A comparative study of how the English A syllabus is interpreted in achievement tests focusing on writing ability

En jämförelsestudie av hur Engelska A kursplanen tolkas i prestationsprov som fokuserar på skriftlig förmåga

Emma Lindström
Richard Silve McLachlan

Handledare: Sarah Jane Gould
Examinator: Björn Sundmark
Abstract

This thesis sets out to do a comparative study of how the English A syllabus is interpreted in achievement tests focusing on writing ability. The tests were collected from vocational and theoretical programs at upper secondary schools. The measurement instrument employed in this study was a content validity scale produced with material originating from Brown (1983), Brown and Hudson (2002) and EN1201 - English A (2000) and was utilized by two judges in order to analyse the collected tests. The analysis displayed numerous noteworthy index mean differences between the two groups: vocational and theoretical. The theoretical programs’ tests received overall a higher index mean compared to the vocational programs’ tests. However, a variance existed between the two judges’ scorings on most of the questions in the validity scale, and some questions’ means contradicted the index means, i.e. one of the judges ranked the vocational programs’ tests higher on more than one occasion. The analysis also revealed two diverse testing approaches, where one had a more prescriptive nature and was closely connected to the vocational programs while the other one, having a close connection to the theoretical programs, was more descriptive in nature. The degree of content validity in each test group (vocational and theoretical), showed that according to the scale used in the analysis of this study, the theoretical programs’ tests display a higher degree of content validity (higher means) compared to the vocational programs’. This result implies a difference in interpretation of the syllabus depending on program (theoretical and vocational) when constructing achievement tests. The difference found could be due to the vagueness of the syllabus.

Keywords: English A, syllabus, achievement tests, writing ability, vocational program, theoretical program,

Svenskt abstract

åt mellan de två grupperna: yrkesförberedande och teoretiska. Proven tillhörande teoretiska program fick generellt ett högre indexmedelvärde än proven tillhörande yrkesförberedande program.Dock förekom det en skillnad mellan de två bedömarnas rankningar på de flesta frågorna i validitetsskalan, och några frågors medelvärden opponerade emot indexmedelvärdena, d.v.s. en av bedömarna rankade de yrkesförberedande programmens prov högre vid mer än ett tillfälle. Genom analysen upptäcktes två olika tillvägagångssätt av testning, där ett var av en mer preskriptiv natur och hade en anknytning till de yrkesförberedande programmen medan det andra tillvägagångssättet som var knyttet till de teoretiska programmen var av en mer deskriptiv natur.

Graden av innehållsvaliditet för respektive grupp (yrkesförberedande och teoretisk) upphäver, enligt validitetsskalan i studiens analys, att proven tillhörande teoretiska program demonstrerar en högre grad av innehållsvaliditet (högre medelvärden) än de yrkesförberedande. Detta resultat tyder på en tolkningsskillnad av kursplanen mellan de två programmen vid konstruerandet av prestationsprov. Skillnaden mellan testen som upptäcktes kan bero på den vaga kursplanen.

Nyckelord: Engelska A, kursplan, prestationsprov, skriftlig förmåga, yrkesförberedande program, teoretiskt program,
Table of content

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introduction</td>
<td>7</td>
</tr>
<tr>
<td>Purpose</td>
<td>9</td>
</tr>
<tr>
<td>Background</td>
<td>9</td>
</tr>
<tr>
<td>The syllabus for English A and the goals for writing</td>
<td>9</td>
</tr>
<tr>
<td>Why do teachers test and what are the benefits of testing?</td>
<td>12</td>
</tr>
<tr>
<td>Achievement testing and proficiency testing</td>
<td>13</td>
</tr>
<tr>
<td>Concerning the constructing of achievement tests</td>
<td>15</td>
</tr>
<tr>
<td>The validity of tests</td>
<td>16</td>
</tr>
<tr>
<td>Validation of Achievement Test</td>
<td>16</td>
</tr>
<tr>
<td>Content validation</td>
<td>16</td>
</tr>
<tr>
<td>Content validity scale</td>
<td>18</td>
</tr>
<tr>
<td>Defining language ability</td>
<td>18</td>
</tr>
<tr>
<td>Types of writing performance</td>
<td>19</td>
</tr>
<tr>
<td>Micro skills for writing</td>
<td>20</td>
</tr>
<tr>
<td>Macro skills for writing</td>
<td>20</td>
</tr>
<tr>
<td>Method</td>
<td>20</td>
</tr>
<tr>
<td>Material</td>
<td>21</td>
</tr>
<tr>
<td>Teachers’ tests</td>
<td>21</td>
</tr>
<tr>
<td>The content validity scale</td>
<td>21</td>
</tr>
<tr>
<td>Procedure</td>
<td>23</td>
</tr>
<tr>
<td>Results</td>
<td>24</td>
</tr>
<tr>
<td>Question one: Is there an overall match between the test and the syllabus regarding writing ability?</td>
<td>27</td>
</tr>
<tr>
<td>Question two: Is there an overall match between the test and the objective that it is meant to test?</td>
<td>28</td>
</tr>
<tr>
<td>Question three: Does the test formulate the questions to allow the students to use their ability to express themselves in writing in different contexts?</td>
<td>28</td>
</tr>
<tr>
<td>Question four: Does the test’s questions allow the students to be “able to formulate themselves in writing in order to inform”?</td>
<td>29</td>
</tr>
<tr>
<td>Question five: Does the test’s questions allow the students to be “able to formulate themselves in writing in order to instruct”?</td>
<td>30</td>
</tr>
<tr>
<td>Question six: Does the test’s questions allow the students to be “able to formulate themselves in writing in order to argue and express feelings and values”?</td>
<td>31</td>
</tr>
<tr>
<td>The two testing approaches found in the analysis</td>
<td>32</td>
</tr>
<tr>
<td>The results of the research questions</td>
<td>33</td>
</tr>
<tr>
<td>Discussion</td>
<td>34</td>
</tr>
<tr>
<td>Interpretation of the syllabus</td>
<td>35</td>
</tr>
<tr>
<td>The difference</td>
<td>35</td>
</tr>
<tr>
<td>Source of errors</td>
<td>38</td>
</tr>
<tr>
<td>Future suggestions and further research</td>
<td>39</td>
</tr>
<tr>
<td>References</td>
<td>41</td>
</tr>
</tbody>
</table>
Introduction

All students are familiar with teacher-made classroom achievement tests. These tests have distinguishing features: they are constructed by the course instructor, cover material taught in a particular course and are used to determine grades. All classroom teachers are constantly confronted with the problem of assessing their students’ knowledge and skills, which often involves achievement testing (Brown, 1983). But, how do the teachers know if the tests represent all of the important aspects that were intended in the syllabus, in other words: do the tests have content validity in reference to the syllabus? The answer according to Brown (1983) is that they most often do not, since it is very seldom that teachers validate their own tests. The reason for this is that it is difficult to validate a test without prior knowledge within the field of validation. Another factor is that it is seen as a waste of time to validate a test that most likely will never be used again, and the fact that the procedure of validation is time consuming.

