Teacher*. From that, comparisons were made by examining individual essays and individual criteria. For the grades, the assigned numerical values from the *Teacher* and *A1, A2* and *A3*, have been compared for each essay. As for the individual criteria, mentioned in the questionnaire, the analysis is restricted to the data from *A1, A2* and *A3*, as no such data is available from the *Teacher*. The data regarding the individual assessment criteria have been analyzed in relation to the grade given, to uncover criteria correspondence with the grade.

The qualitative data have been analyzed by examining what the assessors have commented on, as well as what is not commented on. Commonly occurring comments have been grouped into themes to discover what the assessors have focused on in their assessments. Finally, the quantitative and the qualitative data have been cross-referenced to examine whether they corroborate each other.

5.3. Validity and reliability

Validity in research is concerned with measuring what is claimed to be measured. In quantitative research, validity is concerned with data replicability, whereas for qualitative research, validity is also concerned with the conclusions drawn from the data (Cohen et al., 2018).

While the basis of reliability differs between quantitative and qualitative research, "[R]eliability is essentially an umbrella term for dependability, consistency and replicability over time, over instruments and over groups of respondents" (Cohen et al., 2018, p. 268). Cohen et al. (2018) list stability, equivalence and internal consistency as fundamental principles for qualitative and quantitative research seeking trends and patterns. Here, stability is connected with the instrument for data collection, as "[A] reliable instrument in a piece of research yields similar data from similar respondents over time" (Cohen et al., 2018, p. 268).

The present study consists of both quantitative and qualitative data. To ensure validity and reliability the data have been analyzed both separately and together, by a method of triangulation. According to Cohen et al. (2018) there are several benefits to using triangular techniques, in giving a broader explanation of the phenomenon studied, and reducing researcher bias. "If, for example, the outcomes of a questionnaire survey correspond to those of an observational study of the same phenomenon, the more the researcher can be confident about the findings" (Cohen et al., 2018, p. 265). Moreover, by making use of both qualitative and quantitative data, triangulation can help improving the validity of the study by "demonstrating concurrent validity" (Cohen et al., 2018, p. 265).

In the first stage, the qualitative data were collected and analyzed. In the second stage, the quantitative data, which consist of the external assessors' comments of their assessments, were collected, and analyzed. Finally, the quantitative and the qualitative data were compiled and contrasted to examine how they complement each other, and if there were any discrepancies.

While a reproduction of this study, with different external assessors, could affect the replicability of the results, i.e., resulting in different data, the validity of this study is ensured, as it demonstrates concurrent validity. The reliability of the instrument for data collection, the questionnaire, has been tested in the pilot study, after which, alterations were made to remove ambiguity.

5.4. Ethical considerations

The essays selected for this study have been anonymized, minimizing the risk of peer-assessors having a personal relationship with the student writers (Gustavsson et al., 2012). Essays reveling details of ethnicity, gender, geographical location could have been omitted in an attempt not to provoke bias from the assessors. For instance, an assessor might be tempted to assign a higher grade to a student who is relatively new to Sweden, therefore potentially not having had equal amount of English education as their peers. However, seeing as one of the topics for writing was *Becoming Swedish*, no such decision was made.

It is possible that the selection of students' essays affected the outcome, and that a different selection of essays, with a larger selection of informants could have yielded different results. Nonetheless, the results of this study are considered valid and reliable, as the stability of the questionnaire has been tested and the results of the study demonstrate concurrent validity.

All respondent data is anonymous. Before conducting the study, all informants were sent a letter of informed consent (Appendix 1), in which the aim of the study was explained. The informants were informed that they could cancel participation at any time and that the data received would be treated in accordance with the General Data Protection Regulation (GDPR) and that all data would be deleted upon the completion of the study.

6. Results

The presentation of the results of the study is divided according to type of data, starting with the qualitative data, the grades, and their correlation to the criteria, and ending with the qualitative data, the external assessors' comments.

6.1. The quantitative data

The qualitative data consist of each assigned grade, converted to numerical values, as well as the assigned relationship between the external assessors' grading and the assessment criteria for the task of *written production* (Skolverket, 2023), as documented in the questionnaire. The data analysis displays great variations in both assigned grades and in criteria fulfillment.

In the following sections, grade variation, criteria interpretation, and consensus, as well as the relationship between grades and criteria will be presented. In conclusion, the matter of assessment confidence will be discussed, as this could reveal the assessors' understanding of the grading criteria.

In the first subsection, *the grades*, the analysis is based on data from the *Teacher* and the *external assessors*. In the following subsections, however, the analysis is based on data provided by the *external assessors* only. This is because the *Teacher* graded the students' essays before the questionnaire was designed. Thus, the *Teacher's* contribution to this study consists of providing the assessed students' essays.

6.1. The grades

For data analysis, each grade was assigned a numerical value, F = 1 – A = 6, and the results reveal great differences in grading, with a variation (Δ) of 2 - 4 grade steps per essay. Table 3 demonstrates that the grade difference varies between Δ2 - Δ4, i.e., *low interrater consistency* The variation is greatest for *Essay 1*, where the grades vary from F – B (Δ4) and smallest for *Essay 4*, varying from F – D (Δ2). Essay 4 also has the highest degree of assessor agreement, i.e., *interrater consistency*, with all external assessors agreeing on an F.

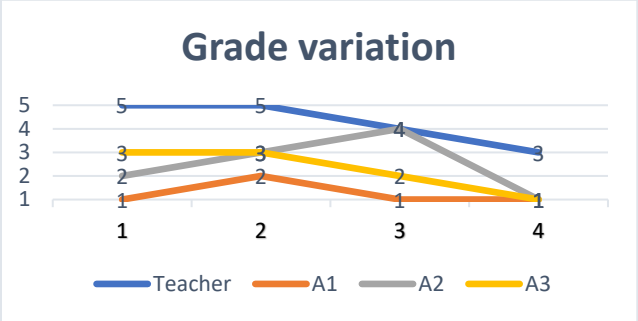
Table 3. *Grade variation, all essays.*

	Essay 1	**Essay 2**	**Essay 3**	**Essay 4**
Teacher	B	B	C	D
A1	F	E	F	F
A2	E	D	C	F
A3	D	D	E	F
Δ	4	3	3	2

By examining the assigned grades, the degree of interrater consistency becomes apparent. In total, there is assessment agreement in 7 out of 16 assessments (see Table 3 and Graph 1). For Essay 2, *A2* and *A3* have graded D, for Essay 3, *the Teacher* and *A2* have graded C, and for Essay 4 *A1, A2,* and *A3* gave graded F.

