



Høgskulen på Vestlandet

MUNDD511: Masteroppgave

MUNDD511

Predefinert informasjon

Startdato:	10-05-2017 10:16	Termin:	2017 VÅR
Sluttdato:	15-05-2017 14:00	Vurderingsform:	Norsk 6-trinnskala (A-F)
Eksamensform:	Masteroppgave	Studiepoeng:	45
SIS-kode:	MUNDD511 1 MG		
Intern sensor:	Kristian Andersen Rusten		

Deltaker

Kandidatnr.: 4

Informasjon fra deltaker

Tro- og loverklæring *: Ja

**Jeg godkjenner avtalen om ja
tilgjengeliggjøring av
masteroppgaven min *:**



**Western Norway
University of
Applied Sciences**

MASTER'S THESIS

Test validity - a teacher's responsibility?

A closer look at which aspects of validity are in the hands of teachers when implementing *The Diagnostic Test of English for Grade Three*

Test validitet - en lærers ansvar?

En nærmere undersøkelse av hvilke aspekt av validitet er læreres ansvar i gjennomføringen av *Kartleggingsprøven i engelsk for 3.trinn*

Anne-Britt Glittenberg Elvegård

**Master i undervisningsvitenskap med fordypning
i engelsk didaktikk
Avdeling for lærerutdanning**

May 15, 2017

I confirm that the work is self-prepared and that references/source references to all sources used in the work are provided, cf. *Regulation relating to academic studies and examinations at the Western Norway University of Applied Sciences (HVL), § 10.*

Acknowledgements

Before I decided to go back to school part time and get a master's degree, writing a master's thesis seemed like an impossible feat to me. And ever since my journey started in the autumn of 2012, my thesis has stood before me like a mountain - a very tall, steep, and intimidating mountain. But it so happens that I am rather fond of mountains, and I see no point in hiking them if I'm not going to reach the top. This "stubbornness" came very much in handy when I stumbled upon a rather big and unexpected bump in the road. Many times I've wanted turn back halfway up "Mt.Thesis" and call it quits. But the best way to reach the summit is by taking it one step at a time. In this case, I had to take it one chapter at a time.

From the top of a mountain you can see how far you've come, and if you are like me, you almost forget how strenuous the hike has been. Seeing the view from the top makes it all worth it.

I would like to thank all my wonderful colleagues at the Research Group for Language Testing and Assessment at UiB for their help, encouragement and support. Thank you for giving of your time and your patience, and for sharing your wisdom and insight with me. Thank you for letting me think aloud and for giving me helpful feedback using the "best words". Last but not least, thank you for making me laugh on days when I was very frustrated.

I would also like to thank my friends and family for their prayers and support, and for believing in me when I did not believe in myself.

You have all been invaluable hiking mates, and I could not have made it to the top without you. But the next time we go hiking, please let it be Mt. Ulriken.

Abstract

The aim of this study is to examine more closely the validity of the KPE3 (*The English Diagnostic Test for Grade Three 2014*), which is a test that assesses young learners' (third grade) listening and reading skills. The test is not obligatory and the purpose of the test is to identify the pupils who fall below the limit of concern, and therefore need extra support. Using Messick's six aspects of validity as a framework and drawing on the analysis model developed by Hasselgreen (2004), the study aims were to consider what the potential threats to the validity of the KPE3 (2014) were, and more specifically, to identify which of those aspects were in the hands of teachers/schools. Finally, the study looked for evidence to indicate that some of the aspects related to teachers' practices may actually be undermining the test's validity.

Two data sources were used for this study: a questionnaire that was sent out to teachers who had administered the test, and a sample of pupils' test booklets. The main finding of the study was that there was evidence indicating that teachers' practices were a threat to the validity of the test for various reasons. Not all teachers surveyed used the practice material to familiarize their pupils with the test, and there was also a common misconception regarding the purpose of the test – over half of the teachers surveyed believed that the test could find the level of the entire class. In addition, only half of these teachers believed that the purpose of the test was to give feedback to pupils. An examination of the test booklets revealed that marking errors were made at all but two schools, which poses a serious threat to validity. Finally, 54.2% of these twenty-four teachers lacked training in both English and language assessment. All of these factors may have had implications for the validity of the KPE3 (2014).

Although test developers and the authorities who commission the tests will always have the ultimate responsibility when it comes to test validity, this study highlights the responsibilities teachers and schools have with regards to ensuring test validity.

Sammendrag

Målet med denne studien er å undersøke nærmere validiteten til *kartleggingsprøven i engelsk for 3. trinn* (2014). Dette er en prøve som måler elevenes lytte og leseferdigheter i engelsk. Prøven er ikke obligatorisk, og formålet med den er å identifisere elevene som faller under bekymringsgrensen og trenger ekstra oppfølging. Ved å bruke Messicks validitetsteori med seks aspekter som rammeverk og analysemodellen utviklet av Hasselgreen (2004), var formålet med studien å vurdere hvilke potensielle trusler mot validiteten av *kartleggingsprøven i engelsk for 3. trinn* (2014) som fantes, og mer spesifikt å identifisere hvilke av disse aspektene som var læreres/skolers ansvar. Studien undersøkte om det var bevis for å indikere at noen av aspektene knyttet til lærernes praksis faktisk kan undergrave testens validitet.

To datakilder ble brukt til denne studien: et spørreskjema som ble sendt ut til lærere som hadde gjennomført prøven, og elevers ferdigutfylte oppgavehefter. Hovedfunnene i studien er at det var bevis på at lærernes praksis var en trussel mot prøvens validitet av ulike årsaker. Ikke alle lærere som er innhentet informasjon fra, brukte øvelsesmaterialet til å forberede elevene og gjøre dem kjent med oppgavetyperne, og det var mange misforståelser om formålet med prøven. Over halvparten av de undersøkte lærerne trodde at prøve kunne identifisere språknivået til alle elevene i klassen. I tillegg trodde bare halvparten av disse lærerne at formålet med prøven var å gi tilbakemelding til elevene. En undersøkelse av prøveheftene viste at det ble gjort rettefeil, med unntak av lærerne fra to skoler. Dette utgjør en alvorlig trussel mot prøvens validitet. Til slutt manglet 54,2 % av de 24 lærerne, som studien fokuserer på, opplæring i både engelsk og språkvurdering. Alle disse faktorene kan ha hatt implikasjoner for validiteten til *kartleggingsprøven i engelsk for 3. trinn* (2014).

Selv om prøveutviklere og myndighetene som bestiller prøvene, alltid har det ultimate ansvaret når det gjelder testvaliditet, fremhever denne studien det ansvaret som lærere og skoler har med hensyn til å sikre testens validitet.

Table of Contents

Acknowledgements.....	i
Abstract.....	ii
Sammendrag	iii
Table of Contents.....	iv
List of Tables and Figures.....	vi
1. Introduction.....	1
2. The test.....	4
2.1 The purpose of the test.....	4
2.1.1 Description of KPE3	5
2.2 The Framework.....	6
2.3 The Curriculum.....	7
2.3.1 Test construct.....	8
2.4 Test content.....	8
2.4.1 Listening section	9
2.4.2 Reading section.....	11
2.5 Test materials.....	14
3. Theory.....	16
3.1 Previous research on low-stakes diagnostic testing.....	16
3.2 Formative Assessment	18
3.3 The KPE3 Construct	19
3.4 Test validity	19
3.4.1 Messick’s six aspects of validity.....	22
4. Examination of the KPE3 in the light of Messick’s six aspects of validity.....	24
4.1.....	24
4.1.1 Content aspect.....	25
4.1.2 Substantive aspect.....	26
4.1.3 Structural aspect.....	26
4.1.4 Generalizability aspect.....	27
4.1.5 External aspect.....	28
4.1.6 Consequential aspect.....	28
4.2 Summary of the potential threats to validity at the hands of teachers/schools	29
5. Methodology.....	31
5.1 Choice of methodology.....	31

5.2 Sample selection	32
5.3 The questionnaire.....	34
5.3.2 Analysis of the questionnaire data	40
5.4 Analysis of the test booklets	40
6. Findings.....	42
6.1 Findings from the questionnaire	42
6.1.1 Block 1: Biodata	42
6.1.2 Block 2: Findings linked to research question.....	44
6.1.3 Block 3: Findings indirectly linked to research question.....	48
6.1 Marking Errors.....	51
7. Discussion.....	54
7.1 The six aspects of validity.....	54
7.1.1 Content Validity.....	54
7.1.2 Substantive Validity.....	57
7.1.3 Structural Validity.....	59
7.1.4 Generalizable Validity	61
7.1.5 External Validity.....	63
7.1.6 Consequential Validity.....	64
7.2 Examples: Linking questionnaire data and marking errors to validity threats	65
7.3 The responsibility of test developers/authorities	69
7.3.1 Content Validity.....	70
7.3.2 Substantive Validity.....	70
7.3.3 Structural validity.....	71
7.3.4 Generalizable Validity	71
7.3.5 External Validity.....	71
7.3.6 Consequential Validity.....	72
7.3.7 Where responsibilities meet.....	72
7.4 Weaknesses and limitations of the study	72
8. Conclusion	74
References.....	79

List of Tables and Figures

Figure 1: Graph of National test and KPE3	5
Figure 2: Draw a line (Listening section)	9
Figure 3: Mark the correct picture (Listening section)	10
Figure 4: True or False (listening section)	10
Figure 5: Mark the correct answer (Listening section)	11
Figure 6: Draw a line (Reading section)	12
Figure 7: Mark the correct picture (Reading section)	12
Figure 8: Mark the correct text (Reading section)	13
Figure 9: True or False (Reading section)	13
Figure 10: Copy the correct word (Reading section)	14
Table 1: Aspects of validity and threats in the hands of teachers/schools	29
Table 2: Summary of data collection	34
Table 3: Summary of questionnaire questions linked to potential threats to KPE3's validity	39
Table 4: Question 10 - Participants by geographical region	42
Table 5: Question 12 - Work experience	43
Table 6: Question 13 - Residence in an English-speaking country	43
Table 7: Question 1 - Priority when preparing for test	44
Table 8: Question 2 - Purpose of test	45
Table 9: Question 3 - How the test is used	45
Table 10: Question 5 - Influence of test on teaching	46
Table 11: Question 8 - Use of answer key	46
Table 12: Question 14 - English education	47
Table 13: Question 15 - Training in language assessment	47
Table 14: Question 4 - Decision to administer the KPE3	49
Table 15: Question 6 - Necessity of KPE3	49
Table 16: Question 9 - Teachers' opinions of KPE3	50
Table 17: Marking errors	51

1. Introduction

Having worked as a language test developer at the Research Group for Language Testing and Assessment at the University of Bergen since 2009, part of my job description has been to work with *Kartleggingsprøven i engelsk for 3. trinn*¹, which was first administered in 2010. In the course of this work, I have become concerned about how the test results were marked and how the test was being used. From 2010-2015, the test was paper-based, and each year approximately 500 pupils' test booklets were collected for data processing. In the spring of 2012, I discovered that teachers were making marking errors, despite the fact that the test was very simple and the teachers were provided with an answer key.

With this in mind, the decision to investigate further was made. In the autumn of 2012, a school which had displayed particularly many correction errors was contacted. Here I was informed that they had administered the test because they wanted some feedback to give to parents at parent-teacher conferences. It was a matter of concern that teachers were using this test as a basis for feedback to parents since the test is only designed to identify pupils who seem to be struggling in the subject. The test aim is to identify those who need extra attention at an early stage in their English learning, whether it is weak reading or listening skills. The test is not constructed to give information about the strong or average pupils.

Diagnostic tests are one measure used by schools to assess and gauge their pupils' learning and can be seen as a component of modified teaching (Kunnskapsdepartementet, 2007-2008). Modified teaching is a core principle in the Norwegian public school as established in the Education Act's principles of learning², and can be defined as "those measures taken by the schools to ensure that pupils get the best possible outcome of their education" (translation my own) (Kunnskapsdepartementet, 2010-2011). It is therefore essential that the tests used by schools "work" as they should, i.e. are valid. While much of the validity is in the hands of

¹ Kartleggingsprøven i engelsk på 3.trinn has been loosely translated to *The Diagnostic Test of English for Grade Three*, although this is not strictly what is normally meant by a diagnostic test. See (Hughes, 2003, p. 15).

² <http://www.udir.no/laring-og-trivsel/lareplanverket/prinsipper-for-opplaringen2/laringsplakaten/>

their designers, as will be briefly discussed in chapters 2 and 7, clearly, flaws in the way tests are administered, marked and/or used can also affect validity.

Further investigation was needed in the hope of identifying which aspects could threaten the validity of *The Diagnostic Test of English for Grade Three* (from this point forward referred to as KPE3³).

The research questions of the study were as follows:

- 1) What aspects of the KPE3 potentially make its validity vulnerable?**
- 2) Which of these aspects are in the hands of the teachers/schools?**
- 3) Is there evidence to indicate that some of the aspects related to teachers' practices may actually be undermining the test's validity?**

The thesis will begin in the next chapter with a closer examination of the KPE3 test itself. The purpose of the test, and what principles the test is based on, is discussed. The test material is presented with examples, as well as a review of the teacher's guide.

Chapter three discusses the theoretical background for this thesis, first presenting previous research and then examining the reading and listening constructs of the test and how they are linked to the competence aims of the Norwegian curriculum. Formative and summative assessment and their relevance to the test are elaborated upon before introducing the concept of validity and finally narrowing the focus to Messick's six aspects of validity.

The following chapter continues to examine Messick's six aspects of validity and how they relate to the KPE3. Drawing on Hasselgreen's analysis model for identifying key elements crucial for identifying various types of validity and what poses a threat to them, the aspects of validity that lie in the hands of test makers/authorities and those that are in the hands of teachers/schools will be outlined.

³ Kartleggingsprøven i engelsk på 3.trinn

Chapter five presents the methodology used in this study and the reasons for the choices that were made. The theory underlying the quantitative approach and the theoretical basis for the construction of the study's questionnaire are introduced, and the participants are presented.

Chapter six focuses on the findings of the study. The results are clearly presented in tables and summarized.

In chapter seven, the findings are discussed in the light of the research questions. The discussion examines the possible causes of areas of concern that the study has revealed, as well as the implications the findings have for the KPE3 and other tests. Thereafter, the data collected for four teachers, including questionnaire answers and marking, is discussed, before briefly considering what responsibilities test developers have for ensuring the test's validity. The chapter concludes with a consideration of the weaknesses and limitations of the study.

The final chapter sums up what has been done and presented in this study, and raises issues that require further research.

2. The test

In this chapter, the focus is on KPE3, presenting the purpose of the test, and a description of the test itself and what it is designed to measure. The framework document setting the parameters for the test, which the test developers followed when developing the test, is summarized, and the curriculum that the test construct is based on is also outlined. Furthermore, the test content is described together with task examples from the test itself. Lastly, the test materials that make up and accompany the test are presented.

2.1 The purpose of the test

The purpose of *The Diagnostic Test of English for Grade Three* is clearly stated on the Norwegian Education Directorate's website. Until January 2015, the purpose was stated as follows:

The Diagnostic Test of English for Grade Three for grade three examines whether there are pupils who need extra support and facilitation in their learning of English. The results of the test provide information about pupils who are at or below the defined level of concern (Utdanningsdirektoratet). (translation my own)

It has since been revised to:

The Diagnostic Test in English for Grade Three is to be used by schools and teachers to identify pupils in need of extra attention in English at an early stage. Most pupils will be able to answer the majority of the tasks, and many will find the test easy. The test does not give information about pupils with good English skills (Utdanningsdirektoratet, 2016). (translation my own)

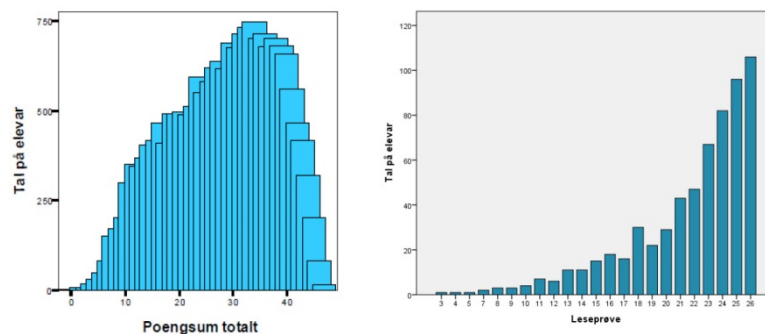
Although both descriptions define the purpose of the test, the latest explanation leaves little room for misinterpretation. The test is designed to identify students who need extra attention at an early stage in their English learning, whether it is weak reading or listening skills. The test is not constructed to provide information about the strong or average pupils.

2.1.1 Description of KPE3

The KPE3 is not a mandatory test. The test is designed to be administered to third grade (7-8 year olds) pupils. It can be categorized as a low-stakes test since it has no significant consequence for the test taker. The test scores are not used, for example, to compare schools, rank teachers, or to determine whether the test taker is accepted to University or granted citizenship to a country. Tests with such consequences would be considered high-stakes. In other words, the only consequence the KPE3 is intended to have for the test takers is to identify those who are at or below the limit of concern so that they can receive the extra learning support that is needed. Since most pupils will manage most of the tasks, the KPE3 does not distinguish between good scores and *very* good scores. One cannot say that a pupil who scores 25 on the reading section is a stronger reader than the pupil who scores 22. The reason for their scores may be random because the test is not constructed to differentiate between the high scorers.