This thesis sets out to do a comparison of how the syllabus for the ‘English A’ course is interpreted in teacher-made classroom achievement tests, belonging to the category of criterion-referenced tests, focusing on students’ writing ability, distributed in theoretical programs and vocational programs. Another aspect of this study will be to analyse to which degree the collected tests demonstrate content validity. The English A course was chosen as it is mandatory for all students attending Swedish upper secondary school. The nature of the comparison will be to investigate if any divergences exist concerning the way the national syllabus’s goals for writing in English A have been interpreted in the tests. This study will compare tests used in the theoretical programs to tests distributed in vocational programs. According to the national curriculum for the course of English level A, all students at upper secondary school are to reach the same goals and no difference should exist between the educational programs. The research question was decided upon the basis of a hypothesis that tests on vocational programs tend to differ from the tests distributed in theoretical programs and this might be a result of the different interpretation of the syllabus teachers might carry out in a vocational class compared to a theoretical class. This hypothesis was formed when the authors of this thesis read Lindberg (2002), who in her interview study found that the interviewed teachers expressed a concern that every student taking a course in the core subjects should be assessed upon the same basis, which
is outlined in the syllabus. The teachers worried that students attending a theoretical program would not be assessed in the same manner as students in a vocational program. Another study that forms the basis for this thesis was Jansson and Carlsson (2005) who found that teachers’ interpretation of the grading criteria depends on the program that the student attends. If as Jansson and Carlsson (2005) suggest the teachers interpret the grading criteria differently depending on the program there is a possibility that they might interpret the whole syllabus, or parts of it, differently. Another dissertation that influenced the hypothesis of this study was Björkander (2006), who found that tests on vocational programs mainly test knowledge of facts.

The two goals of this thesis can be phrased as research questions:

1. Do any differences exist in the interpretation of the syllabus between the tests used in vocational programs compared to tests used in theoretical programs?
2. To which degree do the collected tests demonstrate content validity?

To answer the research questions above, six questions were produced to be used in the content analysis. The questions derived from Brown and Hudson (2002) and the goals for writing in the syllabus for English A (EN1201 - English A, 2000). The tests were collected from four different programs, namely two theoretical and two vocational.

The authors of this thesis consider this subject highly relevant to their coming career since testing is a primary tool of education. While examining the content of the tests the authors of this thesis will not only gain a better and deeper understanding of the nature of language achievement tests, but will improve their general knowledge concerning the testing phenomena in schools today. This area of investigation is highly relevant as the parliament of Sweden suggest that changes affecting the vocational program should be made. A motion was proposed in June 2005 to adapt the goals for the following theoretical subjects: Swedish, English, Mathematics and Civics, to fit the vocational profile better (Leijonborg, 2005; Tolgfors, 2005).

It is however outside the range of this thesis to join the debate concerning the necessity of tests in educational settings as well as the question of whether tests in general are fair and effective. This thesis will be limited to carrying out a comparison between tests and to investigating whether the
English syllabus goals are interpreted differently in this particular form of assessment in the different programs.

**Purpose**

The purpose of this study is to look at the content of English A achievement tests focusing on writing ability employed in theoretical programs and vocational programs at Upper Secondary School and compare them to the goals for writing in the national syllabus for English A. The comparison will be carried out using a scale of content validity constructed with material from Brown and Hudson (2002) and EN1201 - English A (2000). To phrase this as a research question: Are the English A syllabus goals for writing interpreted differently in achievement tests in the different programs?

**Background**

*The syllabus for English A and the goals for writing*

A specific syllabus is constructed for each of the core subjects/courses taught in the Swedish school by the national agency for education. The syllabuses are more subject specific than the curriculum and define goals in the subject. Moreover the syllabuses describe what knowledge or ability level the students should have attained at the end of the course. The syllabuses also contain the grading criteria (Skolverket, 2007). The system in place provides the teachers with a large amount of freedom in interpretation of the syllabuses and as a result there is a risk that the course is taught in very diverse ways depending on the school or program. Many teachers consider the formulations of the goals and grading criteria in the syllabuses to be indistinct and inconsistent (Andersson, 1999).

The syllabus for English A states that the following goals should have been reached at the end of the course: “Pupils should be able to formulate themselves in writing in order to inform, instruct, argue and express feelings and values, as well as have the ability to work through and improve their own written production” (EN1201 - English A, 2000). These goals should have been reached by every student attending this course and therefore they form the basis for this study.
Also included in the national syllabuses for English A are the criteria for the three grade levels:

Criteria for Pass (G):
“Pupils write in clear language, not only personal messages, stories and reflections, but also summaries dealing with their own interests and study orientation” (EN1201 - English A, 2000).

Criteria for Pass with distinction (VG):
“Pupils write letters, notes and summaries of material they have obtained in a clear and informative way that is appropriate for different purposes and audiences” (EN1201 - English A, 2000).

Criteria for Pass with special distinction (MVG):
“Pupils write with coherence and variation, use words and structures with confidence, as well as communicate in writing in ways appropriate to different audiences” (EN1201 - English A, 2000).

In addition to the goals for writing in English A mentioned above, the syllabus states that “English should not be divided up into different parts to be learnt in a specific sequence. Both younger and older pupils relate, describe, discuss and reason, even though this takes place in different ways at different language levels and within different subject areas” (EN1201 - English A, 2000).

These goals and guidelines refer to a general knowledge of writing in English. A more detailed interpretation is then made by the local school. Following the more detailed local goals, the teacher of the course constructs lesson plans, and tests usually follow as a result of these. This thesis will however not use the local course plans, since they are specific to the school where they were created. Since the local plans are based upon the national course/subject syllabus, this becomes a better basis for comparison as the collected tests originate from two separate schools.

As mentioned before, Jansson and Carlsson (2005) found that teachers fail to achieve an equal interpretation of the grading criteria on a specific grading level: as the interpretation depends on whether the student attends a theoretical or vocational program. In their study, Jansson and
Carlsson also discovered that teachers assume students attending the vocational programs to have lower potential to meet the goals of the syllabus. Jansson and Carlsson came to the conclusion that the teachers had different pre-requisites concerning the students depending on which program they attended. Because of this there is a risk that the teachers then indirectly carry out a different interpretation of the grading criteria depending on the program.

Björkander (2006) found in her study on knowledge tests that there is a clear focus of testing knowledge of facts in the vocational programs. Aspects such as reflection and ethical dimensions constitute only a small fraction of the collected tests. Björkander also set out to investigate which cognitive approach is the most prevalent in the collected tests. The results pointed towards a focus upon behaviouristic approach.

Lindberg (2002) carried out an interview study concerning the consequences the new grading system (Lpf 94) for teachers and students. In Lindberg’s study teachers of the core subjects expressed themselves in sceptical terms concerning the syllabuses and the grading criteria. The introduction of collective syllabuses meant that all students regardless of program would have to reach the same goals and be assessed on the same basis. The teachers considered the collective syllabuses and grading criteria an injustice towards the students. The injustice according to the teachers was mainly based upon the notion that the collective syllabuses made the core subjects more difficult for students in vocational programs and ‘the goals to accomplish’ aspect of the syllabuses placed the vocational students in a situation where they did not have the pre-requisites to accomplish the course. Moreover, teachers of the theoretical programs worried that the students attending a theoretical program would not be assessed upon the same basis as the students attending a vocational program (Lindberg, 2002, p. 47). Teachers of the core subjects experienced a decline in the demands for the grade “Pass” (G) since teachers were more or less forced to hand out simplified tasks to students in vocational programs in order for them to receive a “Pass”. In addition, Lindberg (2002) found that teachers considered the syllabuses unclear and vague, which could lead to the risk that the students are assessed differently between teachers, between schools and between programs.
Why do teachers test and what are the benefits of testing?

Even though this thesis does not set out to debate whether tests are a good thing or not, it is essential to mention some of the reasons for testing, since tests are an important tool of education. Tests are constantly criticized and their use called into question as many find them unnecessary and an unsatisfactory form of assessment. Despite the criticism of tests, they are used all over the world on a daily basis. According to Freccero, Hortlund, and Pousette (2005) tests are mainly distributed for the purpose of

- observing the level of knowledge the students are at in order to adapt the tuition and teaching methods accordingly.
- giving the teacher a response concerning whether what she or he has intended to teach to ascertain whether this has truly been understood by the students
- serving as the basis for assessment, grading and evaluation of the effectiveness of the teaching methods
- assisting the students in their evaluation of their own way of learning

Concerning other possible benefits of testing Marton, Dahlgren, Svensson and Säljö (1977) suggests that tests and the formal grade they result in constitute the most important form of influence on people’s choice of learning in most educational settings.