In Graph 1, the grade variations previously discussed are illustrated, making it apparent that the grades of the *Teacher* are higher, or significantly higher compared to the grades of the external assessors. There is only one point of correlation between *the Teacher's* assessment and those of the external assessor, which is for Essay 3. Here the *Teacher* and *A2* both graded C, which equals 4 points.

Graph 1. *Grade Variation*



6.2. Grades and criteria

In this section, the relationship between the assigned grades and the criteria scores will be demonstrated. A high degree of assessment consensus, or *interrater reliability*, indicates consensus in criteria interpretation, which is a prerequisite for assessment equivalence. In contrast, a high degree of interrater criteria consensus, while still grading differently, could be a sign of *factor irrelevant* variation, threatening the validity of the assessments.

For Essay 1 (see Table 4), although having assigned three different grades, there is a consistency between the total score for the criteria and the grade. According to all the assessors there is an increase in criteria points corresponding to a higher grade.

Table 4. *Essay 1, Criteria distribution, and variation*

	Grade	Structure	Clarity	Cohesion	Variation	Adaptation	Flow	Score
A1	F	2	2	3	1	3	3	14
A2	E	4	3	3	3	4	1	18
A3	D	4	4	4	4	3	4	23
Δ	2	2	2	1	3	1	3	9

The same logical correlation between criteria scores and grades does not apply to the assessments of Essay 3 (see Table 5). Here, *A3* assigns high scores for all criteria (24 points), and marks the essay with an E. Simultaneously, A2 marks the Essay with a C, while assigning 21 points. This could be indicative of *intrarater inconsistency* and *factor irrelevant variation*. As the external assessors all have graded, there is no *interrater consistency*. Given that the criteria of the questionnaire correspond with the criteria of the curriculum, this warrants further investigation.

Table 5. *Essay 3, variation of grades and criteria*

	Grade	Structure	Clarity	Cohesion	Variation	Adaptation	Flow	Score
A1	F	2	3	3	2	3	3	18
A2	C	3	3	4	4	3	4	21
A3	E	4	4	4	4	4	4	24
Δ	4	2	1	1	2	1	1	8

From looking at the data one might expect that the highest grade would be assigned by A3, with a total score of 24 points, the second highest score for all the assessments (see Table 15). Instead, the highest grade, C, was assigned by A2, despite scoring three points lower than A3. The data reveal that A3 values Essay 3 higher than A2 for 3/6 criteria: *structure, clarity,* and *adaptation*.

6.2.1. Grade parameters

By calculating the criteria score for each grade, the grade points parameters established in this study have been made visible (see Table 6). With a consensus of criteria understanding, there ought to be a clear correlation between criteria score and the grade, resulting in high *interrater consistency*. At present, the data indicate inconsistencies in the correlation between criteria points and the grade.

Table 6. *Grade parameters.*

Criteria points	Grade
11 – 18	F
18 – 24	E
19 – 23	D
21 –	C

The overlap of E – D – C could be accounted for by the fact that the external assessors have been given the possibility to mediate an understanding of criteria together. Consequently, the outcome can be considered a representation of the necessity for mediation in achieving consensus in criteria interpretation. Nonetheless, the fact that the parameters for *E* and *D* are almost identical is concerning from a grade equivalence perspective. Another explanation may be that grading a student essay by criteria alone is insufficient as six criteria cannot fully capture the qualities of the student performance.

6.2.2. Interrater Consistency

From a total of 12 re-assessments there is very low interrater consistency between the assessments of the external assessors. The highest degree of assessment consensus, *interrater consistency*, can be found for Essay 4 (see Table 7), where all the external assessors agree on the grade: F. Despite grade consensus, there is still some variation in the total criteria scores: *A1* – 13, *A2* – 11 and *A3* – 15, resulting in a variation of Δ4.

Table 7. *Variations of grades and criteria, Essay 4*

	A1	A2	A3	Total score	Δ
Grade	F	F	F		
Structure	2	2	3	7	1
Clarity	2	1	2	5	1
Cohesion	2	2	3	7	1
Variation	2	2	3	7	1
Adaptation	3	3	2	8	1
Flow	2	1	2	5	1
Total score	13	11	15		4
Mean score	2.16	1.83	2.5		

In contrast, there is no interrater consistency for Essay 1 (see Table 8). All the external assessors have assigned different grades: F, E, D, resulting in a grade variation of Δ2. The total criteria scores range from 14 – 23, which is not surprising, given the grade variations. However, there is consistency between criteria score and grade, since the lowest grade has the lowest criteria score, and the highest grade has the highest criteria score.

Table 8. *Variations of grades and criteria, Essay 1*

	A1	A2	A3	Total Score	Δ
Grade	F	E	D		
Structure	2	4	4	8	2
Clarity	2	3	4	9	2
Cohesion	3	3	4	10	1
Variation	1	3	4	8	3
Adaptation	3	4	3	10	1
Flow	3	1	4	8	3
Total score	14	18	23		9
Mean score	2.33	3	3.83		

6.2.3. Intrarater consistency

By comparing all grades set by a single assessor with the corresponding criteria points, the *intrarater consistency* becomes visible. Table 9 shows intrarater inconsistency for the assessments of A3, between Essay 1 and Essay 3. Essay 1 received a D with 23 points, whereas Essay 3 received an E with higher points (24). The assessments of A2 show good intrarater consistency, but with narrow steps between the grades: 11p – F, 18p – E, 19 p – D and 21 p – C.

Table 9. *Intrarater consistency*

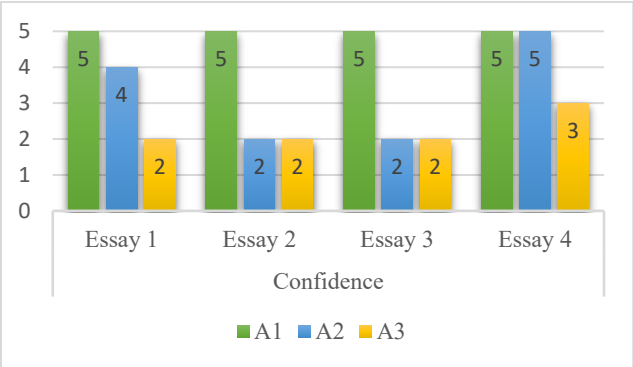
	Essay 1	Essay 2	Essay 3	Essay 4	Conclusion
A1	14 – F	21 – E	18 - F	13 - F	*Consistent*
A2	18 – E	19 – D	21 – C	11 - F	*Consistent, but narrow*
A3	23 – D	26 – D	24 – E	14 - F	*Inconsistent*

6.2.4. Grade and confidence

The question of the degree of confidence in one's assessment can relate equally to assessment skills (Swedish *bedömarkompetens*) and knowledge in relating the student performance with the criteria, i.e., *criteria clarity*. As previously demonstrated, the criteria of the Swedish curriculum require interpretation, which is why a higher degree of assessment confidence could be indicative of a high degree of *proven experience,* (*Swedish: beprövad erfarenhet*), in relation to criteria understanding. The latter proving more important as there are no specified criteria for *D* and *B* (Skolverket, 2022c).

