Analysis Phase in the Development of a Student Monitoring System for Physiotherapy Studies at Saxion Enschede.

Master Thesis
University of Twente
Faculty Behavioral Sciences
Educational Science and Technology
Heleen Hesselink - S1084119
August 2013

UNIVERSITY OF TWENTE.

SAXION

Acknowledgement

This master thesis is the result of my final project carried out to complete my master study Educational Science and Technology at the University of Twente. I would like to thank some people for their contributions and help that resulted in the completion of my thesis. Without their help it was not possible to carry out this research:

Mrs. Amy Lenderink-Hylton, for fulfilling the task of being my first mentor in which she coached and helped me in all kinds of ways during the last half year. For conducting the pilot interview, helping me conduct the interviews, provide me with literature and documents and all other help.

Dr. Bernard Veldkamp, for being my second mentor in which he coached me, helped me with the IRT analysis and my questions and giving feedback.

Dr. T. Eggen, for act as an extra mentor in assessing me and answering my questions about progress test via e-mail contact.

The students, teachers and mentors from physiotherapy studies at Saxion Enschede, for participating in the interviews.

Mrs. E. Laurens, expert in the English language, for giving feedback on the spelling and grammar of this thesis.

Dr. G. aan de Stegge, for providing the graphs from the CITO SMS.

Thank you.

Heleen Hesselink
Enschede, August 2013

Summary
Because of the alarming situation of the quality in higher education some measures have to be taken. One of these measures to improve quality is external validation in higher education and thereby the implementation of progress testing of basic theoretical knowledge. These tests should be developed in a bottom-up manner. So, schools should collaborate and invest in the development of an SMS. The study is aimed at the analysis phase of educational design research in which an SMS for physiotherapy studies at Saxion Enschede will be developed. Problem and context were analyzed by a literature review, document analysis and interviews with teachers, mentors and students from physiotherapy studies at Saxion Enschede. In general, the stakeholders were positive towards the implementation of progress testing and especially about the feedback that can be delivered with this SMS. Furthermore, existing tests are analyzed with IRT to select items and estimate the item parameters. These items can be used for developing the progress tests.

Samenvatting
Door alarmerende situaties over de kwaliteit van het HBO zullen er maatregelen moeten worden genomen. Een van die maatregelen om de kwaliteit te verhogen is externe validering in het hoger onderwijs en daarbij de invoering van voortgangstoetsen van de theoretische basiskennis. Deze zullen bottom-up ontwikkeld moeten worden, dus vanuit samenwerking van de scholen. Hogescholen zullen er dus niet aan ontkomen om een leerlingvolgsysteem te gaan ontwikkelen. Dit onderzoek richt zich op de analyse fase van het onderwijskundig ontwerpproces waarin een studentenvolgsysteem voor de opleiding fysiotherapie van Hogeschool Saxion Enschede zal worden ontwikkeld. Daarbij is het probleem en de context geanalyseerd door middel van literatuuronderzoek, document analyse en interviews met docenten, mentoren en studenten van de opleiding fysiotherapie van Saxion Enschede. Over het algemeen waren de betrokkenen positief tegenover de invoering van voortgangstoetsen en zeker over de gedetailleerde feedback die hierbij gegeven kan worden. Verder zijn bestaande toetsen geanalyseerd met IRT om items te selecteren en de parameters van deze items te berekenen. Deze items zouden gebruikt kunnen worden voor het ontwikkelen van de toetsen.

Abbreviations used

1PLM	1-Parameter Logistic Model
2PLM	2-Parameter Logistic Model
3PLM	3-Parameter Logistic Model
CAT	Computer Adaptive Testing
CITO	Central Institute for Test Development
CTT	Classical Test Theory
HBO	Hoger Beroepsonderwijs (Higher Education)
ICC	Item Characteristic Curve
IPTM	Interuniversity Progress Test Medicine
IRT	Item Response Theory
KNGF	Koninklijk Nederlands Genootschap Fysiotherapie (Royal Dutch Institute Physiotherapy)
NVAO	Nederlands Vlaams Accreditatie Organisatie (Dutch Flemish Accreditation Organisation)
SROF	StudieRichtingsOverleg Fysiotherapie (Study Consultation Physiotherapy)
SMS	Student Monitoring System

Table of contents

1. Introduction

This research presents the analysis phase in educational design research for the purpose of developing a Student Monitoring System (SMS) for physiotherapy studies at Saxion Enschede. In paragraph 1.1 the origins of this research are described. The second paragraph describes the aim of the research. Paragraph 1.3 gives a brief description on the research method.