Why are most tests quite alike in their construction and what do the tests measure?

According to Måhl (1991) the dominant test type in schools generally is prescriptive in its nature and mostly contains questions asking for detailed knowledge of facts. Måhl mentions that some 80 – 90 % of all questions on a test are fact-based questions.

The reason why most tests are written examinations according to Måhl (1991) is that written tests have a more fundamental effect on the overall grade than observation or oral tests. Måhl suggests that most teachers find it easier to construct a written test based upon the educational guidelines than an oral test based on the same guidelines. Therefore this study will only examine written tests, as they are the most common type of test and the easiest to comprehend and analyse.
Another factor in the construction of tests might be that quite often language teachers are blamed when the public and various types of media turn their attention towards the deteriorating standards of students’ mastery of the language, be it actual or perceived standards. This critique tends to focus on the students’ inadequate knowledge of spelling, grammar and vocabulary. This kind of critique carries an important political message, which has a profound effect upon language teaching. Even though the grading system of today provides the teachers with a higher level of flexibility concerning the content and focus of their lessons, most teachers are highly influenced by the views of the public and the will of the politicians concerning what composes a suitable content of lessons. Therefore most tests distributed by language teachers take a prescriptive approach and focus on factors such as spelling, grammar and vocabulary (Allison, 1999).

Achievement testing and proficiency testing

There are two main approaches to testing, achievement testing and proficiency testing. Achievement tests assess what an individual has learnt or achieved during a course of instruction. Proficiency tests measure an individual’s language ability irrespective of how this ability has developed (Hughes, 1989). The former was chosen for this study since teachers more often base their grades on the performance on an achievement test.

To narrow down the perspective of this study, teacher-made classroom achievement tests focusing on writing ability and belonging to the genre of criteria-referenced tests were chosen as the material. This thesis will use Popham’s (2000), McNamara’s (2000) and Brown’s (1983) definitions of achievement test: According to Popham (2000) achievement tests are constructed for the purpose of assessing the knowledge of the student and what she/he can do in a specific subject as a result of schooling. An achievement test is a tool that measures aspects of a student’s developed abilities and process of learning. As a result, achievement tests are often at the end of a project, course or semester (Popham, 2000, p. 45).

Brown’s definition of achievement tests is: “The distinguishing features of these tests were that they were constructed by the course instructor, covered only material taught in a particular
course, probably never have been (or ever will be) administered in exactly the same form to any other class, and were used to determine grades“ (Brown, 1983, p. 232).

According to McNamara (2000) and Brown (1983) achievement tests are closely connected to the procedure of teaching and should therefore strongly support the form of instruction to which they are related. This close bond between teaching and achievement tests is also the reason why they are frequently distributed at the end of a project or a course. The purpose of achievement tests is to investigate if and where progress has been made in relation to the official goals of learning. Concerning the content of an achievement test McNamara observes that “An achievement test may be self-enclosed in the sense that it may not bear any direct relationship to language use in the world outside the classroom”. This kind of test can be used to assess vocabulary used in a specific domain or certain areas of grammar. However, McNamara points out that an achievement test can not be “self-enclosed” if “the syllabus is itself concerned with the outside world, as the test will then automatically reflect that reality in the process of reflecting the syllabus” (McNamara, 2000, p. 156).

The tests this study will cover are achievement tests that belong in the genre of criterion-based tests, also known as content-referenced tests (Brown, 1983). Criterion-based tests compare a student’s test result to clearly predetermined defined skill levels or standards (Popham, 2000, p. 141). The important point is that scores are interpreted in terms of content mastery and to determine an individual’s level of that mastery (Brown, 1983). The opposite of criterion-based tests are norm-referenced tests, which are constructed to illustrate a student’s test result in contrast to the other student’s performance (e.g. she spells better than 75 % of her classmates). Most norm-referenced tests are multiple choice tests (Popham, 2000, p. 141). Since the latter genre of tests is rarely found in the Swedish educational system it seemed irrelevant to include them in this study.

**Concerning the constructing of achievement tests**

Teachers frequently face the challenge of constructing an achievement test that is relevant, carries good validity and reliability and is connected to both the instructions and the relevant syllabus. This poses a problem since relevant and good material covering achievement test construction is
severely lacking. Osterlind (1998) writes that “Test item writers are routinely left to their own devices because there is no developed theory to underpin item writing, nor is there a comprehensive resource identifying the distinctive features and limitations of test items, the function of test items in measurement, or even the basic editorial principles and stylistic guidelines.” (Osterlind, 1998, p. 6). Research in this area has been lacking for some time. Thirty years ago Cronbach (1970) observed that “the design and construction of achievement test items have been given almost no scholarly attention” (p. 509). Furthermore, Bormuth (1970) commented on the lack of concern for information and instruction about constructing test items, and observed that most teachers rely upon nothing more than their intuitive skills when constructing a test.

However there are a few sources of information relating to the construction of achievement tests. According to Larsson (1986) a good model to take into consideration when constructing tests is Blooms taxonomy, which can be studied in its complete form in Bloom, Hastings and Madaus (1981). Bloom constructed a systematic categorization of knowledge levels divided into six categories:

1. **Knowledge.** The ability to remember facts, definitions, methods and principles.

2. **Understanding.** The ability to interpret, understand consequences, translate ideas from one form to another, identify developing trends.

3. **Application.** The ability to use general laws and principles in certain situations.

4. **Analysis.** The ability to identify the components of a structure and present their nature and their internal connection.

5. **Synthesis.** The ability to collect and categorize elements so a new conclusion can be drawn as a result.

6. **Evaluation.** The ability to establish the value of certain material or method through comparison.

Bloom et. al (1981) states that the categorization is hierarchic in its nature so that the latter abilities are built upon the earlier ones and a student has to go through all the levels to reach a complete picture concerning the area of study. Today most scientists and researchers believe that
the levels outlined in Bloom’s taxonomy are not hierarchic, but are stand-alone constructs although they are considered closely interrelated (Korp, 2003, p. 87).

**The validity of tests**

The validity of a test refers to whether the test is actually measuring what the tester set out to measure. According to the most recent edition of the *Standards for Educational and Psychological Testing* (AERA, APA, NCME, 1999), “Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of test scores” (p. 9). The test item ‘Give a short summary of Shakespeare’s life’ might carry low validity if the test is a language test since the restrictions of a student’s ability to convey a correct answer might lie in the student’s lack of knowledge concerning Shakespeare’s life and not the actual language ability required to answer the test item correctly.

**Validation of Achievement Test**

The most appropriate method of validation when validating criterion-referenced achievement tests is content validation, as achievement tests are designed to measure a specific content and/or skills domain and the test items are selected to represent that domain (Brown, 1983).

**Content validation**

According to Brown (2004), there are some critical questions teachers need to ask themselves before constructing a test to raise the content validity of the test. Originally there are five questions, however since this thesis focuses on content validity we will use the four questions relating to the principle of validity. These are as follows:

1. “**What is the purpose of the test?**” (p. 42). According to Brown the teacher needs to reflect upon the reason for constructing the test. Is the reason for constructing “for an evaluation or overall proficiency?” (p. 42). Is the reason for creating the test to accept students into a course or to assess achievement? When the main reason to constructing the test has been decided the teacher should establish the test objectives.

2. “**What are the objectives of the test?**” (p. 42). Brown writes that the teacher should outline what he or she is “trying to find out” (p. 42) with the test and what language abilities are to be tested. According to Brown there exist a number of obstacles to creating relevant
objectives, such as with function and form but also more sophisticated ones where “constructs to be operationalized in the test” (p. 42) constitute one example.

3. “How will the test specifications reflect both the purpose and the objectives?” (p. 42).

With this point Brown comments on the importance of having a solid structure where the objectives of the test are well integrated and where the different abilities to be measured are properly balanced.