To answer the question "I am confident in my assessment of the text" (see Appendix 2), A1 *Strongly agrees* for all four essays. A2 vary between *Agree* (Essay 1), *Disagree* (Essays 2-3) and *Strongly Agree* for Essay 4. A3 expresses the least confidence with *Disagree* (Essays 1-3) and *Neutral* for Essay 4. The variation in confidence is smallest for Essay 4 (Δ2), where there is consensus for the grade, F. For Essays 2 and 3 the variation in confidence is the greatest, Δ3, which is interesting as these two essays have the largest grade variations. As demonstrated in Graph 2, the assessment confidence varies greatly.

Graph 2. *Assessment confidence*



The highest grade assigned by the external assessors is a C, by A2 for Essay 3, despite a low level of assessment confidence. A1 has the lowest assessment score of all assessors and is simultaneously, the most confident in the assessments. A lack of assessment confidence is, according to the data, not synonymous with assigning a lower grade.

6.3. The qualitative data

The qualitative data yields information about the assessment practice and particular difficulties in relation to the assessments of these four essays. The question of what the assessors deemed "the most important features that decided the grade for this essay" (see Appendix 3) gives insight into the reasonings and mediations of each assessor in assigning the grade.

The qualitative data were labelled according to type of comment and then categorized. Thus, the comments could be related to the assessment criteria: *variation, clarity, flow, cohesion, structure,* and *clarity*. By examining the *Commentary material* (Skolverket, 2022b), the

intended meaning of each criterion could be examined together with the external assessors' categorized remarks, to examine their correlation.

6.3.1. Language and structure

The most common remark is that of *language mistakes* and the comments primarily focus on mistakes and errors. Of the total 39 comments, 34 were regarding language mistakes (see Table 10). To understand the intended meaning behind the value words of the curriculum teachers are referred to the *Kommentarmaterial* (Skolverket, 2022b). There it becomes evident that there are criteria overlap in meaning. Both *variation, clarity* and *flow* address how the students' use of words and phrases, sentence structure and text binding contribute to making the presentations easy to follow (Skolverket, 2022b). Text cohesion is also mentioned for *cohesion* and *structure*, which is why further assigning the assessors' remarks with the appropriate criteria is therefore not possible without in-depth interviews. The *Clarity* section of the commentary material addresses to what degree students' word choices, expressions and structure contribute to the clarity of the production. Teachers may also consider how ambiguities and errors affect the presentations and the ability to follow and absorb the content (p. 24). However, *linguistic correctness* is not explicitly mentioned as grading criteria. Instead, teachers are reminded that, in accordance with the action-oriented language view of the curriculum and the CEFR (Council of Europe, 2020; Skolverket, 2023) what the student *can* do should be central to the assessment, rather than what the student *cannot* do.

Table 9. *The Assessors' comments*

Mentions		**Categories**	**Total**
Mistakes	5		
Spelling	5		
Punctuation	2	Language	34
Content	5		
Paragraph	5	Content	5
Clarity*	4		
Flow	2		
Structure	4		
Vocabulary	3		
Linking words	2		
Understanding	2		
Score	**39**		**39**

* Refers to both *clear/unclear*
** includes *spelling* and *typos*

In contrast to the intent of the curriculum, the data from the questionnaire shows that the assessors to a large extent focus on linguistic correctness, contrary to the intent of the curriculum. The few 'positive' comments, not mentioning 'mistake concern students' ability to follow the instructions, conveying the message and text structuring.

6.2.2 Content

There are surprisingly few comments regarding the content of the student productions, given that there ought to be a dual focus on *what* the students write about and *how* they write (Skolverket, 2022b). One possible explanation for this is that the teacher respondents were asked to participate in a study on *summative assessment*, and that they omit the formative aspects of the assessments, such as commenting on content. Regardless, there seems to be a strong focus on 'form over function' from the assessors, emphasizing language correctness. A1

(see Table 11), stands out, having commented on content for every essay, whereas A3 did not comment on content at all. The only other remark about the content comes from A2 for Essay 1 (see Table 12).

Table 11. *Content; A1*

"The student has tried to answer as far as the content is concerned"
"Contentwise the student does not come to the point and quite shortly describes ´´Becoming Swedish`` quite late in the text but they have generally tried to answer what is asked."
"The content is very weak in connection to the title of the text or reflections."
"Contentwise it is very weak."

Table 12. *Content; A2*

"S/he follows instructions, and all parts are there."

6.4. Data conformity

Finally, the quantitative and qualitative data were triangulated to examine the degree of correlation, to which extent they corroborate each other, and the relationship between the data and the assigned grades. To exemplify, the triangulation process for *Essay 1*, is demonstrated below. For *Essay 1*, A2 has commented on a lack of flow and spelling that "disturb the reader's understanding of the text", and graded E, with confidence (*agree)*. The criteria for *Structure,* and *Adaptation* receives 4 points each, *Clarity*, *Cohesion*, and *Variation* – 3 points each and *Flow* – 1 point. The intended meaning of *clarity* is related to linguistic precision, indicating students' ability of using common grammatical structures and patterns and expressing themselves relatively accurately and correctly in oral and written presentations (Skolverket, 2022b). Consequently, as the assessor expresses issues with spelling disturbing the understanding of the text, the score for *Clarity* would presumably be low, indicating issues with the wording and interpretation of criteria (Bergqvist & Mårtensson, 2015; Bloxham et al., 2016).

For the same Essay, *Essay 1*, A3 has graded D, with a lower degree of confidence *(disagree)*. The criteria for *Structure, Clarity, Cohesion, Variation* and *Flow* are awarded 4 points each, and *Adaptation* – 3 points. The most important features in deciding the grade are: "structure, clarity and variation".

6.5. Data validity and reliability

The quantitative and qualitative data have been analyzed separately and together, through a process of triangulation which ensures a higher degree of validity by eliminating research bias (Cohen et al., 2018). The instrument for data gathering has been adequately tested through a pilot study, proving its stability and potential replicability. It is this author's belief that the results have high validity and reliability, although the findings cannot be said to represent the teacher community at large, due to the limited number of respondents.