Figure 1 shows the difference between the National Test in English construct and that of the KPE3. The National Test discriminates on the whole spectrum while the KPE3 only discriminates around the level of concern. The National Test places pupils on all levels, while the KPE3 can only say if a pupil is above or below the level of concern.

Figure 1: Graph of National test and KPE3



*Figure 1: Visualization of difference between National test (left) and KPE3 (right)
In both graphs the x-axis represents the score and the y-axis the number of pupils.*

According to SSB⁴, there were 623 800 pupils attending primary and lower-secondary schools in the autumn of 2015. Dividing that number by the number of grades (10), the average grade has approximately 60 000 pupils. In 2014, a little over 52 000 KPE3 test booklets were sent out to primary schools at their request. That means that roughly 87% of Norway's third-grade pupils took the KPE3, although we cannot know for sure the exact number. In 2016⁵, a little over 47 000 pupils took the digital KPE3, or approximately 78 % of the third graders in Norway.

2.2 The Framework

The Framework for Diagnostic Tests for Grades 1-4 is an unpublished six-page document authored by The Directorate of Education and Training. The Framework was the blueprint that the test developers were required to follow when constructing the paper-based KPE3 first administered in 2010.

The Framework is divided into three parts. The first section, the purpose of the framework, outlines how the task of developing the test has been organized and who is responsible for which aspect of the test. The responsibilities of the test developers, in this case the University of Bergen, Institute of Foreign Languages, Research group for language testing and assessment (hereafter referred to as 'the test development team') are as follows:

The test developer is responsible for the development and piloting of the test items and for involving the Directorate in the entire process (Utdanningsdirektoratet, 2007, p. 1).
(translation my own)

Part two explains the aim and the content of the test. The test development team was to design the KPE3 to cover a broad skill set, and to ensure that it was rooted in the curriculum and the Framework for diagnostic testing. In cases where there are no learning objectives in the curriculum for the test year (the curriculum lists learning objectives after grade 2, 4, 7 and

⁴ <https://www.ssb.no/en/utdanning/statistikker/utgrs>

⁵ Please note that the KPE3 was digitalized in 2016. Although some of the challenges that this study revealed and that will be discussed in later chapters have since been remedied and/or improved upon, this thesis will focus on the original test that ran from 2010-2015.

10), the team should adapt the English learning objectives for reading and listening that the pupils are working towards. As the KPE3 is administered in grade three, it was designed to test the learning objectives for grade two, but considerations were made and the content was adjusted to account for the fact that the pupils were a year older. The KPE3 only tests the pupils' ability to understand and recognize words, phrases and simple sentences, and only covers part of the competence aims in English (Utdanningsdirektoratet, 2007). Part three defines what the requirements are for the test and its implementation. The most important requirement is that the test must be valid and measure the skills it is designed to measure. In the case of KPE3, which tests reading and listening skills in English, the test must identify two "levels of concern" for each skill, and it must demonstrate that it actually tests these two different skills. The scores indicating the levels of concern are determined by The Directorate of Education. The curriculum learning objectives had to be taken into account along with a didactic analysis of those objectives and how the skills assessed in the test are built up (Utdanningsdirektoratet, 2007).

2.3 The Curriculum

One of the overarching goals of the *National Curriculum for Knowledge Promotion in Primary and Secondary Education and Training* (LK06)⁶ is that pupils' academic performance should be improved. Based on the Ministry of Education Report to Parliament no. 31 (Kunnskapsdepartementet, 2007-2008), this should partly be achieved by increasing the educational support to all pupils. According to the Report to Parliament no.18 (Kunnskapsdepartementet, 2010-2011), the results of diagnostic tests should give schools and teachers a better basis for adjusting their teaching to meet pupils' needs.

The competence aims in the LK06 for the end of year 2 include aims in language learning, verbal communication, written communication, and culture, society and literature. Since the KPE3 is a test of English reading and listening, the following aims were focused on in the development of the test:

- to understand simple instructions given in English

⁶ The curriculum used is pre-reform and has since been revised, but this is what was current when the KPE3 and teacher's guide were being developed.

- to understand common words and phrases relating to the pupils' immediate environment
- to recognize some words, expressions and simple sentences in spoken and written texts
- to read some simple sentences
- to write some words (Utdanningsdirektoratet, 2006)

These learning objectives were to be the foundation for the content of the KPE3. The test developers based their work on these learning objectives from the curriculum.

2.3.1 Test construct

The concept or the characteristic that a test is designed to measure is often referred to as the test construct. Bachman and Palmer (1996) consider the construct to be the “unobservable language ability” that has to be operationalized (p. 45).

In the case of the KPE3, reading and listening ability are the constructs, or “unobservable” abilities that are being operationalized, as expressed in the curriculum aims mentioned above. According to the teacher’s guide, these aims, which are very concrete, are what the test items are based on.

2.4 Test content

At the time of this study, the KPE3 was paper-based and came in the form of a booklet divided into a listening and a reading section. The audio section of the test was on a CD. The test had fifty items: twenty-four in the listening section, and twenty-six in the reading section. Twenty minutes was allocated to the listening section of the test, and twenty-five minutes to the reading section. It was possible to get 50 points on the test, one point per task. 24 points could be earned in the listening section and 26 in the reading section.

2.4.1 Listening section

In the listening section, pupils heard the instructions on the CD and were given an example, but they could also read a simplified version of the instructions above each task.

Below are some examples of the listening test items. The pupils heard each task twice. Figure 2 was the first task of the listening section. The classroom setting in the picture was one with which the pupils should be familiar. After hearing the example task, the following instructions were given on the CD: “Look at the picture and listen, then draw a line. Task one. Find number one. Draw a line from the number one to the pencil case.” The instructions were then repeated once. The instructions were the same for the remaining numbers in this test format.

The pupils were asked to identify the following items:

- 1) pencil case
- 2) blue jumper
- 3) fruit on the teacher’s desk
- 4) person wearing white socks
- 5) box where sandwiches are kept
- 6) bag used for carrying school books

Figure 2: Draw a line (Listening section)

Draw a line
Look at the picture and listen.

Example 4 5 2

3 6 1

The next task format asked the pupils to look at the pictures and listen, and then select the correct answer. These tasks involved understanding simple words and phrases. In task 12

(Figure 3), the pupils heard the following: “People wear these on their hands to keep them warm.” They were meant to identify the mittens.

Figure 3: Mark the correct picture (Listening section)

Task 12

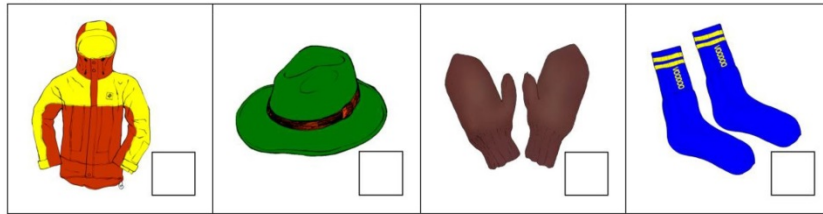
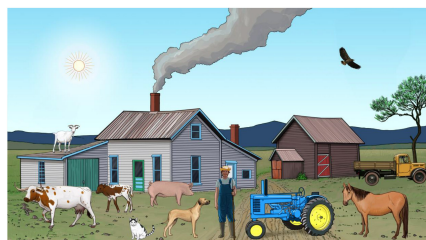


Figure 4 is an example of a true or false task where the pupils heard a statement about the picture and had to decide whether the statement was true or not. The instructions were, “Look at the picture and listen, then mark true or false.” The pupils heard the following statement for Task 13: “The pig is black.” The correct answer was therefore F (false), as the pig in the picture was pink. The remaining tasks in the true or false section were mainly statements where the pupils had to understand colors and animals in order to solve the task.

Figure 4: True or False (listening section)

True or False

Look at the picture and listen.



	True or False	
	T	F
<i>Example</i>	T	
Task 13		
Task 14		
Task 15		
Task 16		
Task 17		
Task 18		


Figure 5 is an example of a task where the pupils heard a short dialog between two people. They were instructed to mark the correct answer and to “look at the question on your paper and listen.” Then they heard the following dialogue between a mother and a daughter:

Mary: Mum, where are my new shoes?
Mum: They’re under your bed, Mary. Can you see them?
Mary: Yes! Here they are!

After listening to the dialogue twice, the pupils had to read the question and answer what Mary was looking for by checking one of the boxes.

Figure 5: Mark the correct answer (Listening section)

Task 23

<p>What is Mary looking for?</p> <p><input type="checkbox"/> Her shoes</p> <p><input type="checkbox"/> Her bed</p> <p><input type="checkbox"/> Her mum</p> <p><input type="checkbox"/> Her t-shirt</p>	
--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------	-------------------------------------------------------------------------------------

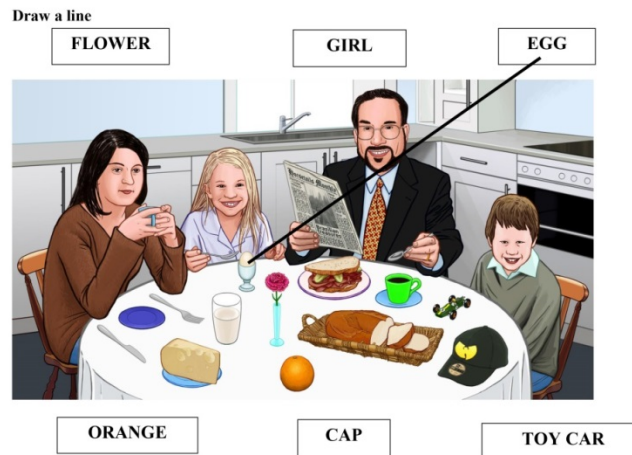
2.4.2 Reading section

Pupils who read fluently can mostly decode words automatically, meaning they recognize the whole word. Most English words, including quite common words that the pupils should know, cannot be sounded out as easily as Norwegian words because the spelling is often not as “logical” or phonetic as Norwegian spelling. This means that without some automatic recognition of whole words, or at least parts of words, the weakest pupils cannot read the tasks even with very well-known vocabulary.

Figure 5 was the first task in the reading section of the test. Here pupils were asked to match items in the picture with its proper name. “Egg” was given as an example. The instructions

were simple and the words and environment depicted in the picture were familiar to the pupil, as stated in the learning aims previously mentioned.

Figure 6: Draw a line (Reading section)

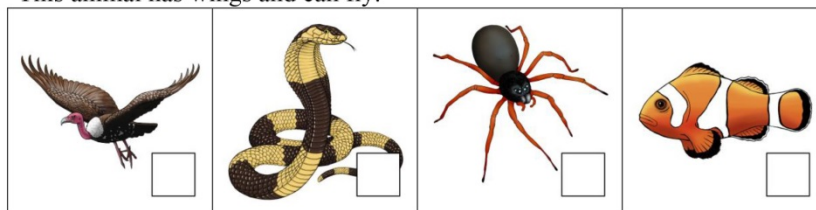


For this next task, the pupils had to choose the correct animal based on the clues given in a simple sentence. The key words they had to recognize and understand were “wings” and/or “fly”, which they then had to connect with the bird.

Figure 7: Mark the correct picture (Reading section)

Task 6


This animal has wings and can fly.



The following task is very similar to the former task type. The pupils had to look at the picture and mark the correct text. For task 11 (Figure 8), the pupils had to recognize the word “puppy”.

Figure 8: Mark the correct text (Reading section)

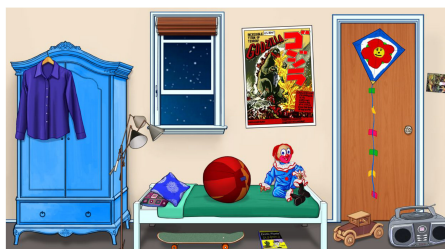
Task 11

	This is a chicken. <input type="checkbox"/>	This is a kitten. <input type="checkbox"/>	This is a puppy. <input type="checkbox"/>	This is a calf. <input type="checkbox"/>
-----------------------------------------------------------------------------------	----------------------------------------------------	---------------------------------------------------	--------------------------------------------------	-------------------------------------------------

The reading section also included a true or false task which is shown in Figure 9. This time the pupils had to read a sentence, look at the picture and decide whether the sentence was true or not.⁷

Figure 9: True or False (Reading section)

True or False



		True or False	
		T	F
Example	<i>There is a ball on the bed.</i>	T	
Task 16	The bed is black.	_____	
Task 17	The door is open.	_____	
Task 18	There is a CD-player by the door.	_____	
Task 19	The wardrobe is blue.	_____	
Task 20	There are four lamps in the room.	_____	




The last task is exemplified in Figure 10. The pupils were asked not only to read, but also to copy the word. It is important to note that this is not a writing task per se. The pupils were being tested on whether they could recognize a word (read and understand its meaning) and match it to the correct picture.

⁷ It might be worth noting that the true or false tasks were removed from the KPE3 when it became digitalized in 2016, as this task type allowed for a 50% chance of getting it correct by guessing.

Figure 10: Copy the correct word (Reading section)

Copy the correct word

sandwich	cake	bicycle	ear	cap
potatoes	bread	shoes	cat	doll

		Your answer cat
Task 21		
Task 22		
Task 23		
Task 24		
Task 25		
Task 26		

2.5 Test materials

The test materials consisted of the test booklets, the audio CD and the teacher’s guide. The teacher’s guide was a document consisting of 26 pages plus a practice test to familiarize the teacher and the pupils with the different task types. It was very detailed and consisted of four sections. These sections included general information about the test, about how to prepare for and administer the test – including implementation instructions such as what to say and how much time to allocate for each section – and about how to follow up the pupils. It explained the different types of tasks and the skills they measure, based on the curriculum. The guide also explained how to mark and score the test and the teachers were provided with an answer key. The KPE3 test booklet and the teacher’s guide can be found in Appendix A and B respectively.

As previously mentioned, the KPE3 assesses the degree pupils to which pupils in grade three have reached the learning objectives for grade two English reading and listening. The main objective of the test is, however, to identify the 20% weakest pupils, i.e. those under the level

of concern, in order to provide them with the necessary help to assist them to achieve the learning aims of the curriculum. That is not to say that the test could not help further learning for the pupils above the level of concern. However, the suggestions and support offered in the teacher's guide was aimed at those needing extra attention. The guide proposed follow-up work to help the teacher to assess the challenges facing a particular pupil, and advised teachers on how to further develop their pupils' skills in English.

Summary:

In this chapter, the purpose of the KPE3 as well as the test itself have been described and presented, together with the test materials. The principles to be used in developing the test and the teacher's guide have been outlined, as these are specified in *The Framework for diagnostic tests for grades 1-4*; and the relationship between the test construct and sections of the *English Curriculum* (LK06) for grade 2(3) has been illustrated.

3. Theory

In this chapter, some theories and principles are discussed that relate to central issues in this study: formative assessment, test construct, and the issue of validity. These are first discussed in general, culminating with Messick's six aspects of validity which provide the basis for the theoretical discussion in chapter four. First, some relevant studies of low-stakes tests are discussed, including diagnostic tests (kartleggingsprøver) used in Norway, since these may shed light on potential weaknesses associated with this type of testing.

3.1 Previous research on low-stakes diagnostic testing

A number of studies have been conducted on low-stakes diagnostic testing for young learners in Norway. Danielsen (2012) has investigated how the results of the mandatory diagnostic tests of Norwegian reading for grade two (kartleggingsprøven i lesing på 2.trinn) are being followed up and what educational value teachers and school leaders consider the test to possess. She has investigated the routines and procedures that are in place to give feedback to school leaders, pupils and parents, as well as to follow up the pupils who come at/under the limit of concern.

Danielsen's study reveals that teachers and school leaders are generally aware of the educational value of the tests. All the school leaders reported that there were routines in place for following up the pupils who needed help, and stated that it was not up to the individual teachers to establish such routines. Teachers reported that feedback to parents was given at parent-teacher conferences. Feedback to pupils was also generally given at parent-teacher conferences.

With a similar aim, the municipality of Stavanger hired an auditing company to write a report on how the results of the diagnostic test in Norwegian reading (kartleggingsprøven i lesing) were being followed up in their school district. The report concluded that schools were taking remedial action that had positive effects on individual pupils, although the statistics indicated that the number of weak readers was increasing over time. The study could not come to any conclusion about the quality of the remedial action, but the 'PP-tjenesten' (special needs

support services for schools) claimed that there was great variation in how schools utilized the test results, and that the action taken was not always very well thought out. Their main concern was that the measures being taken were often too general and not specific enough for the individual pupils (Rogaland Revisjon IKS, 2012).

These findings are supported by studies by Øren Gjelsvik (2012) and Ingebrigtsen, Skarprud, & Stenslund (2015). Their studies of how the diagnostic test (kartleggingsprøver) results have been followed up, and of the routines that are in place to help pupils who score below the limit of concern, reveal considerable variation from school to school. There also seems to be a correlation between the quality of the remedial action being taken and how much the school leadership is involved in this work. Øren Gjelsvik (2012) also reports that her informants were very unhappy that the results were being used to compare not only schools in the same district, but variations from year to year within the same school. If one year had scored at a certain level, the principal expected similar results the following year. Since the teachers differed from year to year, they felt exposed and believed their class's performance reflected on them as teachers (Øren Gjelsvik, 2012, p. 61).