4. “How will the test tasks be selected and the separate items arranged?” (p. 42). Here the teacher should strive for content validity “by presenting tasks that mirror those of the course (or segment thereof) being assessed.” (Brown, 2004, p. 42).

Concerning content validity, Brown (2004) writes that “If a test actually samples the subject matter about which conclusions are to be drawn, and if it requires the test-taker to perform the behaviour that is being measured, it can claim content-related evidence of validity, often popularly referred to as content validity” (p. 22). Brown continues “if a course has perhaps ten objectives but only two are covered in the test, then content validity suffers” (p. 23)

According to Brown (2004) two different ways of testing exist that should be taken into consideration when establishing the content validity of a test, namely direct and indirect testing. Direct testing requires the test-takers to carry out the target task; however, as the name implies indirect testing involves the test-takers in a task that is associated with the target task but not the task itself.

According to Brown (2004) there are two steps to estimating the content validity of a classroom test

1. “Are the classroom objectives identified and appropriately framed?” (p. 32). With this question Brown means to seek information on whether the objectives of the tests are identifiable.

2. “Are lesson objectives represented in the form of test specifications?” (p. 32). Brown relates to the importance of a clear test structure that should be closely related to the unit being assessed. Brown (2004) gives the example of dividing the test into different parts,
constructing an array of diverse test items and striving for balance between each part of the test.

Content validity scale

This thesis’s method of content validation of the tests is a content validity scale taken from Brown (1983), Brown and Hudson (2002) and adapted to the syllabus for English A (EN1201 - English A, 2000). This method was chosen since it focuses on aspects the writers of this study wish to analyse, namely to examine if there exists a difference in the interpretation of the syllabus in the collected achievement tests. A description of the method will follow.

The method is carried out in the following manner: "Scales could be developed for the overall validity of the test and/or for certain dimensions, such as content coverage, stress, unimportant points and appropriateness of item format to content. The ratings of various judges could then be compared statistically with agreement between the various judges’ ratings being an estimate of interrater reliability. If the level of interrater reliability were sufficiently high, the mean rating could be used as an index of content validity” (Brown, 1983, p. 137). That sort of scale has been developed and used for validating the tests in this study and Brown (1983) also explains who those judges might be: “Who is an expert judge depends on the type of test and its intended uses. For example, expert judges of the content validity of an achievement test would be people who teach the courses in the subject area the test covers” (p. 135). According to Brown (1983) those who can judge about a test’s content validity are people who teach that course in the subject area the test covers, here English, as the writers of this study do.

Defining language ability

The aim of the collected tests is to assess the student’s language ability with a focus on their writing ability, this term will be frequently related to in this thesis. As the term ‘ability’ is a very general term it is necessary to clarify what the authors of this thesis imply by the term ‘language ability’. The syllabus for English A outlines one main point of language ability which is the students’ “ability to communicate and interact in English in a variety of contexts concerning different issues and in different situations” (EN1201 - English A, 2000). Bachman (1990) suggests that one should define ‘language ability’ in relation to the cognitive demand of the task and therefore also connect the term to the perceived range of levels that ‘language ability’ entails.
Consequently a student with a high level of language ability should therefore have a higher probability of answering correctly on a test item referring to that language ability than a student with a low level of language ability (p. 4). Given this definition, writing ability will be treated as the skill or level of knowledge required to produce a written text where the students can explain their thoughts. Strong writing ability would then imply that the students will be able to produce a high-quality written text where they succeed in getting their points across.

Types of writing performance
According to Brown (2004) there are four aspects of written performance.

1. **Imitative.** At this stage Brown (2004) places the mechanical dimension of writing, where form is of the most importance thus making content and meaning secondary. The ability to spell correctly and to recognize “phoneme-grapheme correspondences in the English spelling system” is focused upon. This is usually practised by learning to write letters, and then moving on to forming words and using punctuation.

2. **Intensive (controlled).** Form is still the main focus at this stage but more attention has been given to context and meaning while assessing. Abilities such as “producing appropriate vocabulary within a context, collocations and idioms, and correct grammatical features up to the length of a sentence” (p. 220) belong in this category.

3. **Responsive.** In this aspect context and meaning are stressed while form is considered at the discourse level. The types of writing that this aspect includes are “brief narratives and descriptions, short reports, lab reports, summaries and brief responses to reading” (p. 220). The writers can communicate their ideas in several ways and are more concerned with how to express their thoughts in the most suitable manner than spelling correctly.

4. **Extensive.** When the writers have reached this stage they have mastered all the previous stages and are able to adapt a number of approaches towards writing. To be able to view writing as a process is often the trademark of this stage. The type of writing can be “a term paper, a major research project report, or even a thesis. Writers focus on achieving a purpose, organizing and developing ideas logically, using details to support or illustrate ideas, demonstrating syntactic and lexical variety” (p. 220).
Micro skills for writing
According to Brown (2004) writing ability consists of two aspects, micro skills and macro skills. Micro skills are concerned with the correct form of how to represent speech, thoughts or opinions in a graphical form. The focus is on correct grammar, and the ability to “produce an acceptable core of words and use appropriate word order patterns” and “appropriately accomplish the communicative functions of written texts according to form and purpose“ (p. 221).

Macro skills for writing
The macro skills for writing are only available for those who have mastered the micro skills. The macro skills focus on the message the writer intends for the reader and strategies to convey this message in a clear disposition. These macro skills are important if a student aims to master extensive writing in the target language. An example of a macro skill of writing is when students “convey links and connections between events and communicate such relations as main idea, supporting idea, new information, given information, generalization, and exemplification” as well as when they “develop and use a battery of writing strategies, such as accurately assessing the audience's interpretation, using pre-writing devices, writing with fluency in the first drafts, using paraphrases and synonyms, soliciting peer and instructor feedback, and using feedback for revising and editing” (Brown, 2004, p. 221).

Method

As mentioned in the background section, the most appropriate method for this thesis’s content validation of the tests is a content validity scale taken from Brown (1983), Brown and Hudson (2002) and adapted to the syllabus for English A (EN1201 - English A, 2000). This method can be used to analyse whether there is a difference in the interpretation of the syllabus in the collected achievement tests.

The research method in this study is a comparative study since the aim is to perform a content validation of a small sample of tests (16 tests) from the respective vocational and theoretical programs. The tests will be analysed using a scale of content validity to explore if there are any differences exist between the tests used in the respective vocational and theoretical programs.
concerning the interpretation of the national syllabus goals for writing in English A. It will also be used to investigate to which degree the collected tests demonstrate content validity.

The upper secondary schools providing the material for this study agreed to do so after initial contact was made by the authors. Mainly the initial contact was made using e-mail, where a short text explained the needs of this study and posed the question if the school could consider providing relevant tests. The schools that responded and were willing to supply tests were then contacted by the authors. The schools are located in the south of Sweden.

**Material**

The raw data of this study are English A teacher-made classroom achievement tests focusing on writing ability. These tests are criterion-based tests. What the authors of this thesis define as an achievement test is explained in the background section. All the tests are taken from level A courses since that level is compulsory for all programs and will therefore be found in both vocational and theoretical programs. Altogether there were 16 tests, eight taken from vocational programs and the other eight belonging to the theoretical programs. The age of the tests varies from 2000 to 2007 with a mean of 2005.

**Teachers’ tests**

The 16 tests that were constructed by teachers themselves are collected from four different teachers from two schools and from respective theoretical and vocational programs, eight test from each program. The tests are restricted to only include English level A, to be tests of writing ability and created by teachers themselves (not The National Tests of English). The contacted teachers received a detailed explanation of what kind of tests this study required. 18 tests were then collected. However, according to the judges, two tests were not seen as tests focusing on writing ability since other language abilities were included as well, such as grammar and vocabulary. And therefore those two tests were removed from this study’s sample.