7. Discussion

As mentioned, teachers are required to make a comprehensive assessment of students' knowledge when grading. Simultaneously, research points to shortcomings in teachers' grading practices in relation to the governing documents. *Construct irrelevant variance* (Lundahl et al., 2015; Lundgren et al., 2020), *low interrater reliability* (Skolinspektionen, 2010), in

combination with an increased pressure for high test results (Forsberg & Lundahl, 2006; Lundahl et al., 2015) can greatly affect the validity and reliability of assessments.

In this section, the results of the study will be discussed in relation with the previous research and the theoretical perspectives. The results will be discussed thematically, according to the research questions.

7.1. Assessment validity and reliability

7.1.1. Assessment Validity

Assessment validity has been defined as adequately capturing the essence of that which is being assessed, and include content validity, criterion validity, and construct validity (Lundgren et al., 2020). The validity of assessments is at risk if things other than the criteria are being assessed of if the construct is not being sufficiently assessed.

In this study, the assessments by the external assessors reveal differences in grading not correlating with differences in criteria. In this study, the qualitative data show a focus on 'language mistakes', despite *linguistic correctness* not forming part of the assessment criteria. Such a practice could constitute an instance of *factor irrelevant variance*, unless the 'mistakes' get in the way of understanding the text. Based on the grade variation, and discrepancies in correlation between grades and criteria, there are threats to the validity of the assessments in this study.

7.1.2. Assessment reliability

Assessment reliability refers to the replicability of results and reliable assessments should not vary between students and assessors. For assessments to be reasonably equivalent, they need, according to the National Agency for Education, to rest on as secure a foundation as possible and the grades can never be more reliable than the assessments that form the basis for them (cf. Skolverket, 2012a). One means of verifying the assessment validity is by looking at the *interrater and intrarater consistency*.

As previously mentioned, the SSI discovered low levels of interrater reliability when re-assessing the National tests (Skolinspektionen, 2010). They also found the gradings of the regular teachers to be generally higher compared to the external assessors, especially for student essays. One explanation is that although having access to detailed assessment instructions, as is the case for the National tests, teachers lack a common interpretation of the governing documents and how student performance should be assessed (Jönsson & Thornberg, 2014).

In the present study the grades of the *Teacher* are higher or significantly higher when compared to the external assessors. Only in 1/12 of the total number of re-assessments has an external assessor grading the same as the *Teacher* (see Graph 1). For comparison, the degree of assessment correlation between the external assessors is significantly higher. Thus, the results of the present study confirm the findings of the SSI with regards to *low interrater reliability* and teachers generally grading higher than external assessors.

7.2. Assessment guidelines

For grading, teachers are referred to the curriculum for the criteria of knowledge requirements, and the *Kommentarmaterial* (Skolverket, 2022b) for a clarification of the intended meaning of the criteria. The present study reveals large variations in grading, which can be illustrated by

the grading of Essay 1. Here, the *Teacher* graded B, which indicates that all the criteria for C, and most of the criteria for A are met. Simultaneously, *A1* graded F. Without further research the reasons for such a large deviation would be mere speculation, but it remains clear that the assessors have a dissensus in criteria interpretation.

The criterion-referenced grading system is vulnerable to assessment variation resulting from teachers interpreting criteria differently (Bergqvist & Mårtensson, 2015). Additionally, teachers "may not agree with, ignore or choose not to adopt the criteria" (Bloxham et al., 2016, p. 3). Vague definitions of criteria, in combination with the need for criteria interpretations jeopardize the assessment equivalence. In the present study, this is confirmed by the results. The grade variations demonstrated and the fact that the same degree of criteria fulfilment results in different grades demonstrate that understanding the knowledge requirements and the criteria is insufficient to achieving assessment equivalence.

7.3. Assessment equivalence

Consensus of criteria interpretation is central to grade equivalence. While the National Agency for Education acknowledges the need for criteria interpretation in the Swedish criterion-referenced system, the responsibility of achieving grade equivalence is delegated to the teachers. As a result, the criterion-referenced system does *not* ensure assessment or grade equivalence.

> In order for the grades to be more equal, it is central that teachers interpret the knowledge requirements together in relation to purpose, central content and the teaching that has been conducted, and thereby develop a consensus on what is required for a minimum level of knowledge in a subject and what is the minimum level for the different grade levels (Skolverket, 2020).

Research (Bergqvist & Mårtensson, 2015; Lundahl et al., 2015; Roth et al., 2016; Skolinspektionen, 2010) indicates that the main problem is that teachers lack a common interpretation of the governing documents and how student performance should be assessed. True assessment equivalence dictates that equivalent performances should be assessed equally and requires teacher consensus on the governing documents and a shared understanding of criteria. Otherwise, assessment equivalence risks being mere *surface veneer*, or having the appearance of equivalence (Black & Wiliam, 2018; Jönsson & Thornberg, 2014).

In summative assessments, the validity is reliant on teacher consensus of criteria interpretations, otherwise the grade equivalence is at risk (Skolverket, 2020). A lack of grade equivalence can be attributed to the necessity for interpretation of criteria, as shared understanding of criteria is essential for grade equivalence (Bergqvist & Mårtensson, 2015; Lundahl et al., 2015; Lok et al., 2016; Roth et al., 2016). This conclusion in confirmed by the present study.

The results of this study reveal large variations in grading, with a variation (Δ) of 2 - 4 grade steps per essay (see Table 3, Graph 1). These findings support the previous claims about the problems with grade equivalence in the Swedish criterion-referenced system (Gustafsson et al., 2014). In line with the conclusions of the SSI, the grading by the students' teacher is higher or significantly higher compared to the grading of the external assessors (Skolinspektionen, 2010). Given that the findings of the SSI exclusively regard matters of assessment equivalence in upper secondary National tests, the results of the present study offer a broader perspective on assessment equivalence, as it demonstrates issues with assessment equivalence other than those found in the National tests. This study finds no support for the Swedish criterion-referenced system ensuring assessment equivalence.

Teachers' professional judgement in assessment is a cognitive act *and* a socially situated practice (Allal, 2013). Although the National Agency for Education acknowledges the need for local interpretation of criteria in achieving grade equivalence (Skolverket, 2020), Bloxham (2016), argues that consensus of criteria is an insufficient means of reaching assessment equivalence.

In the present study the conditions of *validity, reliability* and *grade equivalence* have been tested through a set of re-assessments, based on the criteria as described in the curriculum. The results show great variations in grades and in criteria interpretation, prompting the conclusion that the requirements for *validity, reliability* and *grade equivalence* are not met. What this study has not been able to examine is how *mediation* could affect the outcome, as the assessments in this study have been conducted in isolation. Allal (2013) argues that teacher assessments are usually observed via measurement theory, and that grading is criticized for not meeting the requirements of *reliability, objectivity,* and *validity*. The assessors in the present study are affected by their local assessment culture (Allal, 2013) but have not been able to create a common *community of practice* in which mediation of meaning can be distributed. Therefore, it would be interesting to conduct as similar study, but with mediation of criteria understanding, and compare the results.