1.1 Origins of the research

Various publications, news items and political debates in the Netherlands discuss the quality of higher education. There is concern and criticism on the quality of assessment in higher education. There are recent cases in which students unjustly received their diploma in higher education. In July 2011 the Dutch secretary reported in the strategic agenda the desirability of external validation in higher education to improve the quality of assessment (Ministerie OCW, 2011). As a result, in December 2011, the HBO Council designated a committee to investigate external validation in higher vocational education. This committee was asked to advise about various options of external validation. The committee discussed different options of external validation in their report (HBO Raad, 2012).

One of these options is to develop national tests in a bottom-up manner. The test would be developed within the organization of the participating schools. Several programmes of different colleges would jointly develop a test. The participating colleges would all supply a part of the test and in this way all participating colleges will be represented in the test. The committee recommended independent longitudinal testing: a series of tests over the whole study which tests the curriculum. In this way the students can be monitored in their theoretical knowledge level progress during their education.

An added value of this measure could be the study behavior of students. Standard domain assessment encourages superficial and rote learning (Blok, Otter & Roeleveld, 2012). With progress testing students are stimulated to learn in a more on-going basis (Hérnandez, 2012). Moreover, students receive feedback of the test results, so they can monitor their own learning process (Isaksson, 2007).

In the Netherlands, a student monitoring system (SMS) has already been implemented in primary education. Almost all primary schools use the SMS from the Central Institution for Test Development (CITO). With the SMS from CITO the teacher and school can monitor the students' progress and get an indication of their educational quality. Besides the monitoring and evaluation system for primary education, CITO also offers an SMS for students of the first two years of secondary education. However, this system is much less used. For vocational education, higher education and universities almost no SMS are known. Progress testing at university level has been implemented at medical schools at the University of Maastricht, Groningen, Leiden and Nijmegen: the Interuniversity Progress Test Medicine (IPTM). The IPTM is administered four times a year to monitor the development of knowledge of the students. The content of the progress test is based on the end terms of the curriculum. Therefore, it is almost impossible to learn test-aimed. It discourages superficial and rote learning. For students, it provides the opportunity to direct their learning process. Furthermore, it gives an indication of the quality of education for the participating universities. For higher education making use of a national SMS is (relatively) new.

1.2 Aim of the research

In the current educational environment it is inevitable that higher education schools will need to invest in the development of an SMS. The study will focus on developing an SMS for testing and monitoring the progress of the students' theoretical knowledge level.

Saxion Enschede, University of Applied Sciences, is initiating the process of the development of an SMS for physiotherapy studies. This research is carried out to investigate the possibilities of developing an SMS for physiotherapy studies at Saxion Enschede. This leads to the research question:

How should an SMS for theoretical knowledge level for physiotherapy studies at Saxion Enschede be developed?

To answer this research question several sub questions are established. These sub questions are:

1. What is the current situation in assessing theoretical knowledge and the feedback at physiotherapy studies at Saxion Enschede?

2. What is the desired situation in assessing theoretical knowledge and the feedback for physiotherapy students at Saxion Enschede?

3. How can item response theory be used in developing an SMS for physiotherapy studies at Saxion Enschede?

1.3 Research approach

This master thesis is part of a PhD from A. A. Lenderink-Hylton which will last about 5 years. Mrs. Lenderink-Hylton is a physiotherapy teacher at Saxion Enschede. In 2010 she finished her master EST at the University of Twente. The PhD is a research with the aim to develop a national SMS for physiotherapy studies.

This study forms the first step in educational design research. "Educational design research can be defined as a genre of research in which the iterative development of solutions to practical and complex educational problems also provides the context for empirical investigation, which yields theoretical understanding that can inform the work of others" (McKenney & Reeves, 2012, p. 7). McKenney and Reeves (2012) present the generic model with three core phases: analysis, design and evaluation (figure 1). The process is iterative and flexible. The model shows that the phases interact with practice and research. The two main outputs are the development of an intervention which matures over time and theoretical understanding. This study contributes, on scientific level, to the theoretical understanding of an SMS in higher education. On practical level, it contributes to the development of a national SMS for physiotherapy studies.