**The content validity scale**

The scale that will be used to create an ”index of content validity” is adapted from Brown (1983), Brown and Hudson (2002) and EN1201 - English A (2000).
Very poor – Moderate – Very Good
1 – 2 – 3 – 4 – 5

The scale 1 to 5 is interpreted as 1 being the lowest grade that could be received if the judges’ experienced that the test poorly reflected the syllabus of English A. A higher number was given to reflect the degree to which the test reflected the syllabus of English A. It is however a subjective judgemental procedure, as all qualitative methods are, and therefore the two judges’ separate means will be taken into consideration in the results and discussion section, in case of any major divergences between the judges.

1. Is there an overall match between the test and the syllabus regarding writing ability? (Brown & Hudson, 2002; EN1201 - English A, 2000).

2. Is there an overall match between the test and the objective that it is meant to test? (Brown & Hudson, 2002, p. 100). The objective in this case is whatever the test is trying to set out to test. If the test is a book test, the test and its questions should relate to that and not include general grammar or translation for example.

3. Does the test formulate the questions to allow the students to use their ability to express themselves in writing in different contexts? (EN1201 - English A, 2000).

4. Does the test’s questions allow the students to be “able to formulate themselves in writing in order to inform”? (EN1201 - English A, 2000).

5. Does the test’s questions allow the students to be “able to formulate themselves in writing in order to instruct”? (EN1201 - English A, 2000).

6. Does the test’s questions allow the students to be “able to formulate themselves in writing in order to argue and express feelings and values”? (EN1201 - English A, 2000).
Procedure

The first step was to contact teachers at the two schools to acquire the tests required for this study. This was done by phoning and e-mailing teachers at target schools. They were fully informed of the aim of this study and their own anonymity. After agreement from the contacted teachers, the tests were either e-mailed, sent by mail or collected from the school.

The analysis consisted of two judges, the writers of this thesis, judging the 16 tests according to the content validity scale. As mentioned before, this study will use the method Brown (1983, p. 137) presented and is carried out in the following manner: the ratings of various judges were compared statistically to construct an index mean that was to be used for comparison between the tests.

However, before the analysis could begin, a third person was involved. This is called double-blind procedure and means that neither of the judges knew before and during the analysis which test belonged to the vocational or theoretical program, which will henceforth be known VPs and TPs respectively. The third person was e-mailed the tests or given the tests before any of the judges saw them. The third person then removed any kind of information regarding which program the tests belonged to, before handing them over to the judges for the analysis. A double-blind procedure is used to guard against both experimenter bias and placebo effects (Christensen, 2004).

The analysis was conducted during one day at one of the judge’s homes, in separate rooms. The tests were randomly divided into two piles of eight tests per pile. When the judges had finished their analysis the tests were swapped, so both judges individually analysed all 16 tests. When all the tests had been analysed and graded, the index mean according to the content validity scale was calculated. The tests’s graded numbers (1 to 5) were compiled and the means for every question in the validity scale were calculated, two means for each question (one for each judge). From those means the index mean was constructed. Up to this point none of the judges knew which test belonged to which program because of the double-blind procedure, though after the analysis and the results were done, program information (VP and TP) about the tests was revealed by the third person involved in the study.
Results

As stated in the method section the index mean for each question and for each program (vocational and theoretical) were produced from both judges’ individual mean. An overview of the final mean for each question is shown in Table 1. Each question in the content validity scale will be accounted for below and the research questions will be accounted for at the end of the results section but they are worth repeating before presenting the results:

1. Do any differences exist in the interpretation of the syllabus between the tests used in vocational programs compared to tests used in theoretical programs?
2. To which degree do the collected tests demonstrate content validity?

The mean computed from each question is based on the five-grade scale:

- Very poor – Poor – Moderate – Good – Very Good
- 1 – 2 – 3 – 4 – 5

Table 1. Overview of the index means for each questions in the validity scale

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theoretical</td>
<td>IM 3.63, N 8</td>
<td>IM 3.88, N 8</td>
<td>IM 3.38, N 8</td>
<td>IM 4.00, N 8</td>
<td>IM 1.56, N 8</td>
<td>IM 3.25, N 8</td>
</tr>
<tr>
<td>Vocational</td>
<td>IM 2.75, N 8</td>
<td>IM 3.19, N 8</td>
<td>IM 2.81, N 8</td>
<td>IM 2.62, N 8</td>
<td>IM 1.31, N 8</td>
<td>IM 1.94, N 8</td>
</tr>
</tbody>
</table>

IM: Index Mean, J1M: Judge One’s Mean, J2M: Judge Two’s Mean, N: Number of tests, JV: The variance between the judges

Table 1 clearly displays that there is a general difference between the index means for theoretical and vocational programs’ graded tests. In all six questions the TPs’ tests received a higher mean, which in turn represents a higher degree of content validity according to the graded scale. Graph 1 below displays the results of Table 1 in a graphical form.
Graph 1. An index of the means according to the content validity scale.

As graph 1 shows, there is a divergence between vocational and theoretical programs. The TPs have the overall highest content validity means.

Table 2 illustrates two major distinctions: number one is the different index means between the TPs’ and VPs’ tests (also visible in graphs 2 and 3). The second distinction is the difference between the two judges’ means. J1’s means are overall lower (with an exception on question number five) than J2’s means concerning the TPs’ tests. When it comes to the VPs’ tests the situation is more or less reversed: J1’s means for the VPs’ tests show a higher mean on four out of six questions. The exceptions are questions four and six. These results will be dealt with in more detail in relation to each question.

Table 2. Overview of the index means and the variance between the two judges’ means

<table>
<thead>
<tr>
<th>Question</th>
<th>Group</th>
<th>IM</th>
<th>J1M</th>
<th>J2M</th>
<th>N</th>
<th>JV</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Theoretical</td>
<td>3.63</td>
<td>3.38</td>
<td>3.88</td>
<td>8</td>
<td>0.50</td>
</tr>
<tr>
<td></td>
<td>Vocational</td>
<td>2.75</td>
<td>3.00</td>
<td>2.50</td>
<td>8</td>
<td>0.50</td>
</tr>
<tr>
<td>2</td>
<td>Theoretical</td>
<td>3.88</td>
<td>3.50</td>
<td>4.25</td>
<td>8</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>Theoretical</td>
<td>Vocational</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---</td>
<td>-------------</td>
<td>------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>3.38</td>
<td>3.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.88</td>
<td>2.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.88</td>
<td>2.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.25</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>1.56</td>
<td>1.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.75</td>
<td>1.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.38</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.37</td>
<td>0.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>3.25</td>
<td>2.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.75</td>
<td>2.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>3.75</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1.00</td>
<td>0.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

IM: Index Mean, J1M: Judge One’s Mean, J2M: Judge Two’s Mean, N: Number of tests, JV: The variance between the judges

Graph 2. Displays the judges’ individual for analysing the vocational programs:
Graph 3. Displays the judges’ individual means when analysing the theoretical programs.

![The judges' means on the theoretical programs](image)

J1M: Judge One’s Mean, J2M: Judge Two’s Mean

Question one: *Is there an overall match between the test and the syllabus regarding writing ability?* *(Brown & Hudson, 2002; EN1201 - English A, 2000).* This question includes all the other questions and therefore gives an overall overview of the test’s content validity. The index mean for the tests belonging to the VPs was M = 2.75 and for the TPs it was M=3.63, which indicates a stronger match between the test and the syllabus regarding writing ability in the TPs’ tests. The VPs’ mean indicate that it was ranked below “moderate” and TPs’ mean indicate that it was ranked above “moderate”. This score suggests that the collected achievement tests belonging to the TPs group do tend to bring up more aspects that are covered in the English A syllabus goals for writing. The collected tests used in TPs put a major focus on writing in an expressive manner, writing in a number of contexts and being able to argue values, thoughts and feelings. This can be compared to the collected tests gathered from VPs, which were more concerned with small, restrictive tasks such as filling in the missing word, translating a short text or answering questions referring to facts. These tasks do not require a deeper level of processing, and do not encourage students to express themselves and their perspective.
As Table 3 and 4 shows, there was a divergence between the two judges, with a variance of 0.50 regarding both programs’ tests, where judge one graded lower on the TPs’ tests but higher on the VPs’ tests. This needs to be considered when interpreting the index means but it also stresses something else, namely how subjective it is to analyse a test as well as how to construct one according to the syllabus. Therefore two things have become clear from this question’s results:

1. The index means show a difference between the two programs’ tests regarding to what extent they reflect the syllabus.
2. The subjective role of the judge and the teacher’s subjectivity while construction and marking of the test.