7.5 Limitations of the study

Although there have been re-assessments of the students' performance in this study, it is not defined as peer-assessments given that there has been no collaboration in the assessment process. The results, such as *mean average grade point*, are relevant for comparison in the discussion of grade criteria and equivalence but cannot be considered relevant in terms of correctness as no mediation of criteria or grade has occurred.

This study adheres to two of the three principles for grade equivalence (Gustavsson et al., 2012, p. 84): students' achievements have been de-identified and the external assessors are not assessing the performance of their own students. Also, three of the four texts were partly chosen because their original grade, B and D require even more from the teacher since no criteria are given for these grades.

The design of this study did not allow for an investigation of teachers' assessment practices from a *situated perspective*. This author would find it interesting to compare the assessment practices of two or more *communities of practice*, in an action study.

Due to the limited number of informants this study makes no claim of being representative, and the results cannot be generalizable. However, the findings of this study support the claims made by the SSI about teachers overmarking their students when compared to external assessors (Skolinspektionen, 2010). This is not to be seen as support of external assessments at the expense of teacher assessments, as this author firmly believes that only the teacher who educates and observes students over the course of a longer period, such as the entire upper secondary education, is the most qualified to provide the best suited (appropriate and relevant) *formative and summative assessments* to motivate, support and enhance student learning.

7.6 Future research

Seeing as assessors represent different assessment and grading cultures it would be relevant, for future studies, to investigate the assessments patters of teachers working in the same assessment culture, as well as comparing two or more different assessment cultures, i.e.,

teachers from two or more schools. As mediation of the interpretation of assessment criteria was not part of the present study, such a study could entail the qualitative study of the mediation process combined with the quantitative study of the assessments resulting from the mediation process in terms of validity, reliability, and assessment equivalence.

Future studies are encouraged to consider adding an interview with the respondents to elicit more information regarding the reasoning behind the grade, and the situated nature of the assessment process.

8. Conclusion

The aim of this study has been to answer the following research questions:

- How valid and reliable is the assessment of student writing in the Swedish upper secondary EFL context?
- Is the Swedish criterion-referenced grading system providing clear guidelines for assessment?
- Does the Swedish criterion-referenced grade system ensure assessment equivalence?

Based on the need for local interpretation of criteria in combination with the grade variations of this study and the re-assessments of the National tests, this study finds a lack of clear guidelines for assessment (Skolinspektionen, 2010; Skolverket, 2020). Moreover, seeing as the validity and reliability of the assessments of student writing can be questioned, as demonstrated in this study, it could be argued that the responsibility of ensuring grade equivalence is transferred from the system onto its teachers. Therefore, this study finds no evidence that the Swedish criterion-referenced grading system ensures assessment equivalence.

Some aspects of the criterion-referenced grading system require attention. As has been demonstrated, the transition from a norm-referenced to a criterion-referenced system has not had the intended effect on grade equivalence. If anything, the attributed value of a grade varies even more (Wikström, 2005). The causes for this are being attributed to criteria vagueness, discrepancies in assessment skills and external pressure to perform. Bloxham (2016) concludes that *assessment equivalence* is unlikely to be reached through criteria alone, since "shared language is insufficient to ensure shared interpretation" (p.3). Additionally, as demonstrated by her study, there is no guarantee that construct-irrelevant factors may not affect assessments or that criteria will be consistently weighed by all assessors. There is strong political motivation to increase grade equivalence, however the result of outsourcing assessment of the National tests to an external third party and defining the degree of variation between the grade of the National test and the final course grade seems connected to accountability rather than equivalence from a student learning perspective, giving it a *veneer of equivalence* (Black & Wiliam, 2018).

External assessments, as conducted in this study and in the re-assessments of the National tests, run the risk of removing responsibility of student learning, with an increased focus on accountability (Black & Wiliam, 2018). It also reduces the assessment practice from being a cognitive *and* situated practice, to solely a cognitive practice. In comparison, the markings of the original teacher in this study are significantly higher compared to those of the external assessors: results which support the conclusions of the SSI (Skolinspektionen, 2010). These findings are sometimes explained in terms of *factor irrelevant variation*, i.e., the teacher making adjustments in their assessments to compensate for their students having 'a bad day' (Lundahl

et al., 2015), or succumbing to the practical consequences of accountability – lower test scores resulting in additional work for the teacher (Black et al., 2010). However, looking at the teacher assessments through a 'situated' perspective allows acknowledging that the teacher hold an understanding of students' knowledge which reaches beyond any language mistakes of a single high stakes test (Allal, 2013; Black & Wiliam, 2018). The action-oriented view of the curriculum clearly focuses on students can do with their language, emphasizing function over form and the degree of understanding of student language production should not be reduced to language mistakes (Skolverket, 2022b). Nor does the curriculum specify a universal degree of understanding. In the Swedish education system the teacher is the primary assessor (Wikström, 2005) and with increased political demands for equivalence reduced to conformity of test scores, a higher degree of interrater reliability, or a predetermined degree of discrepancy between the results of the National tests and the final grades, there is a risk of not only loosing focus of students' learning but also a further de-professionalization, as good teachers and good education becomes even further limited to high test scores (Black & Wiliam, 2018).

From a Sociocultural perspective, learning is "less about internalization and more about appropriation in a local context", and developed through social mediation (Li, 2020, p. 32). The external assessments of this study and the re-assessments of the SSI can be regarded as representing knowledge as a cognitive act, as no mediation of meaning of criteria has occurred directly by the participants. The National tests are designed from a Sociocultural perception of learning and knowledge, with its focus on authentic language use (NAFS, 2022), yet the assessments and especially the re-assessments are evidently more cognitive-based, not acknowledging the variety of contexts or the need for mediation of understanding and knowledge.