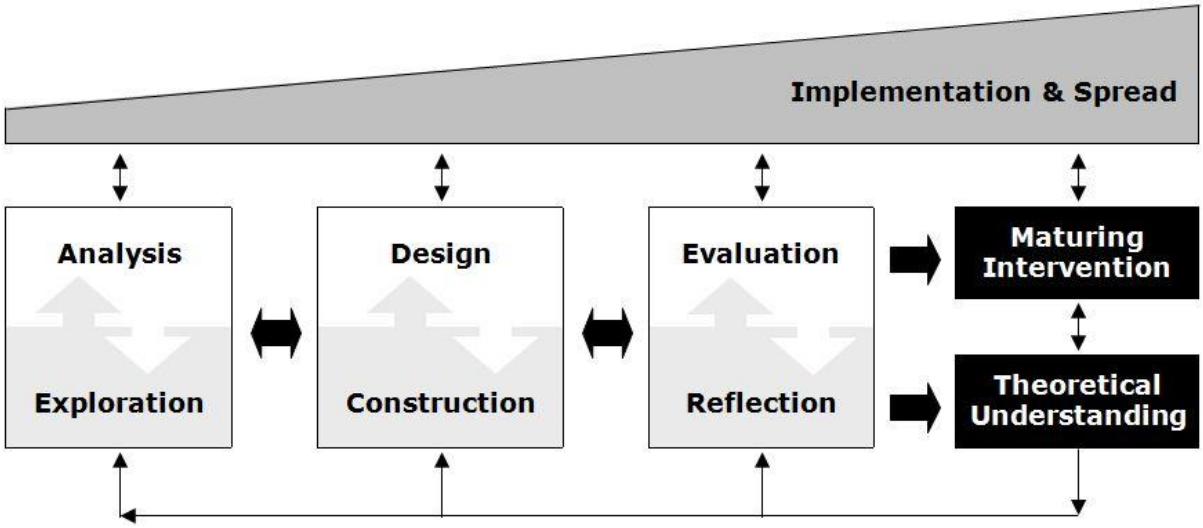


Figure 1. Generic model for conducting design research in education from McKenney & Reeves (2012).

The study focuses on the analysis phase of educational design research. The main goal is to explore the problem and context for the development of a national SMS for physiotherapy studies. The analysis contributes to the (theoretical) understanding of the problem and context. This study involves a literature review, document analysis, interviews with stakeholders and secondary data analysis.

2. Theoretical framework

In this chapter the theoretical background will be described, explained and deepened to set up a theoretical framework for this study. In the first paragraph the student monitoring system (SMS) is outlined in terms of definition, goals, preconditions and the use of item response theory (IRT) in an SMS. Paragraph 2.2 give two examples of an SMS: the SMS of CITO which is used in most primary schools in the Netherlands and the progress tests which are used for medical students at university level in the Netherlands. This paragraph also shows some examples of feedback retrieved from the systems: surveys & graphs. The third paragraph explains the role of feedback in the learning process of students.

2.1 Student Monitoring System

The quality of teaching must be continuously evaluated by teachers. Only if teachers know the level of the students they can adapt their instruction. Assessments and observations will give teachers an impression of the level of their students. But these impressions are often not objective, not systematic and not consistent over time. Also these results cannot be compared to the results in different grades. Therefore, it does not give information on the progress of a student. It is important to monitor students on a regular, reliable and systematic basis (Vlug, 1997). Monitoring development progress is an essential precondition for good teaching (Vlug, 1997). An SMS is an instrument which can be used to determine whether student development and teaching instruction are satisfactory (Vlug, 1997).

2.1.1 Definition

In 1988 the Dutch Primary Education Advisory Council (ARBO) reported the idea of developing an SMS. It was advised to schools to monitor and evaluate the educational quality. It was defined as "a concrete means for identifying discrepancies between progress and the targets and sub-targets set by the school and for registering this progress" (ARBO, 1988, p.64). An SMS can be characterized by several features. It consists of a coherent set of tests for longitudinal assessment used for continuous evaluation (ARBO, 1988; Gillijns, 1991; Janssens, as cited in Vlug, 1997; Kamphuis & Moelands, n.d.). It means that progress can be identified over time by testing at least a few times a year (Vlug, 1997).