Question two: *Is there an overall match between the test and the objective that it is meant to test?* (Brown & Hudson, 2002, p. 100). This question refers to what purpose the test seems to imply. A high mean was given to the extent the objective is clear to the judges’ and if the test succeeds in representing that objective. Although TPs’ test almost reached the level of “good” both programs’ tests received a mean above “moderate” (TPs’ M = 3.88 and VPs’ M = 3.19). The tests in both programs have an overall match between the test and the objective it is meant to test, which in itself is a good result since it implies a high degree of content validity. An example of such a match is a test of a book review from a TP and some of the items it contained: “your review should contain: title and author, a short account of the plot, the message of the book, your own reflections, your opinion”, which according to the judges matches the objective of a book review very well.

The variance between the two judges (JV) was higher regarding the TPs’ tests, JV = 0.75 compared to VPs’ JV = 0.12. However the two judges scored differently. Judge one’s score had a 0.75 lower mean (J1M = 3.50) than the other judge (J2M = 4.25). Which indicate that there exists a difference between the two judges’ grading on this question as well.

Question three: *Does the test formulate the questions to allow the students to use their ability to express themselves in writing in different contexts?* (EN1201 - English A, 2000). Once again the TPs’ index mean was higher and above “moderate”, M = 3.38, than VPs’ M = 2.81 which is
below the “moderate” level. According to those means, TPs’ tests have a stronger emphasis on making the students express themselves in writing in different contexts. In the analysis, TPs’ tests were found to offer diverse contexts such as South Africa, literature review, movie review, book review and translation. Another relevant aspect discovered in the analysis was that the tests in TPs covered not only more contexts, but covered them in more depth. Whereas in a test belonging to a VP there were fewer questions that were formulated to allow the students to use their ability to express themselves in writing in different contexts.

Another aspect that is worth mentioning is the variance between the two judges’ means. The most substantial difference between them two was found in question three (and question six which is accounted for relating to question six) regarding the TPs’ tests. The score was for the J1M = 2.88 and for the J2M = 3.88, which set them apart and show an interpretation difference. It is difficult to know which one of the judges is closest to the truth, since it is an evaluation process and highly subjective. Even though the two judges differed in their interpretation, it also strengthens the use of the index mean, since that is a mix of both judges’ means and probably says more about the tests regarding the question.

Question four: Does the test’s questions allow the students to be “able to formulate themselves in writing in order to inform”? (EN1201 - English A, 2000). This question received the highest index mean, namely IM = 4.00 for the tests belonging to the TPs. The test questions seem to focusing on letting the students formulate themselves in writing in order to inform. Examples of such are taken from TPs’ tests: “your review should contain: a short account of the plot (what it’s about)” and “write a short summary of the book NN”. This question also received the largest divergence between the two groups means with a difference of +1.38 from the highest M = 4.0 (TPs) to M = 2.62 which belonged to the VPs. The reasons for why “writing in order to inform” seems to occur more frequently in the TPs’ tests, could be as stated in the introduction section that the teachers in TPs follow the syllabus more carefully than the teachers in VPs do. The tests from TPs showed a greater tendency to follow the goals provided in the syllabus for English A. Tasks and test items such as “explain to your reader why you have reached these conclusions.”, “write a short description of the authors’ work” and “write a short essay giving your reasons and arguments”, give the students a better possibility to inform the potential reader of facts or the
student’s own opinions. In the tests collected from VPs the tests were mainly focused on the students giving a short answer, writing the correct word or giving the correct translation. The opportunity for the students to inform using facts or opinions was restrictive. One reason may be that the teachers of TPs follow the syllabus to a greater extent than teachers belonging to VPs. Another reason could be that teachers have a different view of how the tests in different programs should be constructed, as in this case where pupils in TPs are allowed to practice the ability to inform, but this is less prevalent in VPs’ tests.

However, there is also a need to investigate if any divergence occurred concerning the variance between the two judges of the two groups’ tests. And it did; the JV for TPs’ tests was 0.75 (J1M = 3.62 and J2M = 4.38) and the JV for VPs’ tests was 0.25. So it seems that the two judges disagreed on the extent to which question four applied according to the content validity scale, one being more lenient than the other.

The conclusion for question number four would be that the question received an overall high grade regarding TPs’ tests, according to both the index mean as well as the judges’ separate means, which indicates a good content validity. VPs’ tests on the other hand did not receive such high scores and must therefore be considered to have a weaker content validity.

Question five: Does the test’s questions allow the students to be “able to formulate themselves in writing in order to instruct”? (EN1201 - English A, 2000). This question received the lowest score for both programs’ tests. The authors interpreted instruct as ‘to tell an individual to perform some kind of task’. This aspect of the syllabus was rarely found in the collected tests regardless of group. Both VPs’ and TPs’ index means fell under the range of “poor”, namely M = 1.56 for TPs and M = 1.31 for the VPs’ tests. What could be the reasons why question received such a low grade? Could it be that teachers of all programs find it difficult to interpret that part of the syllabus (EN1201 - English A, 2000) in the tests? Or that they are not sure of what instruct means and therefore leave it out? Another possible explanation could be that the teachers constructing these tests know what it means, but find it hard to implement it in a test. A third reason why this aspect is missing in the collected tests could be that traditionally it is the teacher’s role to instruct.
and inform and teachers therefore they find it difficult to construct questions that let the pupils instruct.

The index means for the two groups were both low on this question. By studying Table 2, one can observe that J1M’s and J2M’s means were also low regarding both programs’ tests. Neither of the judges found any substantial proof of questions in the tests that matched question five and therefore the tests received a low degree of content validity regarding testing students in order to instruct.

Question six: Does the test’s questions allow the students to be “able to formulate themselves in writing in order to argue and express feelings and values”? (EN1201 - English A, 2000). Question six investigates to which extent the tests let the students create and formulate their opinions and feelings. The analysis clearly shows that the tests belonging to the TPs have in an above moderate way constructed the tests’ questions to make the students “able to formulate themselves in writing in order to argue and express feelings and values” (EN1201 - English A, 2000) with an index M = 3.25. The VPs’ tests received a low score, with only an M = 1.94, which is considered a “poor” score according to the validity scale. Examples of such test items found in TPs’ tests were: “why do you think this book was written?”, “give reasons for every statement that you make” and “write about your own personal opinions”. In the tests collected from VPs only one test of the eight encouraged students’ to write down their own thoughts and perspectives.