References

Allal, L. (2013). Teachers' professional judgement in assessment: A cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 20–34. https://doi.org/10.1080/0969594X.2012.736364

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice*, *18*(1), 5–25. https://doi.org/10.1080/0969594X.2010.513678

Bergqvist, J., & Mårtensson, K. (2015). *Att formulera skriftliga betygskriterier: Bedömarkompetens och att sätta praxis på pränt*. Lunds universitets utvecklingskonferens, 2015. http://lup.lub.lu.se/record/8567366

Biggs, J. (1996). Enhancing Teaching through Constructive Alignment. *Higher Education*, *32*(3), 347–364. https://www.jstor.org/stable/3448076

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 215–232. https://doi.org/10.1080/09695941003696016

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551–575. https://doi.org/10.1080/0969594X.2018.1441807

Bloxham, S., den-Outer, B., Hudson, J., & Price, M. (2016). Let's stop the pretence of consistent marking: Exploring the multiple limitations of assessment criteria. *Assessment & Evaluation in Higher Education*, *41*(3), 466–481. https://doi.org/10.1080/02602938.2015.1024607

Borg, S. (2003). Teacher cognition in language teaching: A review of research on what language teachers think, know, believe, and do. *Language Teaching*, *36*(2), 81–109. https://doi.org/10.1017/S0261444803001903

Cohen, L., Manion, L., & Morrison, K. (2018). *Research Methods in Education* (8th ed.). Routledge. https://doi.org/10.4324/9781315456539

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment (CEFR)—Common European Framework of Reference for Languages (CEFR)—Publi.coe.int*. Common European Framework of Reference for Languages (CEFR). https://www.coe.int/en/web/common-european-framework-reference-languages

Erickson, G. (2018). Att bedöma språklig kompetens. *RIPS*.

Forsberg, E., & Lundahl, C. (2006). Kunskapsbedömningar som styrmedia. *Utbildning & Demokrati – tidskrift för didaktik och utbildningspolitk*, *15*(3), 7–29. https://doi.org/10.48059/uod.v15i3.831

Gipps, C. V., & Farajnezhad, Z. (2022). *Beyond Testing ~ Towards a Theory of Educational Assessment by Caroline V. Gipps*. https://doi.org/10.13140/RG.2.2.10759.47526

Gustafsson, J.-E., Cliffordson, C., & Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan: Problem och möjligheter* (1. uppl). SNS förlag.

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust? - Teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability*, *25*. https://doi.org/10.1007/s11092-013-9158-x

Gustavsson, P., Måhl, P., & Sundblad, B. (2012). *Betygssättning—En handbok*. Liber.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Jönsson, A. (2020). *Lärande bedömning, 5 uppl* (5 uppl). Gleerups Utbildning AB.

Jönsson, A., & Thornberg, P. (2014). Samsyn eller samstämmighet? En diskussion om sambedömning som redskap för likvärdig bedömning i skolan. *Pedagogisk forskning i Sverige*, *19*(4–5), Article 4–5. https://open.lnu.se/index.php/PFS/article/view/1401

Lau, A. M. S. (2014). 'Formative good, summative bad?' – A review of the dichotomy in assessment literature. *Journal of Further and Higher Education*, *40*(4), 509–525. http://dx.doi.org/10.1080/0309877X.2014.984600

Li, L. (2020). *Language Teacher Cognition: A Sociocultural Perspective*. Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-51134-8

Lok, B., McNaught, C., & Young, K. (2016). Criterion-referenced and norm-referenced assessments: Compatibility and complementarity. *Assessment & Evaluation in Higher Education*, *41*(3), 450–465. https://doi.org/10.1080/02602938.2015.1022136

Lundahl, C. (2014). *Bedömning för lärande* (2., [oförändr.] uppl). Studentlitteratur.

Lundahl, C., Hultén, Magnus, Klapp, Alli, & Mickwitz, Larissa. (2015). *Betygens geografi: Forskning om betyg och summativa bedömningar i Sverige och internationellt: delrapport från Skolforsk-projektet*. Vetenskapsrådet.

Lundgren, U. P., Säljö, R., & Liberg, C. (Eds.). (2020). *Lärande, skola, bildning* (Femte utgåvan, första tryckningen). Natur & Kultur.

NAFS. (2022, November 29). https://www.gu.se/nationella-prov-frammande-sprak/prov-och-bedomningsstod-i-engelska/engelska-5-gymnasiet/nationellt-prov-i-engelska-5

Roth, A.-C. V., Gunnemyr, P., Londos, M., & Lundahl, B. (2016). *Lärares förtrogenhet med betygssättning*.

Selghed, B. (2010). Ett omöjligt uppdrag– om lärares bedömningar och betygsättning. In *Dilemman i skolan—Aktuella utmaningar och professionella omställningar*. Kristianstad University Press. https://hkr.diva-portal.org/smash/get/diva2:412219/FULLTEXT01.pdf

SFS 2010:800. Retrieved March 25, 2023, from https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/skollag-2010800_sfs-2010-800.

Skolinspektionen. (2010). *Betygssättning i gymnasieskolan*. https://www.skolinspektionen.se/globalassets/02-beslut-rapporter-stat/granskningsrapporter/tkg/2010/betygsattning-i-gymnasieskolan/betygssattning-i-gymnasieskolan-2010---slutrapport.pdf

Skolverket. (2000). *Nationella kvalitetsgranskningar 2000* (No. 190).

Skolverket. (2020). *Att planera, bedöma och ge återkoppling* [Text]. https://www.skolverket.se/publikationsserier/ovrigt-material/2021/att-planera-bedoma-och-ge-aterkoppling

Skolverket. (2022a). *Betyg och prövning – kommentarer till Skolverkets allmänna råd om betyg och prövning* [Text]. https://www.skolverket.se/publikationsserier/allmanna-rad/2022/betyg-och-provning---kommentarer-till-skolverkets-allmanna-rad-om-betyg-och-provning

Skolverket. (2022b). *Kommentarmaterial till ämnesplanerna i moderna språk och engelska*.

Skolverket. (2022c). *Lgr22* [Text]. https://www.skolverket.se/publikationsserier/styrdokument/2022/laroplan-for-grundskolan-forskoleklassen-och-fritidshemmet---lgr22

Skolverket. (2023). *Ändrad ämnesplan i engelska* [Text]. https://www.skolverket.se/skolutveckling/inspiration-och-stod-i-arbetet/stod-i-arbetet/andrad-amnesplan-i-engelska

Taras, M. (2005). Assessment: Summative and Formative: Some Theoretical Reflections. *British Journal of Educational Studies*, *53*(4), 466–478. https://www.jstor.org/stable/3699279

Skollag (2010:800) Svensk författningssamling 2010:2010:800 t.o.m. SFS 2022:1319-Riksdagen, (testimony of Utbildningsdepartementet). Retrieved January 27, 2023,

from https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/skollag-2010800_sfs-2010-800.

Utbildningsdepartementet. (2023). *2022/23:60 Riksrevisionens rapport om statens insatser för likvärdig betygssättning—Skillnaden mellan betyg och resultat på nationella prov* (p. 101).

Wikström, C. (2005). Grade stability in a criterion-referenced grading system: The Swedish example. *Assessment in Education: Principles, Policy & Practice*, *12*(2), 125–144. https://doi.org/10.1080/09695940500143811

Wiliam, D. (2019). *Att följa lärande: Formativ bedömning i praktiken* (B. Önnerfält, Trans.; Andra upplagan). Studentlitteratur.