A study by Abrams, Pedulla, & Madaus (2003) on state-mandated tests, both high- and low-stake, in the United States, showed that, even if the test was considered low-stake, 63% of the teachers reported that the test influenced their teaching in a negative manner that contradicted their own beliefs about what they considered sound educational practice (p. 23). These results confirm that a low-stake test can be used as if it were high-stake, and that it can have a negative washback effect. The washback effect is the influence a test has on the teaching (McNamara, 2000, p. 72).

There have not been, to the best of my knowledge, published studies on the KPE3 specifically. The Norwegian studies on diagnostic tests in general, all focused on how the test results were followed up after the tests had been administered and none of these studies consider the issue of follow-up in the light of validity theory. Though there were reports of teachers who considered that the tests had a negative washback affect, the attitudes towards the tests were generally perceived as being positive. Nevertheless, there were weaknesses in the way test scores were used, and the quality of the follow-up routines seemed to vary from school to school to the extent that pupils in need did not always receive the help they required to enhance their learning.

Since the stated ultimate aim of the KPE3 was to identify pupils requiring extra support and facilitation in their learning of English, the test can be considered to be mainly formative in nature, i.e. assessment for learning.

3.2 Formative Assessment

Formative assessment, also called assessment for learning, has been an area of focus for the Norwegian Directorate of Education and Training since 2010. This initiative was based on a project entitled “Bedre vurderingspraksis” (Improved assessment practice) that was carried out between 2007-2009. The focus on formative assessment arose due to the fact that international studies had shown that this was one of the most effective ways to strengthen the learning benefit of the pupils and to increase their opportunity to learn (Utdanningsdirektoratet, 2011).

It is generally accepted that assessment *for* learning can be equated with formative assessment and assessment *of* learning with summative assessment. However, Wiliam (2011) argues that this is a rather simplistic view since assessment designed to be summative can also be used formatively (p. 38). Paul Black and Dylan Wiliam define formative assessment as “encompassing all those activities undertaken by teachers, and/or by their students, which provide information to be used as feedback to modify the teaching and learning activities in which they are engaged” (Black & Wiliam, 1998, in Wiliam, 2011, p. 37). The KPE3 can be considered such an assessment. Although tests are generally thought to be summative, the teacher’s guide stresses that the test results can be used to further the learning of those pupils scoring in and around the limit of concern.

Feedback is an important part of formative assessment. In order for feedback to be helpful and effective—to promote learning and change—the teacher must not only point out what is wrong, but provide a “recipe for future action” (Wiliam, 2011, p. 121). According to Cameron (2001):

If assessment feedback is to be helpful to learners and improve their learning, it needs to be specific and detailed enough to make a difference and, equally importantly, it needs to be related to a target performance or understanding towards which the learner can move (p. 238).

When used as intended, therefore, the KPE3 can provide teachers with a tool to guide weak pupils toward further learning.

3.3 The KPE3 Construct

As discussed in chapter 2, the term construct refers to the skills being operationalized in the test. In the case of the KPE3, the construct is actually the curriculum objectives for English reading and listening at the end of year 2:

- to listen and understand simple instructions given in English
- to understand and use some common words and phrases relating to the pupils' immediate environment
- to recognize and use some words, expressions and simple sentences in spoken and written texts
- to use the language through several senses and media (Utdanningsdirektoratet, 2013)

The operationalization of the KPE3 construct is evident in the test content/tasks, as was demonstrated in chapter 2. In addition, the construct frequently underlies the scoring system used. This is most clearly seen in tests of productive skills, where a set of descriptors may be used to represent the way an ability, such as writing or speaking, is defined (Luoma, 2004; Weigle, 2002).

In tests of reading and listening, the score is often given as a sum of all points collected, as is the case in KPE3. However, the operationalized construct is not entirely forgotten here. The teacher's guidelines inform teachers what the various tasks are testing, so that they can interpret a right or wrong answer on any item in terms of the 'subskill' being tested.

3.4 Test validity

Hughes (2003) sums up test validity as follows: "[...] a test is said to be valid if it measures accurately what it is intended to measure" (Hughes, 2003, p. 26). Henning (1987) adds that the preposition *for* should follow the term valid when used to describe a test: "Any test may be valid for some purposes, but no for others" (Henning, 1987, p. 89).

Traditionally, test validity has been categorized according to a variety of aspects—some sources cite as many as sixteen. Moreover, the same terms are often used with slightly different groupings when categorizing validity. Alderson, Clapham, & Wall (1995) have categorized validity according to three main types: internal, external, and construct validity, with subtypes for each category.

Internal validity has the following subtypes: face, content, and response validity. Face validity essentially involves “an intuitive judgment about the test’s content by people whose judgment is not necessarily ‘expert’” (p.172). It is therefore considered by many as superficial and unscientific because it includes the views of school administrators, teachers, students, test-developers – or anybody else who is involved with a test’s use. Secondly, content validity pertains to what degree the content of the test complies with the test specifications. For example, if a specific curriculum goal is being tested, this should be reflected in the test items. Finally, response validity refers to the extent to which the thought processes and actions of the test taker demonstrate that they understand the test construct in the same way as it is defined by the test developers.

External validity includes two subtypes: concurrent and predictive validity. Concurrent validity is related to the level of agreement between the scores achieved by the same person on two tests of the same ability taken at around the same time. Predictive validity refers to the degree to which the test score can indicate the test taker’s future level of ability (Carlsen, 2007).

Construct validity refers to what Gronlund (1985) defines as “How well test performance can be interpreted as a meaningful measure of some characteristic or quality”(p. 58). In recent years, validity has been divided into three slightly different categories: criterion-related, content-related, and construct-related (Carlsen, 2007).

Criterion-related validity refers to whether the test results (score/grade) agree with other dependable assessments of the test taker’s skill in a particular area. This other assessment “is thus the criterion measure against which the test is validated” (Hughes, 2003, p. 27). In a classroom setting, this could mean that a teacher should take other assessments into consideration when looking at a pupil’s test score. There are two subtypes to criterion-related validity: predictive and concurrent validity. These have been defined above as subtypes of

external validity.

Content-related validity is related to the degree to which the test tasks represent the skill being tested. A test is said to be valid if it offers evidence that the content of the test includes a representative sample of the particular skills and structures that it is meant to test. It must also take into consideration the level of the test taker (Hughes, 2003). For example, a vocabulary test for beginners learning a foreign language should not contain advanced vocabulary meant for native speakers.

Construct-related validity, as previously defined, relates to the theory behind the skills being tested. In the case of the KPE3, this would be the curriculum objectives of listening and reading in English as a foreign language.

Messick (1995) was critical of these ways of categorizing the aspects of validity because it failed to take into account the the value implications and social consequences of test score use when evaluating validity. He proposed a unified concept of validity in which the term *construct validity* could be used to describe the overarching notion, encompassing content and criterion-related validity. He maintains that these issues are interrelated, constituting fundamental aspects of a more comprehensive theory of construct validity, one that addresses both score meaning in test interpretation and social values associated with test use. Thus, unified validity integrates considerations of content, criteria, and consequences into a construct framework for the empirical testing of rational hypotheses about score meaning and theoretically relevant relationships, including those of an applied and scientific nature (Messick, 1995, p. 741).

With this unified view in mind, Bachman & Palmer (1996) define construct validity as “the meaningfulness and appropriateness of the interpretations that we make on the basis of test scores” (Bachman & Palmer, 1996, p. 21). This means that there needs to be evidence to support and justify the validity of a test score.

Messick (1995) identifies two threats to construct validity: construct underrepresentation and construct-irrelevant variance. Construct underrepresentation means that the assessment is too narrow, and construct-irrelevant variance means the assessment is too broad (p.742). For example, threats to the construct validity of the KPE3 could occur if the operations in the

curriculum objectives were not adequately covered, or if other skills or abilities in addition to those in the curriculum were included. A case in point: the listening section of the KPE3 may be considered to be testing reading since simplified written instructions are provided, and therefore too broad, if the test taker was dependent upon reading written instructions in order to do the tasks. In this case, the possibility of pupils depending on these could be minimized by using the practice material, so that pupils were familiar with the instructions before taking the test. However, validity extends far beyond the issue of test content, as is discussed in the following section.

3.4.1 Messick's six aspects of validity

Messick identifies six aspects of construct validity: content, substantive, structural, generalizability, external, and consequential

Messick maintains that the content aspect of construct validity shows “how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn” (Messick, 1989, p. 16).

Messick (1989) quotes Loevinger's (1957) definition of the substantive component of construct validity as the “extent to which the content of the items included in the test can be accounted for in terms of the trait believed to be measured and the context of the measurement” (Messick, 1989, p.43). In other words, the test takers go through processes representative of processes they would normally go through in the target language (Hasselgreen, 2004).

The structural aspect of construct validity considers whether the scores are presented in a way that reflects the skill being tested, for example in a scale, and how this skill is constituted.

The generalizable aspect of construct validity is concerned that the test score actually reflects the test taker's ability in a certain area. The test taker should receive the same score if s/he were to take the test again or be assessed in a non-test situation on the same ability. One should be able to rely on the test score.

The external aspect of construct validity highlights the need for correlation between the test scores and other empirical evidence (Messick, 1995). Thus, the scores should reflect what other assessments of the skill(s) reveal. The score of an externally valid test should not come as a surprise to the teacher or test taker.

According to Messick, "The consequential aspect of construct validity includes evidence and rationales for evaluating the intended and unintended consequences of score interpretation and use in both the short- and long-term" (Messick, 1995, p. 746). By implication, the test results should be used *only* as intended.

Summary

In this chapter other studies related to the diagnostic tests (kartleggingsprøver) used in Norway were considered. Thereafter, some principles relating to central issues in this study such as formative assessment, test construct, and test validity were presented and discussed. The discussion concluded with Messick's six aspects of validity, which constitutes the basis for the theoretical discussion in the next chapter.

4. Examination of the KPE3 in the light of Messick's six aspects of validity

This chapter considers validity issues related to KPE3, by addressing the first two research questions:

1) What aspects of the KPE3 potentially make its validity vulnerable?

2) Which of these aspects are in the hands of the teachers/schools?

Messick's six aspects of validity provide a useful framework for investigating which potential 'threats' to validity exist. The approach builds on, and to some extent replicates, the approach used by Hasselgreen (2004) in an *a priori* examination of an oral test of English for young learners; i.e. the test was considered as it was made by the test developers, not taking into account how it was used.

For each aspect of validity, the KPE3 is evaluated in relation to the list of potential threats identified by Hasselgreen, although not all are relevant for the current study. In the few cases where the threats are not considered relevant to tests of reading and listening, these have been removed. The listed threats that are considered to be the responsibility of test developers are not investigated further in this study. Corresponding threats to those listed, which are in the hands of teachers/schools, can then be identified and expanded upon.

The chapter concludes with an overview of areas where validity appears to be threatened in the case of KPE3, focusing on those aspects which are in the hands of teachers and schools, and this is summed up in a table.

4.1 Identifying threats to Messick's six aspects of validity and how they relate to the KPE3 (2014)

In Hasselgreen's (2004) study on testing the spoken English of young Norwegian learners, she identified potential threats to Messick's six aspects of validity. Some of these threats implicitly or explicitly involve test administration by teachers.

4.1.1 Content aspect

Content validity implies that the test items are representative of tasks that elicit evidence of the language ability being tested.

Possible threats to content validity include:

- Faulty or incomplete operationalization of components of the model of CLA (communicative language ability)
- Poor sampling of the language associated with the underlying theoretical model and domain of CLA, when making tasks
- Unclear instruction or unfamiliarity of format (which prevent tasks from being done as intended)
- Test methods and procedures that may prevent testees from performing in the way intended
- Test bias associated with cultural background, background knowledge, native language, ethnicity, age, or gender (Hasselgreen, 2004, p. 30).

Corresponding threats to content validity that are in the hands of teachers include the point related to "unclear instructions or unfamiliarity of format." In this case, a comparable threat in a test such as the KPE3 would be teachers 'failure to use practice material,' since this would mean that the task types were unfamiliar to the pupils.

Another teacher-related threat to content validity is posed in relation to "test methods and procedure that may prevent testees from performing in the way intended"; teaching test content may have had influence on test performance. The fact that the KPE3 was on paper and unchanged from 2010-2015, made it possible for teachers to become familiar with the actual content of the test. The fact that the test booklets could easily be kept from year to year could have influenced how teachers prepared their pupils for the test. This constitutes a potential threat to the test's content validity.

Thus, content validity could be threatened by teachers' failure to use the practice tests and by their use of the test itself or the vocabulary tested to prepare the pupils.

4.1.2 Substantive aspect

Substantive validity is dependent upon pupils going through the same processes when they read or listen on the test that they would normally go through when reading or listening in the target language, for example in a familiar classroom situation.

Possible threats to substantive validity include:

- Tasks that do not fully engage the testees in the processes associated with the underlying theoretical model and domain of CLA
- Tasks that essentially draw on processes that are irrelevant to the underlying theoretical model of CLA
- Tasks that do not enable the testees to actively engage their language ability in a reasonably authentic way
- Tasks that are uninspiring or off-putting and so fail to engage the testee in real communication (Hasselgreen, 2004, p. 30)

In order for a test to show evidence of substantive validity, the testing conditions need to be familiar to the pupils. Test administrators for the KPE3 were in most cases the class teachers. The teacher's guide clearly outlined what teachers were to do in order to prepare their pupils for the test and familiarize them with the testing situation. If the pupils had not been introduced to the different types of test items, or had not been prepared for what the testing situation would be like and how much time they had for each section, this would affect their performance so the test results might not give an accurate indication of what the pupils were actually capable of achieving.

Thus, substantive validity can be threatened by teachers' failure to be fully prepared themselves, with the help of the teacher's guide, as well as their failure to prepare their pupils adequately, including the use of practice material.

4.1.3 Structural aspect

Structural validity is related to the need for the scoring system to fully reflect the abilities being tested.

Possible threats to structural validity include:

- Scoring procedures that do not fully reflect the specified model and domain of CLA
- Clustering and division of constructs in the scoring system that is not supported by primary or secondary empirical evidence (Hasselgreen, 2004, p. 31)

In the KPE3, the scoring procedure was intended to provide the basis for deciding whether a pupil appears to be on or below the “limit of concern” in either reading or listening sections of the test. The scoring of the test was designed to reveal a minimal level of competence in each skill, and not to provide scores that actually reflect the abilities of pupils beyond this level. Thus, structural validity could be threatened if the scores were interpreted as a reflection of the abilities of all pupils.

4.1.4 Generalizability aspect

Generalizability validity implies that the test scores have to be consistent and reliable. This can be affected by the way the test is carried out, as well as the way it is scored.

Possible threats to generalizability validity include:

- Methods and procedures for testing that are unclear or weakly defined, so that the inconsistencies may occur in the way the test is carried out
- Scales or other scoring instruments that are couched in vague terms (and so give rise to different rater interpretations)
- Instructions and procedures for scoring that are unclear or weakly defined
- [...] lack of content, substantive, structural and external validity (Hasselgreen, 2004, p. 31).

Although the scoring instruments and instructions were clear in the KPE3 guidelines, marking errors occurred, which affected the generalizable validity of the test. If the scores are unreliable, the results cannot be trusted. Generalizable validity means that a pupil should get the same mark if s/he were to take the test again or a similar test with a similar test construct. However, if the test is marked incorrectly, the score cannot be relied upon and this has a negative impact on the validity of the test. In the KPE3, this was especially crucial with regard to pupils who score on and around the limit of concern since the test was intended to identify this vulnerable group. In the worst case scenario, marking errors could place a pupil over the limit of concern, and the teacher might not realize that a pupil needed extra support in reading and/or listening.

Thus, generalizable validity can be threatened by incorrect marking.

4.1.5 External aspect

In order to ensure external validity, the tests score should reflect the skills the test taker has, which are also measured using other tests/tasks. There needs to be a correlation between the test scores and other empirical evidence.

Possible threats to external validity include:

- Using external criteria that measure different abilities from the test in question
- Failing to look for discriminant evidence to ensure that the test is not measuring unrelated abilities
- Using external criteria whose validity is unknown (Hasselgreen, 2004, p. 31)

The teacher's guide for the KPE3 instructed teachers to consider the test score in light of other assessments measuring the same ability and not to trust the test result blindly. For external validity to be ensured, the teachers should have been able to compare the outcome of the test based on other assessments they have carried out. This was dependent on their competence in assessing language ability. Thus, external validity can be threatened by teachers' lack of training in the area of language assessment.

4.1.6 Consequential aspect

Consequential validity is only secured if the KPE3 is used in accordance with the purpose for which it was made. According to Bachman (2005), "The single most important consideration in both the development of language tests and the interpretations or their results is the purpose or purposes which the particular tests are intended to serve" (Bachman, 2005, p. 2).

Possible threats to consequential validity include:

- Test tasks and methods that draw on irrelevant abilities
- Scoring procedures that do not encourage the learner to assess his/her own performance
- Lack of any analytic feedback on individual strengths and weaknesses
- Unclear instructions to users on how (and how not) to interpret test results

- Failure to restrict inferences, made from test results, to what the testee can do, in the content domain specified as well as lack of content, substantive and structural validity (Hasselgreen, 2004, p. 31).