The analysis shows that there is a content validity difference between the two groups’ tests, but there were also differences among the judges’ means. The JV for the TP is 1.00 and 0.38 regarding the VPs. J1M = 2.75 and J2M = 3.75 indicates a disagreement concerning the content validity of question six. While one judge finds the tests below “moderate”, the other judge finds them closer to “good”. Again the divergences that can occur in analysing and scoring tests come back in mind. One must not forget the impact such divergences can have if they occur in other contexts beyond this analysis, such as teachers constructing tests, marking and grading students. The finding of these differences between the two judges will be further dealt with in the discussion section.
The two testing approaches found in the analysis

Two main approaches were found in the collected tests. One approach has a more prescriptive nature, which is the dominating type of test according to Måhl (1991). This first testing approach was closely connected to the VPs. The tests belonging to this approach were more focused upon dividing the language up into several sections and assessing the separate language skills one by one. Another aspect of this testing approach found in the analysis was that it tended to concentrate on one of Brown’s (2004) four aspects of writing, namely the “Intensive” aspect where the focus is form but some attention to context and meaning has been given. Moreover, Brown’s (2004, p. 221) list of micro skills and macro skills was considered during the analysis and the first testing approach did show a tendency to focus on testing micro skills that contain the ability to perform short written communication with a basic vocabulary and word patterns. Another aspect found was that the first testing approach did tend to divide the written language ability into separate areas, such as written vocabulary, translation skills, the ability to form correct sentences, etc. However, as mentioned in the background section, the syllabus states that “English should not be divided up into different parts to be learnt in a specific sequence” (EN1201 - English A, 2000). Another discovery found during the analysis was that this first testing approach included the first level of Bloom’s (1981) taxonomy, as this testing approach focused upon written language ability more concerned with the ability to produce the correct answer than to display a deeper understanding of the subject.

The second testing approach takes a completely different stand compared to the first testing approach in that it focuses upon reflective writing, expressing opinions, thoughts and the ability to connect these to established facts. The second testing approach was in close relation with the TPs. In the analysis it was found that the second testing approach includes the “Responsive” and “Extensive” aspects of Brown’s (2004, p. 220) aspects of writing ability. This was concluded since the test items encourage the students to write short narratives and also to respond to a text which had been read. Furthermore, the second testing approach also included the more advanced writing ability where the students are to achieve a purpose with their writing and categorize as well as expand their thoughts using facts to display an opinion. In the analysis it was found that the first testing approach focused upon Brown’s (2004, p. 221) micro skills for writing, in the second testing approach the macro skills for writing possessed a prominent position since this
testing approach allowed the students to use and adapt a variety of writing strategies to establish connection between events and express ideas or present information. As with the first testing approach some of the knowledge levels included in Bloom’s (1981) taxonomy were found in the second testing approach. It was primarily the “Understanding”, “Analysis” and “Synthesis” levels that were found in the second testing approach. These levels were found in the analysis since the second testing approach promoted the students’ ability to make their own interpretation of a subject and identify certain phenomenon in a text with the purpose of organizing and developing these in writing with the students’ own reflections.

The results of the research questions

Regarding research question number one: “Do any differences exist in the interpretation of the syllabus between the tests used in vocational programs compared to tests used in theoretical programs?” The analysis showed that there is a clear difference between the two test groups (see Table 1 and Graph 1). The tests used in theoretical programs received higher index means in all the questions regarding the syllabus. The higher means indicate that the two test groups interpret the syllabus differently. The group with the higher means represent tests that included what is stated in the syllabus to a greater extent. However, both groups received a low mean on question five: Does the test’s questions allow the students to be “able to formulate themselves in writing in order to instruct”? (EN1201 - English A, 2000) which indicates that neither groups succeeded regarding producing test questions in order to instruct.

There is a clear difference between the VPs’ and TPs’ tests, as far as the index mean is concerned. However, according to Table 2, the results show something else. Table 2 displays that the two judges varied in their analysis of the tests and sometimes more than others. The index means provide a picture of the VPs’ tests having lower content validity than the tests belonging to TPs. This is not the picture if one examines the judges’ own means. Instead the results of the JV showed that the highest mean varied between the different programs’ tests.

Concerning research question number two: “To which degree do the collected tests demonstrate content validity?” The analysis showed that the collected tests demonstrate a ‘moderate’ degree of content validity. Eight out of the total 16 means representing the questions in the content
validity scale received a grade of ‘moderate’ or higher, with the other half receiving a lower grade. Although it is difficult to speak of any general degree since vast differences exist among the tests, the clear difference found in the analysis displays a tendency even in small sample of 16 collected tests. However, it is easier to speak of the general degree in each test group (VP and TP), which shows that the content validity according to the scale used in the analysis of this study is demonstrate a higher degree of content validity (higher means) in the TPs’ tests than in VPs’.

Discussion

The primary objective of this study was to set out to investigate whether any interpretation differences of the syllabus of English A between theoretical and vocational programs could be found in the collected tests. When assessing the result of this study the reader should bear in mind that the results can not under any circumstances be generalized outside this study. To create generalized results is however not the aim of this thesis. This study sets out to analyse the collected tests to gain a deeper understanding of how the syllabuses for English A are interpreted in achievement tests focusing on writing ability. And if there might exist a difference between theoretical and vocational programs in the collected tests of this study concerning this interpretation. A natural step to take since the analysis did reveal a difference between the programs was to investigate wherein the differences lay.

It was the goal of this thesis to seek an answer to the two following research questions:
1. Do any differences exist in the interpretation of the syllabus between the tests used in vocational programs compared to tests used in theoretical programs?
2. To which degree do the collected tests demonstrate content validity?

Concerning the results found, the hypothesis this study was based on does appear to contain some credibility. A clear difference between the tests distributed in the theoretical programs and the vocational programs was found. A more detailed discussion concerning the results will be covered below.

*Interpretation of the syllabus*
In light of the results found, it is in the lights of the results found that the difference between the collected tests is due to a difference in interpretation of the syllabus. The clearest differences between the two groups are discovered in questions one, two, three, four and six. In question one the index of content validity indicates a stronger general match between the test and the syllabus in the theoretical programs’ tests. This score suggests that the collected achievement tests belonging to the TP group tend to bring up more aspects that are covered in the syllabus English A goals for writing. Examples of this are that the collected tests used in TPs put a major focus on writing in an expressive manner, writing in a number of contexts and being able to argue values, thoughts and feelings. This can be compared to the collected tests gathered from VPs, which were more concerned with small and restrictive tasks which do not require a deeper level of processing, and do not encourage students to express themselves and their perspective.

The difference

As mentioned above, a difference was found in the interpretation of the syllabus in the collected tests between the programs. Why does this difference exist? It is the authors’ belief that the teachers constructing the tests use two very diverse approaches towards teaching the two groups and as a result of this divergence between approaches the tests relating to the teaching are bound to be dissimilar. The two different testing approaches that the teachers take will be covered in short below. Although this area could be studied in great detail, to go into it in more depth would be a different study altogether.

The first testing approach, which was prescriptive in nature, occurs in the tests distributed in the vocational programs, while the second, which was descriptive in its nature, is prominent in the theoretical programs. Possible reasons for this might be that teachers on VPs make the assessment that the students on these programs are not fully capable or ready for an assessment method that involves the more advanced aspects included in the syllabus. Reflective writing might be an example of a written ability that the syllabus puts a considerable focus on, and is brought up frequently in the TPs tests, yet the tests collected from VPs rarely included this aspect. The tests collected from the VPs tended to focus on, and assess, a more basic written ability. The teachers most probably have a lower expectation on VP students’ written ability compared to the higher expectations of students attending the Tps. This is also supported by the
findings of Jansson and Carlsson (2005) and Lindberg (2002). Although it might be easy to criticize teachers on VPs who might have lower expectations of their students than teachers on the TPs, it must be taken into consideration that these lower expectations most often will prove to be related to a valid assessment of the situation. It is therefore not the aim of this thesis to point out the negative factors of tests distributed on VPs, but to explore the differences that could be found. The authors of this thesis consider the construction of program-specific syllabuses to be a solution to the obvious difficulty of having the same teaching and testing approach in VPs and TPs. By the term ‘program specific syllabuses’ the authors of this thesis imply syllabuses that are constructed in relation to the program where they will be used, be it a theoretical program or a vocational program. The authors take this stand upon the basis that the two different program directions possible in the Swedish school system (VPs and TPs) attract students with very different ambitions, towards academic studies and it is consequently a strange thought that they are to be taught and assessed upon the same basis.