Appendix 1. Letter of Informed consent

Information om deltagande i en undersökning om (summativ) bedömning.

Du tillfrågas härmed om deltagande i denna undersökning. Syftet med undersökningen är att studera summativ bedömning av elevuppsatser på engelska och hur dessa relaterar till begreppen *likvärdig bedömning, reliabilitet* och *validitet*.

Om du väljer att delta blir du ombedd att bedöma 4 autentiska elevtexter, *Engelska* 6, samt att svara på en enkät i Microsoft forms som rör olika aspekter av bedömningen. Dina svar kommer att vara anonyma. Beräknad tidsåtgång: 1–2 timmar.

All insamlad information kommer att behandlas och presenteras i enlighet med Högskolan Dalarnas forskningsetiska regler. Deltagande personer kommer att vara anonyma och endast deltagandes data kommer publiceras. Fullständiga data kommer endast ses av mig, Fredrik Mattsson samt handledare, Jonathan White. Efter studien kommer alla data raderas.
Den färdiga studien kommer att presenteras i form av en uppsats vid Högskolan Dalarna, samt publiceras online för allmän tillgång på DiVA (arkiv för forskningspublikationer och studentuppsatser).

Uppsatsen skriver jag som mitt examensarbete vid Högskolan Dalarna för att bli behörig ämneslärare med inriktning mot gymnasiet. Ditt deltagande i undersökningen är helt frivilligt. Du kan när som helst avbryta ditt deltagande utan närmare motivering.

Jag väljer att delta i undersökningen: JA NEJ

Deltagarens underskrift: _____

Stockholm 2023-02-20
Ytterligare information lämnas av nedanstående ansvariga:

Student:
Fredrik Mattsson
v07frema@du.se
+46 704-939986

Handledare:
Jonathan White
jwh@du.se

_____ _____

Fredrik Mattsson Jonathan White

Letter of Consent

You are hereby asked to participate in a study about (summative) assessment.
The purpose is to investigate aspects of *equality, validity,* and *reliability* in relation to assessment of written student EFL writing.

Should you choose to participate you will be asked to assess (grade?) 4 authentic student essays, *Engelska 6,* and answer a questionnaire in Microsoft Forms, relating to aspects of your assessments. Your answer will be anonymous, and the estimated time is 1-2 hours.

All data collected will be presented in accordance with the Research Ethics of Högskolan Dalarna. All participants shall remain anonymous and only your data will be presented. Participant answers will only be seen by myself, Fredrik Mattsson and supervisor, Jonathan with. Once the study is completed all data will be deleted.
The finished study will be presented in a master's degree thesis and will be available online at Diva (diva-portal.org).

Your participance in the study is completely voluntary and you are free to cancel participation at any time.

I choose to participate in the study: YES NO

_____

Participant signature

Stockholm 2023-02-20
For further information please contact:

Student:
Fredrik Mattsson
v07frema@du.se
+46 704–939986

Supervisor:
Jonathan White
jwh@du.se

_____ _____

Fredrik Mattsson Jonathan White

Appendix 2. The Questionnaire[1]
Thesis study - Engelska 6, Assessment

* Obligatoriskt

1. I assess **Text 1**: *
   - F, E, D, C, B, A
2. I am confident in my assessment of the text *
   - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
3. This text meets the requirement of *struktur* *
   - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
4. The text meets the requirement of *tydlighet* *
   - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
5. The text meets the requirement of *sammanhängande text* *
   - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
6. The text meets the requirement of *variation* *
   - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
7. The text meets the requirement of *anpassning till syfte och mottagare* *
   - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
8. The text meets the requirement of *flyt* *
   - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
9. What were the most important features that decided the grade for this essay? *
Ange ditt svar:

10. I assess **Text 2**: *
    - F, E, D, C, B, A
11. I am confident in my assessment of the text *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
12. This text meets the requirement of *struktur* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
13. The text meets the requirement of *tydlighet* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
14. The text meets the requirement of *sammanhängande text* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree

[1] This version of the questionnaire has been formatted. The original digital version is available upon request.

15. The text meets the requirement of *variation* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
16. The text meets the requirement of *anpassning till syfte och mottagare* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
17. The text meets the requirement of *flyt* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
18. What were the most important features that decided the grade for this essay? *
Ange ditt svar:

19. I assess **Text 3**: *
    - F, E, D, C, B, A
20. I am confident in my assessment of the text *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
21. This text meets the requirement of *struktur* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
22. The text meets the requirement of *tydlighet* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
23. The text meets the requirement of *sammanhängande text* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
24. The text meets the requirement of *variation* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
25. The text meets the requirement of *anpassning till syfte och mottagare* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
26. The text meets the requirement of *flyt* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
27. What were the most important features that decided the grade for this essay? *
Ange ditt svar:

28. I assess **Text 4**: *
    - F, E, D, C, B, A
29. I am confident in my assessment of the text *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
30. This text meets the requirement of *struktur* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree
31. The text meets the requirement of *tydlighet* *
    - Strongly disagree, Disagree, Neutral, Agree, Strongly agree

32. The text meets the requirement of *sammanhängande text* *

- Strongly disagree, Disagree, Neutral, Agree, Strongly agree

33. The text meets the requirement of *variation* *

- Strongly disagree, Disagree, Neutral, Agree, Strongly agree

34. The text meets the requirement of *anpassning till syfte och mottagare* *

- Strongly disagree, Disagree, Neutral, Agree, Strongly agree

35. The text meets the requirement of *flyt* *

- Strongly disagree, Disagree, Neutral, Agree, Strongly agree

36. What were the most important features that decided the grade for this essay? *
Ange ditt svar:

37. Other comments:
Ange ditt svar:

Appendix 3. The students' essays

Eng. 6 Writing

English 6
Test: Writing
Theme: Autobiographical/personal writing
Time: 1 hour 30 minutes

Autobiographical essay

Write an autobiographical essay using one of the following titles:

1. Learning from my mistakes
2. "A lesson from my father" (or "…from my mother" / "…from my sister" etc)
3. Surviving School
4. Becoming Swedish
5. Only in Stockholm! (or "only in Damascus/Ankara/Lima / [var du är uppvuxen]")
6. Fitting in and standing out
7. The misunderstanding
8. A curious character
9. The student becomes the teacher
10. What my parents never understood

Your essay should include anecdotes, opinions, and reflection. Your story must be your own work.
Remember to write the title.
Write 250-700 words

1.