The responsibility for using the KPE3 as intended, which is implicit in points three and five above, does not fully lie with the teachers, but also with school leaders and school owners, since it is often their decision to administer the test at their school. The test was intended to identify the weakest pupils, who needed extra support in English reading and/or listening; it does not discriminate between the average and stronger pupils. A threat to consequential validity would be posed if the KPE3 results had been used to assess and give feedback on the levels of all of the pupils in the class. In addition, failure to give extra support to the pupils identified as being in need could jeopardize consequential validity.

Thus consequential validity can be threatened by misusing the test results. It can also be threatened by teachers' inability to give necessary support to weak pupils, which in turn is dependent on teachers being properly qualified.

4.2 Summary of the potential threats to validity at the hands of teachers/schools

The following table highlights the potential threats to validity identified in relation to the KPE3, specifically those that are the responsibility of teachers/schools, employing Messick's six aspects of validity and Hasselgreen's threat analysis.

Table 1: Aspects of validity and threats in the hands of teachers/schools

Aspect of validity	Threat posed by
Content	<ul style="list-style-type: none"> • failure to use practice tests • using the test itself, or material taken from this, to prepare the pupils
Substantive	<ul style="list-style-type: none"> • teachers failing to be fully prepared themselves with the help of the teacher's guide • failing to prepare their pupils adequately, including the use of practice material
Structural	<ul style="list-style-type: none"> • interpreting the scores as reflecting the abilities of all pupils
Generalizable	<ul style="list-style-type: none"> • incorrect marking
External	<ul style="list-style-type: none"> • teachers' lack of training in the area of language assessment

Consequential	<ul style="list-style-type: none">• misusing the test results• teachers' inability to give necessary support to weak pupils due to lack of training in language pedagogy
---------------	-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Summary

In this chapter, the KPE3 was examined in the light of Messick's six aspects of validity, and the framework developed by Hasselgreen (2004) was employed to identify potential 'threats' to validity. The focus has been on potential threats to the validity of the KPE3 in areas that are in the hands of the teachers/schools. These threats have been summed up and presented in Table 1. Thus, research questions 1 and 2 have been addressed.

5. Methodology

In chapter four, possible threats to the validity of the KPE3 were identified, in particular those that are in the hands of the teachers/schools. These were summarized in Table 1.

Having identified features that may have threatened the KPE3's validity in areas where the responsibility lay with the teachers, only the third research question remains to be answered:

3) Is there evidence to indicate that some of the aspects related to teachers' practices may actually be undermining the test's validity?

The study that is presented here investigates this question, based on information collected from teachers who had administered the test in 2014 and from pupil test booklets.

In this chapter, the methodology chosen for this study is presented and the grounds for this choice are discussed. Data were collected from two sources in this study: a questionnaire completed by teachers and pupils' test booklets. Given the focus on teachers' practices, a questionnaire was considered a suitable data-collection tool for eliciting information from teachers, and the pupils' test booklets served to provide data regarding the marking errors made by teachers. These two data collection methods, and the related data analysis methods, are introduced and discussed in separate sections. In the questionnaire section, the questions have been grouped into three categories, according to topic. The first block of questions focuses on background information about the teachers is presented, and the second block is directly related to the threats identified above. Finally, although not directly related to the proposed threats, the third block of questions sheds light on the teachers' perception and use of the test. In the test booklet section, the focus is on the process by which the data regarding marking errors was collected, analyzed, and recorded.

5.1 Choice of methodology

In order to reveal the significance of all of the threats to validity that had been identified, it was necessary to employ two data collection methods: questionnaire and completed test

booklets. The choice was made, therefore, on the basis of the information required to answer the research question.

In order to obtain information from as many teachers as possible, the decision was made in 2014 to pursue a quantitative approach and develop a questionnaire. A questionnaire can be defined as “any written instruments that present respondents with a series of questions or statements to which they are to react either by writing out their answers or selection them among existing answers” (Brown, 2001, p. 6). This method was considered suitable because it was a good way of obtaining a larger amount of data in a relatively short time. The method has its origin in the social sciences and the basic idea behind survey research is “the recognition that the characteristics, opinions, attitudes, and intended behaviors of a large population can be described and analyzed on the basis of questioning only a fraction of the particular population” (Dörnyei & Csizér, 2012, p. 74). According to Dörnyei & Csizér (2012), this method is a good way of gaining insight into teachers’ feelings, opinions, and prior knowledge, as well as collecting some background information and biodata. Since it is the aim of this study to identify teachers' practices, the questionnaire was considered a useful tool.

In addition to the questionnaire, KPE3 test booklets from pupils who had completed the test were collected in order to investigate whether or not the teachers were making marking errors. Class sets of test booklets were collected from schools that had administered the test. Like the questionnaire, this type of research can also be described as quantitative.

5.2 Sample selection

The Directorate of Education was contacted to obtain a list of all the schools that had administered the KPE3 in 2014. From this list, schools across the country were randomly selected and asked to submit the test booklets needed. In order to obtain as diverse a sample as possible, schools all over the country, in both urban and rural areas, were contacted, as well as schools of various sizes. The test booklets were collected experimentally, in the sense that the participants were randomly selected. The data collected was quantitative, and the results were derived from statistical analysis.

In addition to the booklets, some teachers were contacted and asked to complete the questionnaire. However, it was not possible to obtain responses from a representative sample, which means that it is difficult to draw definite conclusions about the country as a whole. Nevertheless, the sample was sufficient to provide some tentative answers to my research question, giving an indication of the validity issues threatening KPE3 in 2014.

Between May 15 and June 11, 2014, eighty-six schools were contacted to request their pupil test booklets, and, at the same time, the opportunity was taken to ask whether third-grade English teachers at these schools would be willing to answer the questionnaire. Schools were contacted by telephone. After receiving the email addresses of each school's third-grade teacher(s) in the schools that had agreed, an email with the questionnaire attached was sent out. The email informed the teachers that the study was anonymous and that no information would be able to be traced back to themselves or their school. Each questionnaire was given a number so as to protect the privacy of the teacher and school.

NSD, the Norwegian Centre for Research Data, was contacted, but there was no need to submit a notification form as long as all correspondence with the teachers was destroyed. All emails were deleted and questionnaires destroyed upon completion of the study. The decision was made to anonymize both the teachers and the school since their identity was of no relevance to the study. It was also assumed that more teachers would be comfortable answering the questionnaire when their anonymity could be guaranteed.

Thirty-five schools agreed to be part of the study and sent their test booklets by post, resulting in a 40.7% return rate. In total, 910 booklets were submitted. Some schools wanted the booklets to be returned after the study was complete, and the remainders were destroyed.

From the thirty-five schools that submitted the test booklets, eighteen schools refrained from answering the questionnaire, while seventeen schools responded positively. Three schools chose only to answer the questionnaire, but not submit their test booklets, resulting in a 23.3% return rate for the questionnaire. One school sent five questionnaires, resulting in a total of twenty-four questionnaires from schools in nine different counties.

The data collection information is summed up in the Table 2.

Table 2: Summary of data collection

Only questionnaires	Questionnaires + test booklets	Only test booklets
3 schools	17 schools	18 schools
3 questionnaires	21 questionnaires + 409 test booklets	501 test booklets
TOTAL:	24 questionnaires	910 test booklets

5.3 The questionnaire

The questionnaire included two types of questions: closed-item questions and open-ended items. A closed-item question is one in which the researcher provides the participants with possible answers and all they need to do is tick a box or circle an answer. According to Mackey and Gass (2015), “Closed-item questions typically involve a greater uniformity of measurement and therefore greater reliability. They also lead to answers that can be easily quantified and analyzed” (p. 93). In contrast, open-ended items are questions that the participants can respond to freely, writing their own thoughts and ideas; this often results in more unexpected and insightful data. (p. 93)

The questionnaire of this study was inspired by an unpublished questionnaire that was developed by colleagues at the University of Bergen in 2012 to elicit information from English teachers about the English skills of young learners. The aim of the questionnaire was to collect data regarding the grade-five National Test in English. The questionnaire was adapted to ensure its relevance for this study. A 1981 survey (Yeh, Herman, & Rudner, 1981) related to teachers and test use, in which teachers were asked their thoughts and opinions on testing, also influenced the wording of questions in the questionnaire. The questionnaire can be found in its entirety in Appendix C.

Great care was taken to follow the five key strategies for producing items that function well, as described in Dörnyei & Csizér (2012). In accordance with the first strategy, the items were kept short and simple. Second, an effort was made to use simple and natural language. Moreover, since it is thought that “the quality of the obtained data improves if the questionnaire is presented in the respondents’ own mother tongue” (Dörnyei & Csizér, p. 79), the questionnaire was written in Norwegian. This eliminated language as a factor that might

discourage potential participants from answering the questionnaire. Lastly, ambiguous and loaded words and sentences were avoided, as well as negative constructions and double-barreled questions (Dörnyei & Csizér, p.78).

5.3.1 The questions

The questionnaire consisted of eighteen questions and a wide variety of question types were used in the questionnaire. The first eight questions were closed-item multiple-choice questions, in which the participants were given the option of ticking one or more boxes. Questions 1-3, 5 and 7 included an “other” option that was open, allowing the participants to write their own answer if they felt that the given options were inadequate. Questions 4, 6, and 8 did not have this option since the wording of the questions did not require it. Question nine consisted of a list of eight statements about the KPE3 that the participants were asked to respond to on a Likert scale. A Likert scale “consists of a characteristic statement accompanied by five or six response options” (Dörnyei & Csizér, p. 76). This allowed the participants to indicate the extent to which they agreed or disagreed with the statements. The scale was as follows: strongly disagree (1), generally disagree (2), somewhat disagree (3), somewhat agree (4), generally agree (5), strongly agree (6). Question 18, the very last question on the questionnaire, was an open-ended question that allowed the participants to provide any other information they considered relevant or to comment upon the KPE3 or the questionnaire.

Block 1 consisted of questions 10-13, which were biodata questions. Question 10 was the only open-ended question in this section requiring the participants to write the answer. The remaining questions were all closed-item multiple-choice. In this block, the questions focused on work experience, education, whether they had lived in an English speaking country, if they had any formal English education and whether or not they had chosen to teach English themselves. The information was intended to shed some light on their attitude towards English as a subject in general and to determine whether they had any formal training in language assessment (Dörnyei & Csizér, 75).

To ensure that the questionnaire was eliciting the information it was meant to elicit (Mackey & Gass, 2015), the questionnaire was piloted on five teachers. Taking their feedback into consideration, a few minor adjustments were made regarding the wording of the questions

The three topic categories identified in the introduction to this chapter are considered in greater detail in the following sections. Block 1 consisted of questions 10-13 and was included to provide some general background information about the participants. The questions in block 2 were directly linked to the research question (questions 1-3, 5, 8, 14-15), while block 3 (questions 4, 6-7, 9, 16-18) was not, but gave insight into how they used the test.

5.3.1.1 Block 1 questions: Biodata

Questions 10-17 collected biodata in order to gain insight into the background of the teachers. However, two questions (14 and 15) asked the teachers about their formal education in English and language assessment, and are therefore directly linked to the research question. Question 16 was included in block 3. The other questions included in block 1 gathered the following information: question 10 asked which county the participants worked in; question 11 asked their gender; question 12 asked the duration of their work experience as teachers; and question 13 asked whether they had lived in an English speaking country. These questions were specifically chosen since they elicited data regarding factors that had the potential to influence the findings.

5.3.1.2 Block 2 questions: Directly linked to the research question

Clearly, the key questions in the survey are those that can provide data regarding potential threats to the KPE3's validity, and thereby address the research question directly. As identified in the previous chapter, questions 1-3, 5, 8, 14-15 on the questionnaire fall into this category. These are considered in some detail.

The first question on the questionnaire was: *What is important to you when preparing for the test?* This is a closed-item question in which the following five tasks were specified:

- Read the teacher's guide
- Make practical arrangements
- Do the practice test with pupils

- Do the practice test alone
- Inform parents

The participants were able to tick off more than one option. The answers given to this question provide evidence related to threats 1 (*failure to use practice tests*), 3 (*teachers failing to be fully prepared with the help of the teacher's guide*), and 4 (*failing to prepare their pupils adequately, including the use of practice material*). The phrasing of the first question regarding test preparation, the term “what is important to you” was chosen rather than simply asking “did you do the following” because it was feared that no teacher would admit to not following instructions.

Questions two and three on the questionnaire were also closed-item questions. They were similar in the sense that the teachers were given the same options to choose from. Question two asked the teachers their opinion on the purpose of the KPE3, while question three asked them how they believed the tests were being used. The options they could choose from were:

- Find the weakest pupils and give extra attention
- Find the level of the entire class
- Give feedback to pupils
- Give feedback to parents
- Give feedback to principal
- Give feedback to school owners

These options check awareness of the intended function of the test and also correspond to potential threats 5 (*interpreting the scores as reflecting the abilities of all pupils*) and 8 (*misusing test results*). Given the intended function of the KPE3, only answers *find the weakest pupils and give extra attention* and *give feedback to pupils* would be correct. Feedback to parents, principals and school owners would only be relevant in cases where pupils score at or below the level of concern.

Question five was also directly linked to the potential threats to validity. Teachers were asked whether the test influenced their teaching. This was a closed-item question, and they were given the following options:

- No
- Yes, I adjust the teaching for the weak students
- Yes, we practice words
- Other

Under “other” they had the opportunity to add points of their own, so it was not entirely closed. Point 2 in this question was linked to the test aim, and point 3 to potential threat 2 (*using the test to prepare the pupils*).

Question eight was linked to potential threat 6 (*incorrect marking*). Teachers were asked if they used the answer key when marking the test. While not all teachers who do not use the key make mistakes, those who do make mistakes are definitely not using the key.

This closed-item question had four options for the teachers to choose from:

- Always
- Usually
- Some of the time
- Never

The last questions that were directly linked to the potential threats were questions 14 and 15 in which teachers were asked about their education in English and in language assessment. Question 14 asked whether teachers had a college or university degree in English, and if so, how many credits. Question 15 asked if they had any formal training in language assessment.

These were closed-item questions, offering the following options:

- None
- In-service course
- Included in degree
- Academic course with exam and credits

However, question 14 allowed teachers to write how many credits they had obtained. These questions corresponded to potential threats number 7 (*teachers’ lack of training in the area of language assessment*) and 9 (*teachers’ inability to give necessary support to weak pupils due to lack of training in language pedagogy*).

Table 3 summarizes the questionnaire questions and the corresponding potential threats to test validity.

Table 3: Summary of questionnaire questions linked to potential threats to KPE3’s validity

Potential Threat to KPE3 Validity	Corresponding Questionnaire Question
1. failure to use practice tests	1
2. using the test itself to prepare the pupils	5
3. teachers failing to be fully prepared themselves with the help of the teacher’s guide	1
4. failing to prepare their pupils adequately, including the use of practice material	1
5. interpreting the scores as reflecting the abilities of all pupils	2, 3
6. incorrect marking	8
7. teachers’ lack of training in the area of language assessment	15
8. misusing the test results	2, 3
9. teachers’ inability to give necessary support to weak pupils due to lack of training in language pedagogy	14

5.3.1.3 Block 3 questions: Indirectly linked to research question

Questions 4, 6, 7, 9, 16, and 17 on the questionnaire were all closed-item questions that were indirectly linked to the research question, and included in order to shed light on the teachers’ attitudes and practices with regard to the KPE3 and how they used the test, issues that will be returned to in the discussion chapter of this thesis.

Question four asked who decided that the KPE3 was to be administered at their school, and question six followed up by asking whether the teachers felt that it was necessary to have an English diagnostic test in grade three. Question seven asked who corrected the KPE3 papers.

In question nine, a Linkert scale with six options was used. The teachers were given the following eight statements and asked to indicate the degree to which they agreed with the statements or not:

- The teacher’s guide gives adequate information.
- My school focuses a lot on KPE3.
- It is difficult to correct KPE3.
- The results of KPE3 reflect the reading skills of my pupils.
- The results of KPE3 reflect the listening skills of my pupils.

- The results of the KPE3 reflect the general English competence of my pupils.
- KPE3 is a good assessment tool of my pupils.
- It is difficult to follow up the results.

Questions 16 and 17 were close-item questions that were included in order to gain insight into the teachers' attitude to teaching English. They were asked whether they themselves had chosen to teach English (question 16) and whether they enjoyed teaching English (question 17).

5.3.2 Analysis of the questionnaire data

The answers to each of the questionnaire items were plotted onto an excel spreadsheet which was used to develop the tables that are presented in the next chapter. Each respondent's questionnaire was coded, assigned a number from 1-24 and listed vertically on the spreadsheet. The questionnaire items were listed horizontally. Starting at number 1 (participant number 1) the answers for the multiple-choice items were plotted by giving each tick the number "1". If a participant failed to answer a questionnaire item, it was given the number "9". For question number 9, the Likert scale from 1-6, the number each participant wrote for each statement was recorded on the spreadsheet. Upon completing the spreadsheet, the answers for each questionnaire item were summed up and converted it into a percentage.