However, there were more differences that were to be found during the analysis than just between the teachers constructing the tests: namely the divergences between the two judges’ scoring. As Table 2 and graphs 2 and 3 in the results show, the judges differed in their interpretation more than once and regarding some questions also differed from the index means to the extent that one of the judges graded the VPs’ tests higher than the TPs’. The differences found between the judges are interesting in themselves. This indicates that the analysis is highly subjective and therefore raises the issue of subjectivity, not only in evaluating tests but also in constructing, marking and grading. It leads to the question of whether teachers should double mark and grade each others exams to avoid such subjectivity.

The syllabus is written in a way that allows teachers to adapt it to their own teaching approach: however, a drawback of this freedom is the vagueness of the syllabus. It does not instruct the teacher in detail on how to teach, what to test and how to assess. There are, however, goals and guidelines for the teachers to follow in each subject and course, such as the ones mentioned in this thesis regarding ability of writing in English A. Though, as this study found, even those guidelines were interpreted differently by the teachers constructing the tests and the judges evaluating the tests according to the syllabus. Can it be that the syllabus is too vague and non-
specific and would do well to include some guidelines in relation to assessment? This statement is supported by Bormuth (1970) who commented more than 30 years ago on the lack of concern for information and instruction about constructing test items, and observed that most teachers rely upon nothing more than their intuitive skills when constructing a test. This seems to be the case according to Osterlind (1998) who wrote that “Test item writers are routinely left to their own devices because there is no developed theory to underpin item writing, nor is there a comprehensive resource identifying the distinctive features and limitations of test items, the function of test items in measurement, or even the basic editorial principles and stylistic guidelines” (Osterlind, 1998, p. 6) and this still today, leastways in this study.

As mentioned before, Jansson and Carlsson (2005) found that teachers fail to achieve an equal interpretation of the grading criteria on a specific grading level. The interpretation depends on whether the student attends a theoretical program or a vocational one. In their study Jansson and Carlsson also discovered that teachers assume students attending the vocational programs have lower potential to meet the goals in the syllabus. This is supported by the index means of the content validity scale in this study, which are different depending on what program the test belonged to.

Lindberg (2002) found that teachers of the core subjects experienced a decline in the demands for the grade “Pass” (G), since teachers were more or less forced to hand out simplified tasks to students in vocational programs in order for them to receive a “Pass”. In addition, Lindberg (2002) found that teachers considered the syllabuses unclear and vague, which in the light of this study could be leading to the risk that the students are assessed differently between teachers, between schools and between programs.

The consequences of the syllabus being too vague may lead to different tests depending on which teacher constructs the test. This leads towards different grading material and grades, even though it is the same course (in this case English A). As Jansson and Carlsson concluded: different prerequisites concerning the students might be interpreted as the teachers having a different basis for assessment depending on the course of study the student is engaged in and therefore indirectly carrying out a different interpretation of the grading criteria (2005, p. 52). Moreover, teachers of
the theoretical programs worried that the students attending a theoretical program would not be assessed upon the same basis as the students attending a vocational program (Lindberg, 2002, p. 47).

The notion of the syllabus being unfair to students belonging to separate programs has been mentioned before in the background section. Among others Lindberg (2002) carried out an interview study concerning the consequences the new grading system (Lpf 94) had on teachers and students. In Lindberg’s study teachers of the core subjects expressed themselves in sceptical terms concerning the syllabuses and the grading criteria. The teachers considered the collective syllabuses and grading criteria an injustice towards the students. The injustice according to the teachers was mainly based upon the notion that the collective syllabuses made the core subjects more difficult for students in vocational programs and ‘the goals to accomplish’ aspect of the syllabuses placed the vocational students in a situation where they did not have the pre-requisites to accomplish the course. This is however not a findings according to this study. The analysis of the collected tests for this study show that the VPs’ tests received lower index means than the TPs’ tests, in other words the syllabus was not followed to the same extent in the VPs’ tests as in the TP’s.

Something the authors of this thesis will bring along to their future profession as English teachers, is a new awareness of one’s own subjectivity and knowledge of the tools to check and regulate it. Actions such as validating tests and double marking students’ work could contribute to a better educational system. In addition, the writers of this study believe that the information and knowledge gained concerning tests, testing, constructing and validation will enrich their pedagogical occupation on a day-to-day basis.

Source of errors

The results can not under any circumstances be generalized, since the primary objective of this study was to set out to investigate whether any interpretation differences in the syllabus of English A between theoretical and vocational programs could be found in the collected tests. The index of means shows a distinct divergence on how the two groups’ tests differ. It is a interesting
result since the tests should not differ granted that they use the same syllabus and goals for English A.

The content validity scale has not been previously used as it is composed in this study, however it is based on the syllabus for English A and literature regarding content validity (Brown, 1983, Brown & Hudson, 2002). The judges of the content validity scale were the writers of this thesis and to increase the analysis’s interrater validity, a double-blind procedure was used. The procedure itself is covered in more depth in the method section. However it needs to be stated that content validation will generally be a judgemental evaluation and no generalizations can be drawn from such an investigations, which Brown (1983) acknowledges.

The writers of this study are unable to present any of the tests as an example due to the anonymity promised to the teachers whose tests were collected.

**Future suggestions and further research**

The authors of this thesis consider the construction of program-specific syllabuses to be a solution to the obvious difficulty of having the same teaching and testing approach in VPs and TPs. By the term ‘program-specific syllabuses’ the authors of this thesis imply syllabuses that are constructed in relation to the program where they will be used, be it a theoretical program or a vocational program. The authors take this stand upon the basis that the two different program directions possible in the Swedish school system (VPs and TPs) attract students with very different ambitions towards academic studies, and it is consequently a strange thought that they are to be taught and assessed upon the same basis.

The differences found between the judges are interesting in themselves. This indicates that the analysis is highly subjective and therefore raises the issue of subjectivity, not only in evaluating tests, but also in constructing, marking and grading. It leads to the question of whether teachers should double mark and grade each others exams to avoid such subjectivity. The writers of this thesis suggest double marking since there is such a vast difference between teachers’ judging. Double marking of one’s’ pupils tests and work is something the writers of this thesis will try to implement in their own teaching. One has to be aware of the subjectivism that occurs across the
full spectrum of teaching activities. To strengthen the content validity regarding the syllabus it is helpful to evaluate oneself and one’s work with the help of others, such as double marking or validating tests given to students.

Further research within the field would be needed since only a small number of previous studies have been made. Such research suggestions could be to carry out a quantitative study where one investigates a larger sample of tests taken from different schools and programs to see if any of the results from this study can be found in a larger study. Moreover, it would be interesting to interview teachers constructing the tests and the pupils taking the tests to explore their thoughts concerning the issue of content validity and the tests’ link to the syllabus. Another focus future research could take, is to explore the possibilities of differences in interpretation of the syllabus within the specific programs rather than between programs.

More research into the two testing approaches would be interesting as this thesis found a clear difference in testing approaches. If one takes the subject further the difference in testing approach is maybe a product of a difference in teaching method. A study focusing upon these different teaching methods would assist in giving a more complete picture of the research area.

The writers of this thesis suggest double marking since there is such a vast difference between teachers’ judging. This thesis has shown that the interesting point lies within the two judges’ different grades, however, it is still interesting to observe that the two programs’ tests did tend to differ according to the analysis made in this study.
References


Bedömning av kvalitativ kunskap Konkreta exempel från gymnasieskolan. Hässleholm: Gymnasietidningen/Grundskoletidningen/Fortbildningsförlaget. Internationella gymnasiet och fortbildningsförlaget


http://www3.skolverket.se/ki03/front.aspx?sprak=EN&ar=0708&infotyp=5&skolform=21&id=3199&extrald