Fitting in and standing out

I originally come from a smal town in Greece where everyone knows eatchother and everyones lives with out having to ever talk to that person. Now I can say by experiens that the people that always have an opinion on everything that you do, where you are going and what you wear, are the old people. They are like cameras on every street that "send" the things they saw to the other cameras of that town that have chosen to take a break from their porole so that they can meet for cafe and discass everything they saw. After that infomation has left the first ones mouth it cant help but change a little in the prosses, everytime it travels from one person to an other and just like that it ends up with your parents finding out a hole different story and you getting your ass wooped.
It wasnt so much discation about the clothes that we were wearing becase we got those comments before we left our house or simplly our mothers handpicked what outfitt to wear. Eventho we hade a slight disagriment about what chlothes to wear, in the age of eleven it was always about kids chlothes.
From the moment I came to Sweden has it been a presure to fitt in, no one talks about it but everyone feels it exept for the freespirited people that just dont care about the jugment in everyones eyes whenever you do, or chlothe a surten way that they dont apruve of.
From the age of twelve had kids started to wear simular chlothes as a twenty year old and that surpriced me, because I was walking around with orange jeans, a hello kitty navy blue longsleeved shirt and a zebra grey and white jeanjacket.
After the first year I started to really care about what to wear just because I didnt want to be an outsider and the weird new girl with pug leggings. As the years went by, the pressure incrised, it incrised so much that after the age of 7 you dont get a childhood. Now I walk down the street and see a bunch of seven year olds smoking with reviling chlothes. Its crazy to think that in todays secity its cool to be a trublemaker and everyone talks about how fun it is to end up in hell.
If you dont dres like everyone else then you are stiking out and that can make some people so jelous that jugment becames boling. We have to normalize that everyone has a diftrent tast in the way they speak, act or chose to dres. You dont have to follow what the others are doing because they mighit end up in rehab or picking up trush from the grownd. If you are your self, there will be two tipes of people that will look at you. The ones that wish the hade the curige to do the same, admire you and want to be your friend, or the jelous ones that are going to make you feel bad and boly you just because you like your self so much that you dont have a problem showing how confident you are and they are selfcontios about themselvs.

2.

Becoming a Swedish

In summer 2015 a plane landed at the Arlanda airport, it was 14 July and it was so hot. Me and my family were among those passengers. Everything and everyone was new for me and I felt like an alien, I couldn't speak Swedish and I had stress because I was responssible for my little sister and my sick mother to take care of them. It was my first time traveling by plane so I didn't know so much about the rules and regulations, that is why I had stress.
My mother was sick and she had headache so she couldn't handle the situation. I was only 12 so I had to take care of my mother and my little sister. I could just few words in English but that was not enough to ask someone for help or guidness. We were waiting at the airport for my older sister to come and drive us to her house. I had a streng feeling at that moment becuase I was 5 years old when my sister came to Sweden. At that moment of stress and worries I was just waiting for my sister to see her after a long time. I was just thinking about her because I didn't remmember her so much. I was thinking about my sister's face and voice because it has been 7 years.
I was thinking and stressing, suddenly I saw someone who hugged my little sister and my mother. She was my older sister. I can't explain that feeling and I'm speechless about that because that feeling was a combination of happiness and sadness. We were happy and after check out and those stuffs we went to our house. It was Monday when we came to Sweden and we had a lots of fun those first days in Sweden. We met some great places in Stockholm, we went to shopping and bought a lots of stuffs for our new house.
My sister bought me those stuffs that I was dreaming about. She bought me a ps4, an Iphone, and a hp laptop. I was so excited for my new home and my new room but after a couple of weeks it was time to fit into the society and become a Swedish.
I had to go to the school and meet new people so I was excited for that but I was a little bit shy becuse I couldn't speak Swedish and I thought it was difficult to adjust myself and fit into the society. I stressed and worried a lot at the first day but I saw that everyone is so kind to me, some guys came to me and then some teachers and the started talking to me, they were kind and I can say that they were angles. By passing the time I could learn Swedish and I could fit into the society.
I found good friends and they helped me to learn Swedish. In my point of view it is easy to know about a society and the people of a society when you have someone from that society. My friends, my teachers and my family helped me and supported me a lot to fit into the society and become a part of this society.
I got best friends, good memmories, great times, and useful experience to become a Swedish. I learned a very valueable thing in Sweden and that was sympathi, to help people, and to respect everyone's thoughts and privecies. Sweden is a great place when you have goals in your life and when you want to achive them, it has facelities for work, study, health, and the most important a good life. That is not difficult to become a humanbieng, you just need a heart and a brain to understand everything and everyone.

3.

What my parents never understood

When I was a child, my family was forced to leave Afghanistan because of insecurity and war, we emigrated to Iran. Living in Iran also had meny problems and hardships like discrimination but I never had experienced and didn't know about it before I step into society. But when I was seven years old in the first grade, everything changed. I remember that I was the only foreign student in our class.It was the last day of school, all teachers and students was getting ready for the celebration and gathered in a big hall. I was very excited and looked araund happily, colorful ballonns and flowers were all over the hall and right in the meddle of a white table there were some presents for students. Our teacher gave everyone a present. My present was a storybook called" Cinderella". I was extremly happy and felt weightless when I got it. I was eagery looking the pictures in my new book but suddenly the teacher came back and said give back this book, I will give you another present later. I didn't want to give it back but she took the book from my hands and gave it to another girl who was from Iran because there was no present left to give her. It was such a embbarrising and sad situation because on that day everyone had presents but expect me. I was absolutely gutted, my smile disapeared of my face, my eyes were watering and everyone was looking at me. I was very sad and cry infront of everyone. It was the worest experienc that I have ever had in my childhood. I did never talked about it to my parents and they never understood it. I thing discrimination at school is the worest thing that kan tacke away childrens connfidens in future.

4.

What my parents never Understood

All of parents want the good things for their children. They think we can make best decicions for our children in every moment because they know better how live is then their children and they have more experience then their doughter or son. Yeah this is a fact at all pererts can do everything to make thier children happy and they know a allt of things that we can't understand, but in every moment this is not wright because we all people create different from each others, think different and have different whishes.
The importint point is at our maind is chanching everyday like me before i live here in Sweden. When i had lived in Afghanistan i thouth exactly like my parents like girls cant't get married without of their parents permestions and this is not okey to have a boyfriend or girlfriend, girls mast have hijab or etc , but right now i growed up in a different socaity and i don't think like my parents and i know they are right in thier own ways because they have lived in a different sociaty and culture then swedish culture This differentsy creath problems between us like they says you don't have permission to make relation with boys or you can't get marriad without of muslims boys and i remember i speak up with my parente for two mounths before about can i do to be friend with a Soni boy, they said ofcourse not. What the hells is going wrong with you, this is right at you live here i freedom country and u can do what you want but u can't across our religion and forget our agrekulture and our backgrount.