5.4 Analysis of the test booklets

Every year since the KPE3 was first administered in 2010, data has been collected from the test in order to ensure the test's reliability from year to year. A sample of completed test booklets was collected from selected schools that had administered the test. From 2010-2105, the test was paper-based, and each year test booklets from approximately 500 pupils were collected. What each pupil had answered for each question was plotted in an excel spreadsheet and statistics were run to see whether the test seemed to get easier for each passing year. In 2014, the year this study was conducted, thirty-five schools sent in their pupils' test booklets, 910 booklets in total.

To ensure anonymity, each school was asked to remove the pupil's names from the booklets, but to note the pupil's gender on the cover. Each school was given a code number, and each

pupil was recorded as being boy or girl plus a number, for example boy1 or girl 4. For each pupil, individual answers were recorded in an excel spreadsheet. This served two purposes: it facilitated the analysis of pupils' answers and enabled researchers to compare pupils' answers from year to year. However, this was not relevant to the current study.

For the purposes of this study, pupils' answers were recorded to determine whether there were any discrepancies between their answers and the teachers' marking. Each pupil's answers to the different tasks were recorded in the spreadsheet. If a pupil had answered incorrectly but had not been deducted a point by the teacher, this was marked in a red color. The same was done if a pupil had answered correctly but had deducted a point by the teacher. This way it was easy to see how many marking errors there were per pupil and per teacher.

Summary

In this chapter the data collection and analysis methods used in this study have been presented. A quantitative research design was employed to investigate whether there is evidence to indicate that, in 2014, some teachers' practices might have been impairing the KPE3's validity. A questionnaire was designed and sent out to schools across the country. The content of the questionnaire was divided into three topic-based sections: one section included the block of questions that provided background information about the teachers; the second block of questions was directly linked to the research question regarding teachers' impact on test validity; and the third section contained a block of questions that were indirectly linked to the research question. Test booklets were also collected in order to identify marking errors.

6. Findings

The findings with regard to the third research question from the study are of central importance to this study: Is there evidence to indicate that teachers' practices are impairing the validity of the KPE3? An analysis is undertaken of the data obtained from the various question blocks in the questionnaire and that obtained from the test booklets.

6.1 Findings from the questionnaire

The data from the questionnaire has been divided into three sections, as in the chapter 5: the first section considers the findings related to the teachers' biodata; the second presents the findings directly linked to the third research question; and the third examines the indirectly-related findings that shed light on teachers' attitudes to the KPE3.

6.1.1 Block 1: Biodata

The findings for questions 10-13 provide data regarding the teachers' backgrounds that is not directly linked to the research question. These findings are summed up in this section, while the findings for the two biodata questions (14 & 15) that were particularly relevant to the third research question are discussed in the next section. It should be noted that one participant chose not to answer any of the biodata questions. The results are therefore based on 23 participants rather than 24.

Data was received from every region in Norway, except Southern Norway. Table 4 sums up the results from question 10, showing the number of participants from each region.

Table 4: Question 10 - Participants by geographical region

Region	Participants
Western Norway	8
Central Norway	5
North Norway	3
East Norway	7

Question 11 asked about the participants' gender. Nineteen women and four men participated in answering the questionnaire. One participant did not disclose his or her sex. This meant that 82.6% of the participants were women.

The teachers had varied teaching experience. Over 50% of the participants had more than 10 years teaching experience, while fewer than 25 % had less than 5 years. Responses to question 12 regarding the teachers' work experience are summarized in Table 5 below.

Table 5: Question 12 - Work experience

Work Experience	Participants
Less than 2 years	3
2-5 years	2
6-10 years	6
More than 10 years	12

While the majority of the participants had never lived in an English speaking country, 17.4% had done so for a period of 1-4 years. 8.7% had lived in an English speaking country for 7 or more years. Responses to question 13 are summarized in Table 6.

Table 6: Question 13 - Residence in an English-speaking country

Lived in English-speaking country	Participants
Never	16
6 months	1
1-4 years	4
7-10 years	2

6.1.2 Block 2: Findings linked to research question

The results from questions 1-3, 5, 8, 14-15 are directly related to the research question: **Is there evidence to give an indication that some of these aspects may actually be impairing the test's validity, due to some teachers' practices?**

The first question on the questionnaire asked teachers what they considered to be important when preparing for the KPE3. 100% of the teachers answered that they felt it was important to read the teacher's guide. The teacher's guide strongly advised teachers to do the practice test together with the pupils before administering the test (Utdanningsdirektoratet, 2010, p. 2), and yet 29.2% did not consider this to be important. In addition, approximately half of the teachers stated that preparing for the test by doing the practice test themselves was not a priority. These findings provide evidence that indicates that the test's validity may be compromised by the teachers' practices. This corresponds to potential threats 1 and 4 discussed in chapter four: *failure to use practice tests*, and *failure to prepare their pupils adequately, including the use of practice material*.

The results from question 1 are summarized in Table 7 below.

Table 7: Question 1 - Priority when preparing for test

1. What is important to you when preparing for the test?	
Read the teacher's guide	100 %
Make practical arrangements	70.8%
Do the practice test with pupils	70.8%
Do the practice test alone	54.2%
Inform parents	41.7%

Questions 2 asked the teachers their opinion about the purpose of the test. All but one teacher understood that the purpose of KPE3 was to identify the weakest pupils needing special attention in English, but over 60% incorrectly assumed the test could identify the level of the entire class. Furthermore, only 50% understood the purpose of the test was to give feedback to pupils. This is summed up in Table 8.

Table 8: Question 2 - Purpose of test

2. What, in your opinion, is the purpose of the KPE3?	
Find the weakest pupils and give extra attention	95.8%
Find the level of the entire class	62.5%
Give feedback to pupils	50 %
Give feedback to parents	50 %
Give feedback to principal	25 %
Give feedback to school owners	8.3%

For question 3 (Table 9) there was a slight discrepancy between the teachers' understanding of what the purpose of the test was and how it was being implemented. While 87.5 % assumed it was used to identify the pupils who needed extra attention in English, 54.2% still assumed it was used to find the level of the entire class. Slightly more than half stated that the test was used to give feedback to pupils. Interestingly, this result revealed that teachers were aware that the test was being used in a way that was to some extent at odds with its intention, which in itself is a threat to validity.

Table 9: Question 3 - How the test is used

3. How is the test being used?	
Find the weakest pupils and give extra attention	87.5%
Find the level of the entire class	54.2%
Give feedback to pupils	45.8%
Give feedback to parents	54.2%
Give feedback to principal	33.3%
Give feedback to school owners	0

The results from questions 2 and 3 on the questionnaire provided evidence indicating that threats 5 and 8 to validity (*interpreting the scores as reflecting the abilities of all pupils and misusing the test results*) were occurring and this might invalidate the KPE3 test results. The test was not designed to find the levels of the entire class, nor could it provide information to parents and school administration except for pupils who scored on or around the level of concern.

Another potential threat to the validity of the KPE3 involved *using the test itself to prepare the pupils*. Question 5 on the questionnaire asked if teachers let the test influence their teaching. While 50% did not let the KPE3 influence their teaching, 16.7% said they practiced the words on the test together with their pupils. This is a clear violation of the test's validity.

Table 10 shows the results for question 5.

Table 10: Question 5 - Influence of test on teaching

5. Does KPE3 influence your teaching?*	
No	50%
Yes, I adjust the teaching for the weak students	33.3%
Yes, we practice words	16.7%

*Percentages out of 18 because three did not answer this question and three had not done the test before.

Yet another potential threat to the test was marking errors. Question 8 on the questionnaire asked teachers if they used the answer key when marking the test. 70.8% answered that they *always* or *usually* used the answer key when marking the test. A further 25% used the answer key some of the time, while one teacher admitted to never using the answer key. The result from this question is summarized in Table 11. Below is a summary of question 8 on the questionnaire.

Table 11: Question 8 - Use of answer key

8. Did you use the answer key when marking?	
Always	37.5%
Usually	33.3%
Some of the time	25 %
Never*	4.2%

*This school had seven marking errors.

Questions 14 and 15 were included in the questionnaire in order to find out what educational background the teachers had in English and language assessment. This was considered to provide evidence pertaining to threats 7 and 9: *teachers' lack of training in the area of language assessment*, and *teachers' inability to give necessary support to weak pupils due to the lack of training in language pedagogy*.

Two teachers had one year or more of English training, four had six months (30 credits) and one had 15 credits of English. Sixteen participants (69.6%) had no training whatsoever in English. These findings are summarized in Table 12.

Table 12: Question 14 - English education

English education	Participants*
None	16
15 credits	1
30 credits	4
60 or more credits	2

*One person did not answer this question

The results were not much different when it came to formal language assessment training. 65% had no training, while the remaining participants had little or some training, as can be seen in Table 13. Please note that this study cannot make any claims about teachers' ability to follow up their pupils, but can only comment on their training in language assessment or lack thereof.

Table 13: Question 15 - Training in language assessment

Formal language assessment training	Participants*
None	15
course	4
Included in degree	3
Own subject	1

*One person did not answer this question

It is a matter of concern that nearly 70% of the teachers in question did not have formal training in English. But perhaps the fact that only 65% of the teachers in the sample do not have formal training in language assessment is an even greater concern. It raises the question whether they have the competence that is needed to follow up the results of the test. In fact, 54.2% of the twenty-four teachers lacked training in *both* English and language assessment.

The results presented in this section indicate that there is evidence that teachers' practices are impairing the validity of the KPE3. There was the possibility that threats to both the content

and the substantive validity of the test were posed by the fact that 29.2 % of the sample teachers did not consider it important to do the practice test together with their pupils. Also, 16.7 % of the teachers stated that they practiced the words of the test with their pupils, which violated the content validity of the test. With regard to score interpretation, 62.5% of the participating teachers thought that the test results reflected the abilities of all pupils, and 54.2% thought that that was how the test was being used, violated the structural and consequential validity of the test. Lastly, there was evidence to show that teachers lack formal training in English, as well as language assessment; 54% of the teachers had no formal training in either which threaten external and consequential validity.

6.1.3 Block 3: Findings indirectly linked to research question

The questions in the questionnaire that were indirectly linked to the research question have been summarized here in order to gain an understanding of teachers' knowledge and opinions of the KPE3 (2014).

With regard to question 4 that addressed the issue of who had made the decision to administer the KPE3 to their pupils, 25% of the teachers responded that it was their decision; while 75% answered that it was the decision of either the principal or the school owner. Although 54.2 % said the decision to take the test was made by the school owner, none of the participants assumed the test was being used to give feedback to the school owner. It should be noted that more than one option could be indicated here; two teachers checked off both "school owner" and "principal", and three others checked off "principal" and "yourself".

Although the KPE3 is not a mandatory test, some counties have made it so. The high percentage of teachers being instructed to administer the test, combined with the fact that over half of the teachers assumed the test was being used to find the level of the entire class, raises the question as to whether the true purpose of the test being misunderstood – by not only teachers but also the school owner. It is, however, interesting that school owners do not seem to be receiving feedback on the test results, according to the findings from question three above. 54.2 % of the teachers answered that the choice to administer the test was made by the school owner, while 41.7 % stated that the principal made the decision. Only 25 % of the teachers made the choice themselves. It is important to note that in three cases, teachers

checked off more than one box, meaning that in two cases the decision to administer the test was a joint one between principal and teacher, and in one case principal and school owner. The results of question four are summarized in Table 14.

Table 14: Question 4 - Decision to administer the KPE3

4. Who decides that you are to administer the KPE3?	Participants
School owner	13
Principal	10
Yourself	6
Head of department	0

When asked if the KPE3 was a necessary test, 66.6% of the participants felt that it was necessary, while 25% were undecided; 8.3 % believed the test to be unnecessary. This can be seen in Table 15.

Table 15: Question 6 - Necessity of KPE3

6. Do you think it is necessary to take the KPE3?	Participants
Yes	16
I don't know	6
No	2

Question 7 on the questionnaire asked the participants who marked the tests. This was to determine whether someone other than themselves was involved in the marking. 100 % said they were the ones who corrected the tests.

The only questionnaire item using a Likert scale was question 9. The summary of this question is presented in Table 16. All of the participating teachers agreed that the teacher's guide provided adequate information, although only 58.3% strongly agreed with this statement. 54.1 % disagreed with the statement that their school focused too much on the KPE3, while 45.9% thought their school did so. The vast majority did not feel it was difficult to correct the test; the one teacher who did consider it difficult was the same teacher who never used the answer key when marking.

The answers to question 9 also revealed that 41.7% of the participants generally agreed that the KPE3 reflected their pupils listening and reading skills, while 33.3% agreed that the test reflected the general English competence of their pupils. There was only a 4.3% difference between the number of participants who agreed and those who disagreed with the statement that the test is a good assessment tool of their pupils, with those who agreed having a slight majority. Based on the data collected, it is not possible to explain why nearly half the teachers surveyed do not believe the KPE3 is a good assessment tool. It is worth noting that, although only 4.3% strongly agree that the KPE3 is a good assessment tool, 66.7% stated that they thought the test was necessary. One possible explanation for this discrepancy could be the fact that teachers see the need for a test, but perhaps what they need is a test that can discriminate on all levels.

Table 16: Question 9 - Teachers' opinions of KPE3

9.	Strongly disagree	Generally disagree	Partly disagree	Partly agree	Generally agree	Strongly agree
	1	2	3	4	5	6
A: The teacher's guide gives adequate information				12.5%	29.2%	58.3%
B: My school focuses a lot on KPE3	20.8%	20.8%	12.5%	29.2%	12.5%	4.2%
C: It is difficult to correct KPE3	66.7%	16.7%	8.3%		4.2%	4.2%
D: The results of KPE3 reflects the reading skills of my pupils		8.3%	29.2%	16.7%	41.7%	4.2%
E: The results of KPE3 reflects the listening skills of my pupils		4.2%	33.3%	12.5%	41.7%	8.3%
F: The results of the KPE3 reflect the general English competence of my pupils		12.5%	29.2%	20.8%	33.3%	4.2%
G: KPE3 is a good assessment tool of my pupils *		13 %	30.4%	13%	39.1%	4.3%
H: It is difficult to follow up the results**	9.1%	36.4%	22.7%	27.3%		4.5%

* One person did not answer this question--percentage out of 23.

** Two people did not answer this question—percentage out of 22.

Questions 16 and 17 asked the participants whether the decision to teach English was their own or whether they were simply assigned the task. The participants' were divided 50/50 on this, although all but one said they enjoyed teaching the subject. As mentioned at the beginning of this chapter, one teacher refrained from answering these questions.

Although the results presented in this section are not directly linked to the research question, they provided insight into teachers' knowledge and opinions with regard to the KPE3. Taken together with the results from the former section, they paint a better picture of how the test is being used and administered. The findings also raise new questions, such as why did so many marking errors occur when 66.7 % strongly disagree with the statement that the KPE3 is difficult to mark. The data collected in this study does not provide an answer to this question and possible explanations would only be speculative. However, the following section examines more closely the marking errors made by all but two of the participating schools.

6.1 Marking Errors

The data collected from the pupils' test booklets that is relevant to the third research question is presented in this section. The marking errors that were made by teachers were identified since these impair the validity of the test. Thirty-five schools submitted their pupils' test booklets and twenty-four teachers answered the questionnaire; seventeen of the questionnaires can be matched with corresponding pupil data (*See Table 3 at the end of the previous chapter*). Table 17 shows the results from all of the schools, including the number of pupils and the number marking errors per school.

Table 17: Marking errors

School #	Number of pupils	Marking errors	Comments
1	25	31	3 errors on 1 pupil
2A	16	2	
2B	18	8	
2C	17	11	3 errors 1 pupil
2D	17	1	
2E	18	3	
Total from school 2*	86	25	
3	19	7	3 errors on 1 pupil

4	25	3	1 pupil read for
5	4	0	
6	32	11	Gave ½ points
7	44	9	
8	20	30	7 errors 1 pupil 5 errors on 1 pupil 4 errors on 1 pupil
9	33	7	
10	4	3	2 errors 1 pupil
11	13	5	
12	9	4	2 errors 1 pupil
13	11	5	3 errors 1 pupil
14	41	0	
15	30	3	
16	5	2	
17	8	6	6 errors on 1 pupil
18			
19	Only Questionnaire		
20			
21	37	16	4 errors 1 pupil
22	6	3	2 pupils had the <i>match word with picture</i> section read to them
23	48	12	
24	14	6	3 errors 1 pupil
25	8	1	
26	46	13	
27	5	1	
28	9	1	
29	56	24	3 errors 1 pupil
30	20	15	3 errors 1 pupil (x2)
31	23	4	One pupil was undergoing testing for dyslexia and was read to
32	14	8	1 pupil received reading support
33	33	5	
34	44	15	
35	58	35	
36	19	2	
37	42	30	3 errors 1 pupil
38	19	3	
TOTAL	996	370	

*School 2 cannot match the classes (2A, 2B, 2C, 2D, 2E) with the individual teacher's questionnaires as they arrived separately from the test booklets.

Key:

White background = questionnaire + test booklets

Dark shading = only questionnaire

Light shading = only test booklets

Out of all the thirty-five schools that submitted their pupil data, there were only two (5.7%) schools (schools #5 and #14) that displayed no marking errors. The schools in the first part of the table with no shading submitted both test booklets and questionnaires. School 2 submitted five questionnaires as well as all their pupils' test booklets, but since the booklets and questionnaires arrived separately, the test booklets cannot be matched with the teacher who marked them. The schools marked in the darkest shading submitted the questionnaire but no test booklets. The section marked in the pale grey shading sent in the test booklets, but refrained from answering the questionnaire. Marking errors are discussed in more detail in the following chapter.

Summary

In this chapter the results from the questionnaire were presented in the three blocks introduced in the previous chapter. The participants' answer(s) to each question are presented and summed up in tables. Both numbers and percentages have been calculated. In addition, the results from the test booklets collected from schools and analyzed with respect to marking errors were summarized.

Analysis of the data in relation to the threats to validity identified in chapter four, indicated that there was evidence that teachers' practices constituted a potential threat, except for threat number 3 (*teachers failing to be fully prepared themselves with the help of the teacher's guide*). Although this question had not been explicitly asked in the questionnaire, the answers given by many teachers raise questions as to whether or not teachers read the guidelines thoroughly. There were examples of teachers who did not consider doing the practice test together with their pupils to be important, something the guidelines stressed; as well as teachers who incorrectly assumed that the test could establish the level of all the pupils. All but two schools made marking errors, and over half the teachers had no formal training in English or language assessment, all of which could potentially compromise the validity of the test. The results are discussed further in the next chapter.

7. Discussion

According to Hughes (2003), a test is valid if it accurately measures what it is supposed to measure. However, validity is a complex phenomenon, which can be approached from many different angles. Answering the question ‘does the test measure what it is supposed to’ with a simple ‘yes’ or ‘no’ is not necessarily enough, but merits a discussion. For example, whose responsibility is it to ensure that a test is valid? A test can be valid in every respect in theory, and yet not be valid in practice due to human error. As discussed in chapters 3 and 4, Messick’s (1989, 1995) unified concept of construct validity links the social aspects and their consequences to test use. In other words, the users of the test and the test developers share the responsibility.

In chapter 6, evidence indicating that teachers’ practices were threatening aspects of the test’s validity was presented. In this chapter, the results from the questionnaire are discussed in the light of the theory presented, and on which the study was based on. The first section considers each of the six aspects of validity, expanding upon the findings and suggesting possible reasons and implications for these findings. In order to gain deeper insight into the complexity of the problem, the second section focuses on four teachers, for whom both questionnaire data and marking data were available. Their data is summarized and linked to the possible threats their practices pose to the KPE3’s validity. The third section briefly considers which aspects of validity are in the hands of the test developers and/or authorities behind the tests, since validity are not only in the hands of teachers/schools. In conclusion, some weaknesses and limitations of the study are examined.

7.1 The six aspects of validity

The finding will now be discussed in the light of Messick’s six aspect of validity.

7.1.1 Content Validity

As indicated in the theory chapter, content validity implies that the test items are representative of tasks that elicit evidence of the language ability being tested.

Questions 1 and 5 in the questionnaire corresponded to this. Possible threats to content validity were: *failure to use practice tests* and *using the test itself to prepare the pupils*.

With regards to the teachers who answered the questionnaire were not asked directly whether and why they used the practice tests before administering the KPE3 to their pupils. The decision was made to ask instead what they thought it was important to do in order to prepare for the test; giving them the option of ticking off the boxes for both “doing the test themselves” and “doing the test with their pupils”. It was feared that if they had been asked directly whether or not they took the practice test, they would have been reluctant to acknowledge that they had not done so.

It is interesting to note that, although 100% of the participants stated that they thought it was important to read the teacher’s guide, only 70.8% checked the box for doing the practice test with their pupils, a practice that the teacher’s guide strongly recommends in order to familiarize the pupils with the different test item formats and the instructions. For the listening section of the test, the instructions are given on the audio CD. If a child has not heard these instructions or become familiar with the task type, he or she may be confused and not know what to do. If he or she has to rely on reading the instructions, at which point one may question whether their listening comprehension skills are being tested or their reading skills.

The ethical issues associated with language testing are complex and complicated, and an in-depth discussion of these is outside the scope of this study. The study focused mainly on the purpose of the test, how the test was used, and the practical implementations of the test, and not the ethical backdrop of testing. Nevertheless, the questionnaire revealed that some teachers were teaching test content, i.e. using the test itself or content elements to prepare pupils; this is considered negative washback since it threatens the content validity of the test. In the case of KPE3, the teaching of test content refers to the teaching of vocabulary taken directly from the test. This is possible because teachers may have had access to the test prior to it being administered, seeing as the test was the same from 2010-2015. It is evident that the teaching of test content not only affects the validity of the test, but may also be considered unethical behavior, or even “cheating”.

According to the study findings, 16.7% (3 out of 18) of the surveyed teachers answered that the KPE3 influenced their teaching (washback) and that they “practiced words”. In retrospect, the wording that was used in the questionnaire could have been more clearly and concisely

formulated. The phrasing could be interpreted to mean several things. It could be taken to mean that the teachers prepared the pupils for the KPE3 by practicing words from the KPE3. This would have been possible because the test was paper-based until 2016, and was available in the schools. However, it could also be interpreted as meaning that the teachers used the KPE3 to practice words after the test had been taken, for example, if the teacher noticed a trend in vocabulary errors. The latter interpretation would be in keeping with the test function and could be considered positive washback, while the former would constitute an ethical issue and be considered negative washback. Regardless of how the teachers interpreted the question, the fact that the KPE3 was on paper and was not changed for several years made it possible for teachers to teach the content of test. Teaching test-specific vocabulary reduces the test's content validity; the score cannot be relied upon if the pupils have practiced the words prior to the test. The score, therefore, would not reflect the true skill of the pupil, and pupils needing extra attention would not be identified.

A study by Øren Gjelsvik (2012) revealed one possible explanation for this practice. She found that teachers feel pressure from their superiors to ensure that their pupils do well on diagnostic tests, and feel that their pupils' failure to do well reflects badly on their teaching. It is possible, therefore, that some teachers feel pressured into preparing their pupils by teaching the test content.

One can assume that this is a problem only if the test is mandatory. Some municipalities, such as Oslo, have made the KPE3 mandatory, despite the fact that the Norwegian Directorate of Education has explicitly said that it is not (Oslo Kommune/Utdanningsetaten, 2015, Oslo Kommune/Utdanningsetaten, n.d.). Oslo has been criticized concerning kartleggingsprøver (diagnostic tests) for teaching to the test, by the staff at the Reading Centre at the University of Stavanger; their concern was that the Oslo Education Department was not providing adequate information about the purpose of the test, and teachers, who were under a lot of pressure to obtain good performance reviews, coached their pupils before the tests (Mellingsæter, 2014). If test results are being wrongly used to compare teachers, schools and municipalities, it is crucial that more information be provided by the test developers or the Norwegian Directorate for Education and that teachers receive training related to the purpose and use of the test.

It should be mentioned, however, that by making diagnostic testing mandatory in Oslo, they have been able to identify pupils who are not experiencing optimal learning outcomes at an early stage (Aarseth, 2013). Nevertheless, this positive outcome must be weighed up against the possible negative consequence that teachers may be teaching test content. Although the questionnaire did not explicitly refer to teaching test content, this is an important issue that must be addressed since it can seriously affect the validity of the test. Thus, in cases where the tests are mandatory, there need to be safeguards to prevent this from happening in order to ensure the content validity of the test.

7.1.2 Substantive Validity

Substantive validity means that the pupils use the same processes for reading and listening when taking the KPE3 (2014) as they would when reading or listening in the target language, for example in a familiar classroom situation. The corresponding question on the questionnaire was question number one, which was used to elicit information about what teachers considered important when preparing. Substantive validity could be threatened by: *teachers failing to be fully prepared themselves with the help of the teacher's guide and failing to prepare their pupils adequately, including the use of practice material*

Alderson, Clapham, & Wall (1995) stress the importance of the test administrators' role in a testing situation:

The administrators of a test are those who 'deliver' the test to the candidates, and they are responsible for seeing that the conditions in which the test is given provides all candidates with the best chance possible to display the abilities which are being tested. Though the training of administrators need not be as complex... it is still important that the administrators understand the nature of the test they will be conducting, the importance of their own role and the possible consequences for candidates if the administration is not carried out correctly (Alderson, Clapham, & Wall, 1995, p. 115).

In order for a test to show evidence of substantive validity, therefore, the testing conditions should be optimal for the pupils. Test administrators in the case of the KPE3 (2014) were in most cases teachers. The teacher's guide clearly outlines what teachers are to do in order to prepare themselves and familiarize their pupils with the testing situation. If the pupils have not been acquainted with the different types of test items, or have not been prepared for what the testing situation would be like, and how much time they have for each section, this can

affect their performance and may not give an accurate indication of what they are actually capable of achieving.

Particular challenges can arise with regards to timing of the test. Since the KPE3 was paper-based in 2014, and the audio for the listening section had to be played on a CD player, which meant that the whole class had to work at the same speed. A specific time limit had been set for the two sections of the test, and it was therefore paramount that teachers were aware of this and stopped the CD at the correct intervals.

In 2016, the researcher was an observer in four schools evaluating the administration of the new digitalized KPE3 to third graders. It was very evident which teachers had read the guidelines and which had not by taking note of how the test was implemented in the classroom. One teacher even admitted to not having read the guide; there was a lot of chaos in that classroom, and the test guidelines were not followed when it came to giving directions and timekeeping. It is no unreasonable to assume that similar situations had arisen when the test was paper-based—and were occurring in other schools as well.

Danielsen's (2013) study on the diagnostic test of reading for grade two revealed that there was great variation in teachers' and school leaders' familiarity with the teacher's guide. She reported that two out of seven teachers did not know about the guide, and one teacher had only recently been made aware that such a document existed. Of the remaining four that knew about the guide, only one teacher said that she used the guide actively before and after administering the test (Danielsen, 2012, p. 69). If this is happening with a test that is mandatory, there is reason to be concerned that even more teachers who are administering the KPE3 are not familiar with the content of the KPE3 teacher's guide, and some may not even know of its existence.

As previously mentioned, 100 % of the teachers who answered the questionnaire stated that they thought it was important to read the teacher's guide when preparing for the KPE3. In spite of this, responses to other questions on the questionnaire can lead one to wonder if they have read it thoroughly. A possible explanation for teachers not using the guide may lie in their motivation to have their pupils do the test. The questionnaire revealed that only six teachers stated that it was their decision to administer the test to their pupils, an indication that the majority were instructed to do so by their superiors. It can be questioned whether or not

teachers are given adequate information as to why they are to administer the test and time to properly prepare for the test. If teachers are not properly prepared, this affects the substantive validity of the test. It is up to teachers to ensure that the testing situation is the best possible for their pupils.

The distribution of relevant information by the test developers and other responsible parties is paramount to ensuring the substantive validity of the test. More importantly, it is paramount that this information reaches the teachers. If the test developers have provided sufficient information, it is not unreasonable to claim that the responsibility for ensuring that the teachers receive this information lies with the school administration. However, it seems that there are misconceptions about the test, even among school administrators.

As mentioned, few teachers are making the decision to administer the KPE3. Since only 25 % of the teachers indicated that they were the ones who decided to administer the test to their class, the remaining 75% were being told to do so. Unfortunately, the data obtained from the questionnaire does not indicate why teachers are not making this decision themselves. It could be possible, however, that this may help to explain many of the findings: why there seemed to be misconceptions among the teachers about such basic issues as the purpose of the test, why they didn't seem to read the teacher's guide, and perhaps even why there were so many marking errors. While some school administrators might, in accordance with the test aim, have wanted to identify the pupils who need special attention so that they could receive the help they needed, others might, misguidedly, have wanted to gain an overview of how the third graders at their school were doing in English reading and listening. If that was the case, one can question whether they were aware of the fact that the KPE3 cannot provide that information due to its limitations and purpose.

7.1.3 Structural Validity

Structural validity refers to the degree to which the scores presented fully reflect the skill being tested, and how this skill is made up. Questions 2 and 3 on the questionnaire pertained to the structural validity of the test and the threat posed was the issue of *interpreting the scores as reflecting the abilities of all pupils*.

In order for the KPE3 to be structurally valid, the way in which the results are presented must reflect the way the skills being tested (reading and listening). In the KPE3, the scores in these two skills provide the basis for deciding whether a pupil appears to be on or below the limit of concern in either reading or listening. Although several aspects of reading and listening are tested, the test results do not present any subskill scores; and no claims are made related to any skills but reading and listening. The scoring of the test is designed to reveal a minimal level of competence in each skill, and not to provide scores that reflect the abilities of pupils beyond this level. A threat is posed to the structural validity of the KPE3 if the scores are interpreted as reflecting the abilities of all pupils.

Study findings indicated that 62.5% of the participants considered that the KPE3 was capable of identifying the level of the entire class, while 54.2% assumed this was how the test actually was used. The questionnaire revealed that 41.7% of the participants generally agreed that the KPE3 reflected their pupils listening and reading skills, while 33.3% generally agreed that the test reflected the English competence of their pupils. There was virtually no difference between the number of participants who agreed and those who disagreed with the statement that the test is a good tool for assessing their pupils. The study data did not reveal why half the teachers surveyed did not consider the KPE3 to be a good assessment tool. One possible reason could be that they assumed the test could discriminate on the whole learning spectrum, and therefore considered the KPE3 too easy for their strong pupils. This conception of the test was a further result of misinterpretation of the purpose of the test, which is an example of a threat the structural validity of the test.

In order to assist teachers' identification of where their pupils' weaknesses may lay, the teacher's guide states what each item format is designed to test. The different aspects are in accordance with current thinking regarding what subskills comprise reading and listening. Teachers are given advice on how to go through the KPE3 with pupils, in order to establish why the pupil has given an incorrect answer. Nevertheless, 45.8% of the teachers who answered the questionnaire did not think the test was being used to give feedback to pupils. Thus, the KPE3 was not being used as a form of formative assessment since the teachers were not sitting down with their pupils and going through the test to make them aware of their mistakes and inquire why they had given the answers they did. It is important to take into account that guessing can occur, and one way a teacher can determine for certain whether a pupil has understood a task is to review the test together with the pupil. According to Wiliam

(2011), in order for feedback to promote learning, the pupil must be guided and be given a “recipe” for how s/he can improve.

7.1.4 Generalizable Validity

In order to ensure the generalizable validity of a test, the test scores have to be consistent and reliable. This can be affected by the way the test is carried out, as well as the way it is scored. The question on the questionnaire pertaining to this was number 8 and the threat to generalizable validity was posed by *incorrect marking*.

Marking errors affect the generalizable validity of the test because, if the scores are unreliable, the results cannot be trusted. The score of the test should reflect the test taker’s ability in a certain area (in this case listening and reading). This means that a pupil should get the same score if he or she were to take the test again or a similar test with a similar test construct. However, if the answers that are incorrect are marked as correct, or vice versa, the score cannot be relied upon. This is especially crucial with regards to pupils who score in and around the limit of concern, since it is this group that the KPE3 is intended to identify. It is not a matter of concern when pupils who should have scored a total of 47 of the possible 50 points are awarded +/-2 points due to marking errors. Not only are these pupils far above the limit of concern and unlikely to need extra support in their reading or listening skills acquisition, but their proficiency level is average or above-average and the test is not intended to precisely identify these groups. This being said, teachers are doing these pupils a disservice by not correctly assessing their work.

As can be seen in Table 17 in the previous chapter, there were sixteen cases where two or more marking errors were made on the same pupil’s test. In school #17, six marking errors were made on one pupil’s test. This pupil was given more points than he should have received, putting him over the limit of concern in reading. There was also one case (in school #6) where the teacher awarded half points, despite the fact that the guidelines specify that answers are either right or wrong and only whole points are allowed.

Four schools indicated (schools #4, #22, #31, #32) reading support had been given during the test to five pupils in total. Such support is not general practice during a test that is designed to

test pupils' reading skills. If a text is read to a pupil, it is their listening skills that are being tested rather than their reading skills.

The limit of concern for the listening section of the KPE3 was set between 16 and 17 points, and between 18 and 19 for the reading section. The pupils from schools #4, #22, and #32 who had received assistance in the reading section, all scored at or under the limit of concern in the listening section. However, this was not necessarily the case in the reading section, which could indicate that their reading score may not be accurate. For example, while only scoring 2 out of a possible 24 in the listening section, the pupil from school #4 scored 25 of a possible 26 in the reading section.

The pupil from school #31 was undergoing testing for dyslexia and had all but the last six questions read to him or her, resulting in 21/26 on the reading section. In each of these examples, the pupils scored well above the set limit of concern on the reading section, a level that they would not likely have achieved without assistance. This is clearly a threat to generalized validity.

It is important to note that the KPE3 was not designed to provide very specific information for weak pupils with learning disabilities, and the results may indicate problems that are not specific to English. Although the teacher's guide gives some general advice for working with very weak pupils, the pupils falling in this category must be assessed further by experts in the field (Utdanningsdirektoratet, 2010).

Errors have been detected in all the task types in the KPE3; wrong answers have been marked as correct and vice versa. This can be illustrated by examples taken from the last part of the reading section. The last six tasks on the KPE3 are of the type: "copy the word". Pupils have to look at a picture, find the corresponding word in a list of words and write it beside the picture. Spelling was an issue that was not always taken into account by the markers. In one case where the correct answer was "sandwich", the pupil wrote "sandjis" and it was marked as being correct. Several other pupils wrote "bycycle" and this was not marked as incorrect by the teacher. There was even one pupil who wrote "sandwich" next to the picture of the bicycle and was given a point. This causes one to wonder why so many marking errors are being made when the test is very easy and the teachers are provided with an answer key. One explanation could be that teachers are overworked and simply do not have time to mark the test, and are therefore making careless errors. However, regardless of the reason, it is

reasonable to assume that the weaknesses in marking revealed in this study are not confined to the context of this test alone but are likely to occur in many other assessment situations.

When the findings from the questionnaire were compared to the grading on the submitted test booklets, it was found that the teacher who never used the answer key when marking had seven marking errors. The teachers who had zero marking errors on the test papers used the answer key “some of the time”, and “usually”. Although it is not possible to claim that all teachers who did not use the answer key made a lot of marking errors, one conclusion that can be drawn from this finding is that teachers did not always use the answer key, and that marking errors have been made by teachers in 94.3% of the 35 sampled schools. This is a very clear indication that teachers’ practices are invalidating the KPE3.

Incorrect marking affects the generalizable validity of the test and out of the thirty-five schools that submitted their pupils’ test booklets, only two schools showed zero marking errors.

7.1.5 External Validity

In order for a test to be considered externally valid, the score should have some correlation with other evidence of a pupil’s ability. The test score should therefore not come as a surprise to the teacher or test taker. In the KPE3, a threat to external validity was posed by teachers’ lack of training in the area of language assessment, and question number 15 was the corresponding question on the questionnaire.

The KPE3 teacher’s guide instructs teachers to consider the test score in the light of other assessments and not to trust the test result blindly. For external validity to be evident, the teachers should roughly be able to predict the outcome of the test based on other assessments they have done. It is important, therefore, that teachers are competent in language assessment.

The study revealed that 65.2% of all the participating teachers had no formal training in language assessment. Other research has confirmed that such training is important. A study from Colombia (Lopez Mendoza & Bernal Arandia, 2009) found that teachers’ training in language assessment clearly influenced their perceptions of language assessment and how

they used language assessment in their classroom. It also showed that the lack of training in language assessment often resulted in tests being used in other ways than initially intended. Moreover, a study by Moss, Girard, & Haniford, (2007) has shown that teachers' conceptions of valid assessment varies; what is valid for one teacher may not be valid for another. It is doubtful, therefore, whether teachers who lack training in language pedagogy or assessment can make the necessary judgments about pupils' abilities. If Norway is going to continue to use standardized testing in schools, and teachers are going to be expected to carry out formative assessment with their pupils, teachers need to be adequately trained to use tests correctly and to employ assessment practices that promote learning. Only when teachers are competent in assessing their pupils' performance can external validity be achieved.

In 2015, the Norwegian government signed a document entitled "*kompetanse for kvalitet*" (competence to ensure quality). This strategy is intended to raise teachers' qualifications by the year 2025. For primary schools, this will mean that teachers who are to teach English must have at least 30 credits in the subject (Kunnskapsdepartementet, 2015). Although this is a step in the right direction, one can question whether 30 credits (equal to one semester) are sufficient to qualify teachers to teach and assess pupils in English. Some may consider this adequate, since grades 1-4 are only allocated one hour of English per week. Rixon's (1999) research on English as a foreign language in primary schools in a number of European countries stresses the significance teacher competence and contact hours with pupils for teaching methodology, and how proper training and sufficient time is paramount in meeting the demands of teaching a foreign language. Even though the KPE3 is administered in the third grade where pupils are only allocated one hour of English per week, teachers still need the proper training in order to ensure the external validity of the test.

7.1.6 Consequential Validity

The last aspect of validity to be discussed is that of consequential validity. A test is only consequentially valid if it is used in accordance with the purpose for which it was produced. Thus, consequential validity can be threatened by teachers *misusing the test results* and/or their *inability to give necessary support to weak pupils due to lack of training in language pedagogy*. Questions 2, 3, and 14 on the questionnaire corresponded to this aspect of validity.

The test developers can control the content of a test, but cannot control how it is used, although they can have some influence. The responsibility for implementation lies not only with the teachers, but also with school leaders and school owners, since they are often the ones making the decision to administer the test. According to Bachman (2005), “The single most important consideration in both the development of language tests and the interpretations or their results is the purpose or purposes which the particular tests are intended to serve” (Bachman, 2005, p. 2). Therefore, the consequential validity of the KPE3 is only ensured if it is used in accordance with the purpose for which it was made: to identify the pupils who are struggling in English reading and listening.

In 2014, teachers who completed the questionnaire seemed to be confused as to what the purpose of the KPE3 was. One teacher did not identify the purpose of the KPE3 as being to find the weakest pupils, and two teachers did not think the KPE3 was being used to find the pupils who needed extra attention. In addition, 62.5% of the participants thought that they could gain information about the level of the entire class from the test, and 54.2% considered that this was how the KPE3 was being used. The high percentage of teachers who responded in this manner raises the question as to whether teachers want another type of test. It almost seems as if teachers want a “National Test in English” for grade three.

It is difficult to say whether teachers’ misconceptions about the purpose of the KPE3 affect the test’s validity in relation to how they implement it in their classrooms. While understanding the purpose of the test is the teacher’s responsibility, it is hard to say with certainty that it directly affects the test validity in all cases. It is entirely possible, and one may suspect that this is often the case, that misconceptions influence feedback, and the general way they administer the test. However, this is hard to measure and one cannot therefore say whether or not it affects the consequential validity of the test directly.

7.2 Examples: Linking questionnaire data and marking errors to validity threats

Teacher’s practices can potentially affect the validity of the KPE3. In this section a closer examination of four teachers’ backgrounds, practices and attitudes toward the test are linked to their marking errors. This is done in order to provide deeper insight, through concrete

examples, into the validity issues discussed in the previous sections. These teachers' questionnaires can be matched to the first section of Table 17 in the previous chapter which showed schools that had submitted both test booklets and questionnaires.

The first example is a **female teacher from school # 8** who had more than 10 years of work experience, had chosen to teach English and enjoyed doing so. She had 15 credits in English, and had taken a credited-bearing course in language assessment. In addition, she herself had made the decision to administer the KPE3 to her class, while maintaining that she partially agreed that her school focused too much on the test. She considered the KPE3 to be necessary, but partially disagreed that the test reflected the pupils' reading and listening skills, as well as their general competence in English. In addition, she partially disagreed with the statements that the KPE3 was a good assessment tool or that follow-up was difficult.

She considered that the purpose of the test was not only to find the weakest students, but also find the level of the entire group, as well as to give feedback to the pupils, their parents, the school leaders and the school owners. In reality, she assumed this was the way the test was being used, although she did not think that school owners were given feedback about the test.

She stated that, in order to prepare for the test, it was important to inform the parents about the test and to read the teacher's guide, which she somewhat agreed gave adequate information. She also took the practice test alone, as well as with her pupils.

The test influenced her teaching in that she adapted the teaching to suit the weakest pupils.

She marked the test herself, using the answer key only some of the time and strongly disagreed with the statement that the test was difficult to mark. However, there were 30 marking errors in the 20 test booklets that were submitted. There was one case of seven marking errors on the same pupil's test, as well as two cases where four and five marking errors were made on the same pupils.

The potential threats identified in the case of the teacher from school #8 would be related to structural validity. This teacher thought that the score of the test would reflect the abilities of all her pupils, and not just the weakest. The test's generalizability validity was also violated in that there were many cases of incorrect marking.

The second example is a **male teacher from school # 14** who had less than 2 years teaching experience, and no formal training in English or language assessment. He had not chosen to teach English himself, but he enjoyed doing so. He commented on the questionnaire that he did not think that the KPE3 was relevant to the majority of the learning objectives in the curriculum (LK06).

To prepare for the test, he considered that it was important to read the teacher's guide, which he generally considered to give adequate information. In addition, he prioritized organizing the practical aspects of the test and going through the practice test together with his pupils. He did not consider it important to familiarize himself with the practice test on his own prior to doing it together with his pupils.

With regard to the questions regarding the purpose of the test and how it was used, he felt that these were one and the same, in his opinion. He thought that the purpose was to find the weak pupils, to find the level of entire group, and to give feedback to school leaders. The principal was the one who decided that the class was to take the KPE3, but he strongly disagreed that there was too much focus on the test at his school.

He partially disagreed with the statement that the KP3 reflected the pupils' reading and listening skills, and mostly disagreed that it reflected their general English competence. In addition, he partially disagreed that the KP3 was a good assessment tool and partially agreed that it is difficult to follow up the results. With regard to whether the KP3 was necessary, he answered "I don't know" and refrained from answering the question regarding whether the test influenced his teaching.

He marked the test himself, something he did not consider difficult, and used the answer key most of the time. He made zero marking errors on 41 pupils' papers.

This teacher is one of only two teachers that made no marking errors. He did not, therefore, violate the generalizable validity of the test. However, his belief that the test discriminated on all levels threatens the structural validity of the test. Moreover, his lack of training in both language assessment and language pedagogy threatens the external and consequential validity of the test respectively.

The third example is a **female teacher from school #17** who considered making the practical arrangements and reading the teacher's guide to be important when preparing for the test. She did not indicate whether she had done the test herself or together with her pupils.

She generally agreed that the KPE3 reflected the pupils' reading and listening skills, as well as their general competence in English. She also generally agreed that the test was a good tool to assess her pupils. She had only eight pupils and marked their tests herself, always using the key. However, she made six marking errors, all of which were on the same pupil's test. If marked correctly, this pupil would have been exactly on the level of concern in reading with 18 points. However, the pupil was awarded 20 points, putting him above the limit. She generally disagreed with the statements that the test was difficult to mark and that following up the results was difficult.

This teacher had more than 10 years teaching experience, but had no formal education in English, although she had taken a course in language assessment. She chose to teach English and enjoyed doing so. She stated that the purpose of the test and how it was used was one and the same thing: to find the weakest pupils and to give feedback to pupils and their parents. On the issue of whether or not the test influenced her teaching, she indicated that said that they had practiced the words before the test.

In this teacher's case, the content validity of the test was compromised. Although she did not indicate whether or not she had familiarized herself and her pupils for the test by doing the practice test as the guidelines encourages, she did not consider doing the practice test to be important. This failure to properly prepare herself and her pupils potentially affected the substantive validity of the test. Her marking errors meant that the generalizable validity of the test was impaired, and her lack of training in language pedagogy draws the consequential validity of the test into question.

The last example is a **female teacher from school #9**, who, when preparing for the test, considered reading the teacher's guide to be important, as well as doing the practice test alone and together with her pupils. She thought that the purpose of the test was to find the level of the entire class, and to give feedback to pupils and parents. She assumed the test was used to find the weakest pupils, as well as the level of the whole group. The decision to take the test

had been made by the principal at her school, and it was her first time administering the test. She did, however, consider the test to be necessary.

She strongly agreed with the statements that the teacher's guide gave adequate information and that the test was difficult to mark. She marked the test herself but did not refer to the answer key. She made seven marking errors in her class of 33 pupils.

She generally agreed with the statements that her school focused too much on test, and that the test was a good tool to assess pupils. However, she only partially agreed that the KPE3 reflected the pupils' reading and listening skills, as well as their general competence in English.

This teacher had 6-10 years teaching experience, but had no formal education in either English or language assessment. She had, however, chosen to teach English and enjoyed doing so.

The practices of the teacher from school #9 constituted a threat to the structural validity of the test since she assumed that the test scores could be interpreted as reflecting the abilities of all of the pupils. In addition, incorrect marking pose a threat to the generalizability of the test. Moreover, despite having taught for several years, she had no formal training in language assessment or English pedagogy, which potentially influences the external and consequential validity of the test,

7.3 The responsibility of test developers/authorities

This study has considered the ways in which the validity of the KPE3 could be compromised by certain teacher practices. The full responsibility for validity does not lie solely with teachers, however, so mention should be made of some of the obligations the test producers have with regard to safeguarding the test's validity. It is their responsibility to ensure that a test itself is a valid assessment tool—that it actually assesses what it is supposed to. Hamp-Lyons (1997) takes it even further by urging test developers to take responsibility for all of the consequences that we are aware of, including those related to practice:

Furthermore, there needs to be a set of conditions and parameters inside which we are sure of the consequences of our work and we need to develop a conscious agenda to push outward the boundaries of our knowledge of the consequences of language tests and their practices (Hamp-Lyons, 1997, p. 302).

In chapter four, Hasselgreen's analysis model was used to identify the key elements crucial for various types of validity and what poses a threat to them in the KPE3. Having focused on the aspects of validity that lay in the hands of teachers, a brief consideration of the ways in which test developers and the authorities behind the tests can safeguard the validity of tests such as KPE3 may be useful at this point.

7.3.1 Content Validity

It might seem obvious that test developers are responsible for a test's content validity. It is they who ensure that the test content is actually testing the skills and abilities it is designed to test. When a test development team is commissioned to make new tests, strict guidelines are provided, such as the Framework for Diagnostic Testing for the KPE3 that is presented in chapter two. Part three of the framework strictly outlines the requirements that have to be met in relation to what is to be tested, among other things. For example, in order to meet the reliability requirement, the test items have to be piloted before compiling the items that discriminated the best into one test.

7.3.2 Substantive Validity

With regard to substantive validity, the KPE3 was designed to include reading and listening tasks of the type that the pupils are expected to be able to carry out in a classroom situation, and test developers must bear this in mind. In the case of the KPE3, the curriculum and the framework underlie the test. For example, the extensive use of pictures provides a context for the reading/listening tasks that supports the children in a familiar way. In this test, all of the tasks were supported by pictures, a practice with which the pupils are familiar, and the themes and the vocabulary were familiar, being limited to "common words and phrases relating to the pupils' immediate environment" (translation my own) (LK06). In addition, example tasks are provided before and during testing in order to ensure that the test formats are not new to pupils.

7.3.3 Structural validity

The structural validity of the KPE3 has also been taken into account in developing the test. Since it tests two skills, reading and listening, separate scores are given for each. The test is designed that the listening section comes first in the test, followed by the reading section. The number of items in each section is approximately the same. For the listening part of the test, the reading skills of the pupils are not tested since the instructions are given audibly, although simplified instructions are written before each task type as a support. In the reading section, pupils have to read the instructions in order to understand and execute the tasks, listening is not required, apart from the instructions given by the teacher such as how much time they have to complete the test. Thus, the test structure is intended to ensure that the skills being separated do not overlap.

7.3.4 Generalizable Validity

The primary threat to generalizable validity in the original KPE3 test was marking errors. Following a decision made by the Directorate of Education, the KPE3 was digitalized in 2016, so this is no longer an issue; teachers are no longer required to mark the tests manually as a computer does this automatically. The answer is either right or wrong, as coded by the test developers, and pupils receive one point for every correct answer. These are then totaled and pupils receive one score for the listening section of the test, and one for reading. The scores can therefore be trusted, thereby ensuring the generalizable validity of the test.

7.3.5 External Validity

Although the test developers cannot influence the external validity of the KPE3 since they do not control over what other assessments teachers have of their pupils, the school authorities do. Teachers are required to document their pupils' progress and competence, which ensures that some form of evaluation is undertaken. This is referred to in the teacher's guide provided by the test developers instructs the teachers to use their own judgment when considering the pupils' individual scores and to view them in light of their other assessments.

7.3.6 Consequential Validity

There is only so much test developers and authorities can do to ensure the consequential validity of a test. How the test is used and whether it is used in accordance to the purpose for which it was intended is beyond the control of the test developers, and not fully in the control of the authorities either. However, there are steps that can be taken in order to safeguard the consequential validity of a test. In the case of the tests commissioned by the Directorate of Education, their website⁸ is very informative about the diagnostic tests, including the KPE3. The test is not obligatory and the purpose of the test is very clearly outlined. In addition, substantial changes were made to the teacher's guide when the KPE3 was digitalized; new instruction and examples were required. The guide also very clearly defines the purpose of the test and how the teacher can use it as formative assessment tool to facilitate learning. There is also an online practice test that teachers are strongly advised to do together with their pupils before administering the test. Nevertheless, all of these efforts may not eliminate the human factor and its impact on consequential validity.

7.3.7 Where responsibilities meet

The crossroads where teachers, authorities and test developers' responsibilities meet is a gray zone that is not easy to define. In trying to establish a concise and clear division of responsibility, one quickly discovers why this is no easy feat. If a test is a valid assessment tool if used correctly, one can claim that it is the test developers' responsibility to develop a valid test and that it is the teachers and the school administrators responsibly to implement the test in accordance with its purpose. However, as this study has shown, the areas of responsibility overlap.

7.4 Weaknesses and limitations of the study

Looking back at the study, there are some weaknesses and areas that could have been improved. The wording in the questionnaire that was sent out to teachers was, on occasion,

⁸ <http://www.udir.no/eksamen-og-prover/prover/kartlegging-gs/>

ambiguous and should have been clearer. For example, the statement, “find the level of the entire group,” could be interpreted to mean that the test can discriminate between all the pupils and give information about all the levels within the class, but could also mean that the test is able to find an average that represents the entire class. However, regardless of how this was interpreted, the fact remains that the test purpose was misunderstood. Furthermore, as already mentioned, the question, “Does the test influence your teaching?” could be referring to before or after the test.

It would also have been useful to interview some of the teachers who both answered the questionnaire and submitted their pupils’ test booklets. This might have provided more insight into their attitudes and practices, as well as into the underlying reasons; for example, the reasons why marking errors were occurring. It might also have opened up new paths of enquiry.

It is important to stress that this study is not based on a large representative sample so the results cannot be generalized to the whole population. Nevertheless, it provides indications of potential threats to validity that can be followed up. There is clear evidence that things are happening that threaten the validity of the KPE3. Most importantly, even if one child misses out on the support he or she needs because the test is not being used correctly, that is one child too many. Finally, steps can be taken to eliminate these threats even without further investigation; for example, the digitalization of the tests has eliminated the problem of marking errors.

Summary

In this chapter, the results from the questionnaire have been discussed in light of the theory earlier presented. The first section considered each of the six aspects of validity, expanding upon the findings in relation to these and suggesting possible reasons and implications. Next, the questionnaire responses of four of the participants were summarized in order to provide a clearer picture of teachers' practices and their knowledge about the KPE3, as this might pose a potential threat to the validity of the KPE3. The following section discussed which aspects of validity are in the hands of the test developers and/or authorities behind the tests. Finally, some weaknesses with the study were mentioned, as well as what could have been improved.

8. Conclusion

The aim of this thesis has been to examine more closely the KPE3 (2014) and the potential threats to its validity, in relation to teachers' and schools' responsibilities in particular. The research questions for the study were as follows:

- 1) What aspects of the KPE3 potentially make its validity vulnerable?**
- 2) Which of these aspects are in the hands of the teachers/schools?**
- 3) Is there evidence to indicate that some of the aspects related to teachers' practices may actually be undermining the test's validity?**

Following a brief introduction, background information related to KPE3 was presented, including the purpose of the test, the test itself, and the material that accompanied the test. Two key documents guiding the development of this test were also introduced: the *Framework for diagnostic tests for grades 1-4*, which outlines the principles for test construction and for the teacher's guide; and the sections of the *English Curriculum* (LK06) for grade 2(3) that lay the foundation for the test construct.

In regards to the research and theoretical background for this study, previous research on the low-stakes diagnostic tests (kartleggingsprøver) used in Norway was considered, before presenting and discussing some of the test-related concepts associated with central issues in this study: formative assessment, test construct, and test validity. The perspective then narrowed to focus on Messick's six aspects of validity – content, substantive, structural, generalizable, external, and consequential validity – which lay the foundation for the theoretical discussion in this study. In answer to the first research question, therefore, the KPE3 (2014) was examined with reference to the aspects of validity identified by Messick and aspects that might make the test's validity vulnerable were outlined.

Hasselgreen's analysis model for identifying the key elements of various types of validity, and the potential threats these face, was then introduced; and on the basis of this model, potential threats to the validity of the KPE3 (2014) were identified, focusing particularly on those that are the responsibility of the teachers/schools. In answer to the second research question, the following threats were identified:

- failure to use practice tests
- using the test itself to prepare the pupils
- teachers failing to be fully prepared themselves with the help of the teacher's guide
- failing to prepare their pupils adequately, including the use of practice material
- interpreting the scores as reflecting the abilities of all pupils
- incorrect marking
- teachers' lack of training in the area of language assessment
- misusing the test results
- teachers' inability to give necessary support to weak pupils due to lack of training in language pedagogy

These threats formed the backbone to the rest of the study, in which the data collected was analyzed to identify any aspects related to teachers' practices that might actually have been impairing the validity of KPE3 (2014). The research method chosen was quantitative, employing two data sources: pupils' test booklets and a questionnaire sent out to teachers who had administered the test. The discussion probed further into the findings, seeking possible explanations for the problems identified. This included a deeper analysis of the questionnaire answers and marking errors of four example teachers. As validity is not only in the hands of the users, but also in the hands of test developers and authorities, this was briefly discussed as well.

The overarching finding related to the third research question indicated that all of Messick's aspects of validity were being impaired by teachers'/schools' practices or lack of training. All of the potential threats that had been identified were evident in relation to KPE3 (2014), except *teachers failing to be fully prepared themselves with the help of the teacher's guide*; this information was not explicitly asked for in the questionnaire since it was expected that few teachers would admit to not doing so. The participating teachers were asked, therefore, how important they thought it was to read the guide when preparing for the test. The answers given by many teachers raised questions as to whether or not they had read the guidelines thoroughly.

Analysis of the data from the questionnaire revealed that the main threats to the test's validity were the failure to use the practice material, misinterpreting the purpose of test, and lack of teacher qualifications. In addition, the participants seemed to share a misconception about the purpose of the test; over half them assumed that the test could identify the level of the entire class. Another interesting finding was that only half of the teachers surveyed believed the

purpose of the test was to give feedback to pupils. It is difficult to say whether teachers' misconceptions about the purpose of the test directly affected the validity of the test in all cases. It is quite possible, and perhaps safe to assume, that the misconceptions influence feedback and the general administration of the test. However, this is hard to measure and no claims can be made based on this study, although it is the responsibility of teachers to know and understand the assessment tools they are using in their classrooms. Furthermore, it seems that more training is needed as over half of the participants had no formal training in both language assessment and English, which is a matter of great concern. The questionnaire data, therefore, was an important source of evidence that teachers' attitudes and practices were negatively affecting the validity of the KPE3 (2014).

This was further supported by the analysis of the test booklets, which revealed that there were marking errors in all but two schools, raising concern that other assessments happening in the language classroom might also be suffering. The reason for these errors was not clear from the questionnaire, but the responsibility for marking the test correctly lay with the teachers, and they were supplied with an answer key.

Kubiszyn & Borich (2007) state that “[...] if a test is to be used in any kind of decision making, or indeed if the test information is to have any use at all, we should be able to identify the types of evidence that indicate the test's validity for the purpose it is being used for” (p. 307). The purpose of the KPE3 is to identify the pupils who need extra support in English listening and/or reading. This is the only purpose the test was designed for, although many of the study participants were under the misconception that it could identify the level of a class in English. This study, although small scale, has shown that there is evidence to indicate that the validity of the KPE3 is being threatened by teachers' beliefs and practices.

With reference to the third research question, therefore, the analysis of both questionnaires and test booklets has confirmed that some of the aspects related to teachers' practices may actually be undermining the test's validity. Having identified that a problem exists in relation to validity, the next question to be addressed is: What can be done to rectify this situation?

Although this study was not large enough to provide more than indications that teachers' and schools' practices constituted a potential threat to the validity of the KPE3 (2014), these issues are a matter of concern in all areas of assessment in the primary education sector. It

would be useful, therefore, to examine the teacher-education programs around the country in order to find out whether students are being adequately taught about assessment at the different institutions. Since the use of standardized testing is on the increase in Norway, the institutions that educate teachers have a responsibility to their students to train them to use these tests correctly. Further research should be conducted in order to determine whether such programs are in place, since this is a necessary prerequisite to developing a valid assessment practice that promotes learning. A comparative study of the training programs offered at the teaching institutions in Norway would provide insight into the different institutions' practices, and provide a basis for developing assessment courses that would minimize teachers' impact on test validity.

Another area in which further research is needed relates to how test results are being followed up. This study did not investigate how the test results were being interpreted or followed up by the teachers. However, it was found that 68.2% of the teachers disagreed in various degrees with the statement that it was difficult to follow up the KPE3 (2014), indicating that 31.8% actually found it difficult. Various studies (Aarseth, 2013; Danielsen, 2012; Øren Gjelsvik, 2012, Rogaland Revisjon IKS, 2012) report that there is great variation in the way test results are being followed up. To gain a better understanding of this in relation to the KPE3, further research is needed. A similar study to those mentioned above would be useful to examine whether the KPE3 is actually identifying the pupils who need extra attention and whether they are receiving the help they need. In addition, it would be useful to investigate the quality of that remedial action, and whether or not pupils are receiving the same support regardless of external factors such as where they live in the country.

The analysis of the data from this study has shown that there appears to be vital information – about the test purpose and/or how to prepare for the test – that is not reaching all the teachers and schools. As of 2016, the KPE3 has been digitalized and, although the generalizability aspect of validity is no longer an issue, this may give rise to other challenges. One thing that test developers and teachers alike must be aware of is the fact that technical problems may pose a threat to the substantive validity of the test. When a test is computerized, another stakeholder is added to the group, the software developers (Fikke & Helness, 2012), which adds another “link” to the information chain. When technical problems arise, it is not always easy to know whose responsibility it is to fix it; for example, the school's IT-department, the test administration unit or the software developers. Regardless of who is responsible for the

problem, it is vital that the testing situation is as optimal as possible for the pupils. The teacher's guide clearly outlines who is responsible for what, but as this study has shown, misconceptions arise even with clear information. As long as the various stakeholders do not follow through on those responsibilities, the validity of the KPE3 will continue to be compromised. Further research is needed to identify where weaknesses lie in the chain of information, and what the effects of digitalizing the test have been in relation to safeguarding the test's validity.

Lastly, this study was carried out on the 2014 paper-based KPE3. It would be useful to carry out a similar study on the digital tests to see whether teachers' practices are still affecting the validity of the test, now that they have less responsibility and have had more experience using the tests.

Given the increasing demand for evidence of pupils' performance in relation to subject competence goals, and given the increasing emphasis on assessment and testing of this competence, it is essential that studies such as this one are prioritized. It is only by identifying potential threats to validity that the test scores can be relied upon.

References

- Abrams, L. M., Pedulla, J. J., & Madaus, G. F. (2003). Views from the Classroom: Teachers' Opinions of Statewide Testing Programs. *Theory Into Practice*, 42(1), 18–29.
http://doi.org/10.1207/s15430421tip4201_4
- Alderson, J. C., Clapham, C., & Wall, D. (1995). *Language Test Construction and Evaluation*. Cambridge University Press.
- Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly: An International Journal*, 2(1), 1–34.
- Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.
- Brown, J. D. (2001). *Using Surveys in Language Programs*. Cambridge University Press.
Retrieved from https://books.google.no/books?id=W8y_1D59SyIC
- Cameron, L. (2001). *Teaching Languages to Young Learners*. Cambridge: Cambridge University Press.
- Carlsen, C. (2007). The place of reliability within a unified (Messickian) view of construct validity. In C. Carlsen & E. Moe (Eds.), *A Human Touch to Language Testing*. Oslo: Novus AS.
- Danielsen, K. (2012). *MED KARTLEGGING UNDER LUPEN. -En kvalitativ undersøkelse av læreres og skolelederes uttalte oppfatninger av, og rutiner omkring, gjennomføringen av den obligatoriske kartleggingsprøven i leseferdighet.* (Masteroppgave) University of Stavanger. Retrieved from <https://brage.bibsys.no/xmlui/bitstream/handle/11250/185812/Danielsen,Kjersti.pdf.pdf?sequence=1>
- Dörnyei, Z., & Csizér, K. (2012). How to Design and Analyze Surveys in Second Language Acquisition Research. In A. Mackey & S. M. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 74–94). Chichester: Wiley-Blackwell.
- Fikke, A. J., & Helness, H. L. (2012). Synergies and Tensions in Computerised Language Testing. In D. Tsagari & I. Csépes (Eds.), *Collaboration in Language Testing and Assessment* (pp. 159–170). Frankfurt: Peter Lang GmbH.
- Gronlund, N. E. (1985). *Measurement and evaluation in teaching*. Macmillan.
- Hamp-Lyons, L. (1997). Washback, impact and validity: ethical concerns. *Language Testing*, 14(3), 295–303.
- Hasselgreen, A. (2004). *Testing the Spoken English of Young Norwegians: A Study of Testing Validity and the Role of Smallwords in Contributing to Pupils' Fluency*. Cambridge: Cambridge University Press.

- Henning, G. (1987). *A guide to language testing: development, evaluation, research*. Cambridge: Newberry House Publishers.
- Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.
- Ingebrigtsen, A., Skarprud, T., & Stenslund, K. (2015). *Skolen som lærende organisasjon Hvordan bruker skolen sine kartleggingsresultater i arbeidet med å utvikle en lærende organisasjon?* (Masteroppgave) University of Agder. Retrieved from [https://brage.bibsys.no/xmlui/bitstream/handle/11250/299924/Anita Ingebrigtsen%2C Tor Skarprud og Karen Stenslund.pdf?sequence=1&isAllowed=y](https://brage.bibsys.no/xmlui/bitstream/handle/11250/299924/Anita%20Ingebrigtsen%20Tor%20Skarprud%20og%20Karen%20Stenslund.pdf?sequence=1&isAllowed=y)
- Kubiszyn, T., & Borich, G. D. (2006). *Educational Testing and Measurement: Classroom Application and Practice*. Hoboken: John Wiley & Sons.
- Kunnskapsdepartementet. (n.d.). Stortingsmelding 18. Retrieved from <http://www.regjeringen.no/templates/Underside.aspx?id=639547&epslanguage=NO-SE>
- Kunnskapsdepartementet. (n.d.). Stortingsmelding 19. Retrieved from <http://www.regjeringen.no/nb/dep/kd/dok/regpubl/stmeld/2009-2010/Meld-St-19-20092010.html?id=608020>
- Kunnskapsdepartementet. (n.d.). Stortingsmelding 30. Retrieved from <http://www.regjeringen.no/nb/dep/kd/dok/regpubl/stmeld/20032004/stmeld-nr-030-2003-2004-.html?id=404433>
- Kunnskapsdepartementet. (2015). *Kompetanse for kvalitet*. Retrieved from https://www.regjeringen.no/contentassets/731323c71aa34a51a6febdeb8d41f2e0/kd_kompetanse-for-kvalitet_web.pdf
- Lopez Mendoza, A. A., & Bernal Arandia, R. (2009). Language testing in Colombia: A call for more teacher education and teacher training in lanugage assessment. *Profile*, 11(2), 55–70.
- Luoma, S. (2004). *Assessing Speaking*. Cambridge: Cambridge University Press.
- Mackey, A., & Gass, S. M. (2005). *Second Language Research: Methodology and Design*. Mahwah: Lawrence Erlbaum Associates, Inc.
- McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.
- Mellingsæter, H. (2014). Advarer osloskolen mot å øve på kartleggingsprøver. *Aftenposten*. Retrieved from <http://www.aftenposten.no/osloby/Advarer-osloskolen-mot-a-ove-pa-kartleggingsprover-583519b.html>
- Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York: American Council on Education.

- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749.
- Moss, P. A., Girard, B. ., & Haniford, L. C. (2007). Validity in Educational Assessment. In J. Green & A. Luke (Eds.), *Rethinking Learning: What Counts as Learning and What Learning Counts* (pp. 109–162). SAGE Publications.
- Oslo Kommune/Utdanningsetaten. (n.d.). Prøver og Kartlegginger. Retrieved October 31, 2016, from <https://www.oslo.kommune.no/skole-og-utdanning/eksamen-og-elevvurdering/prover-og-kartlegginger/>
- Oslo Kommune/Utdanningsetaten. (2015). Rundskriv 3-2015 – Prøveplan Oslo – årshjul for skoleåret 2015 - 2016. Retrieved from <https://nordstrand.osloskolen.no/siteassets/nyheter/rundskriv-3-2015---proveplan-oslo---arshjul-for-skolearet-2015-2016.pdf>
- Rixon, S., & Council, B. (1999). *Young learners of English: Some research perspectives*. Longman.
- Rogaland Revisjon IKS. (2012). *Forvaltningsrevisjon av Skole-Oppfølging av Kartlegginger*. Stavanger.
- Utdanningsdirektoratet. (n.d.). Grunnlagsdokument Satsingen Vurdering for læring 2010 - 2014. Retrieved November 9, 2016, from [http://www.udir.no/PageFiles/35141/Grunnlagsdokument for satsingen Vurdering for læring okt 2011.pdf](http://www.udir.no/PageFiles/35141/Grunnlagsdokument%20for%20satsingen%20Vurdering%20for%20l%C3%A6ring%20okt%202011.pdf)
- Utdanningsdirektoratet. (2007). Rammeverk for kartleggingsprøver på 1.-4. trinn.
- Utdanningsdirektoratet. (2010). Lærerveiledning.
- Weigle, S. C. (2002). *Assessing Writing*. Cambridge: Cambridge University Press.
- William, D. (2011). *Embedded Formative Assessment*. Bloomington: Solution Tree Press.
- Yeh, J. P., Herman, J. L., & Rudner, L. M. (1981). *Teachers and Testing: A Survey of Test Use (CSE Report No.166)*. Retrieved from <http://eric.ed.gov/?id=ED218336>
- Øren Gjelsvik, J. (2012a). *Kartlegging – og så kva? Kva blir gjort for elevar som kjem under kritisk grense i leseferdigheit i 2.klasse?* (Masteroppgave) University of Oslo. Retrieved from <https://www.duo.uio.no/handle/10852/31491>
- Aarseth, I. M. (2013). *Fra bekymring til tiltak. En kvalitativ studie av skolens rutiner ved bekymring for elevens læringsutbytte*. (Masteroppgave) University of Oslo. Retrieved from <https://www.duo.uio.no/handle/10852/36576>