2.1.2 Goals

An SMS serves several purposes in education. The primary goal is to monitor the development of students because an SMS provides data on the progress of the student (Gillijns, 1991; Glas & Geerlings,2009; Kamphuis & Moelands, n.d.; van der Drift, 1995; Vlug, 1997). Through the continuous collection of data, there can be decided whether a students does well compared to previous results, with fellow students or nationally (Gillijns, 1991; Glas & Geerlings, 2009; Kamphuis & Moelands, n.d.; Vlug, 1997). Whether educational objectives have been met can be monitored and also whether the subject matter fits the level of the student (Gillijns, 1991; Vlug, 1997). For early recognition and identification of any learning problems it also is of great importance (Kamphuis & Moelands, n.d.). Next to the information about the students, an SMS also provides information about the educational quality. The data can indicate whether the quality of education is sufficient and if there are indications for improving teacher's instruction (Glas & Geerlings, 2009; Kamphuis & Moelands, n.d.; van der Drift, 1995; Vlug, 1997)

2.1.3 Preconditions

An SMS has to meet certain quality criteria. Gillijns (1991) analyzed these different criteria into subtopics: content, technical measurement, registration and practical criteria.

First, the content criterion: an SMS should involve the entire education curriculum. For example in primary education an SMS should be used in all grades to monitor the progress. It has to be directed to the basic skills. Furthermore, it should not depend on content of the educational publishers but on the national curriculum. Therefore, another indirect precondition for developing an SMS are

longitudinal learning progression with concrete end- and sub-targets for the different subjects (HBO Raad, 2012; van Berkel & Bax, 2006; Vlug, 1997). For higher education it is also important that the study programmes are homogeneous: sharp specializations hinder successful implementation of an SMS (van Berkel & Bax, 2006).

The technical measurement criterion involves the standardized test to be reliable and valid. The test instruments must make use of a scaling technique so that longitudinal monitoring is possible. This can be done by using item response theory (IRT).

According to Gillijns (1991), a SMS must have a good registration system: the registration criterion. A good system is characterized by: high information value, clarity, simplicity and accessibility.

The last criterion for an SMS is the practical criteria. An SMS should administer the test for the whole group at once. The amount of time for administering and processing should be limited. An SMS must have the possibility to implement in phases.

Furthermore, it is important to note that an SMS should not depend on computers and software. However, this could be a useful tool. It has the advantage of registering and organizing the test data and it can mark tests and create surveys and reports (Vlug, 1997). Moreover, nowadays digital use is not desirable to be avoided.

2.1.4 Item Response Theory

The main purpose of an SMS is to follow student progress over the years. Therefore it is necessary that in an SMS the results from students on different test measuring the same ability must be put on the same scale to monitor student progress and compare with previous measurements (Kamphuis & Moelands, n.d.; Vlug, 1997). The measuring technique IRT can make this possible. It has become the mainstream theory in measurement issues. IRT presents a framework for constructing measurements, establishing validity measurements, estimating item and test characteristics, estimating abilities of individuals and spread of abilities in populations as well as providing a framework for interpreting test results (Kamphuis & Moelands, n.d.; Vlug, 1997).

IRT provides several advantages over the traditional Classical Test Theory (CTT) methods. The conclusions drawn with CTT are population dependent. With IRT there is no need for a random sample from a population (Kamphuis & Moelands, n.d.). Pilot tests determine the item characteristics. The item parameters can be estimated without making any assumptions about the distribution of the latent ability (Kamphuis & Moelands, n.d.).

IRT provides the possibility to separate the ability of a student (the chance that an item can be solved) and the item characteristics (e.g. the difficulty of an item) and put them on one scale (Glas & Geerlings, 2009). The ability score is expressed by the Greek letter theta (θ). The probability that the item will be answered correctly is expressed as P(θ). Figure 2 shows an example of an item characteristic curve (ICC). The probability of a correct response increases with the ability parameter.

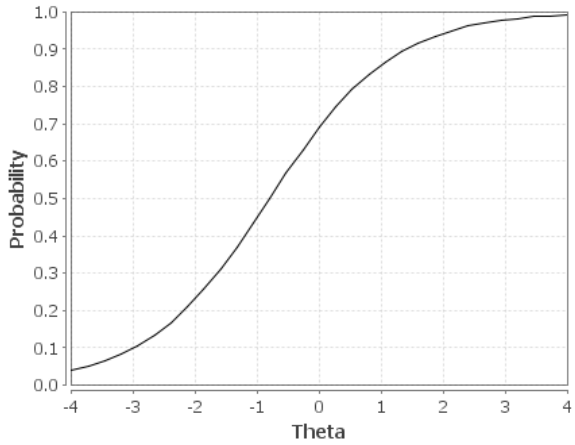


Figure 2. An example of an Item Characteristic Curve (ICC). Retrieved from: http://www.itemanalysis.com/sample-irt-plots.php

According to Vlug (1997), this parameter separation provides several advantages for an SMS. First, the test results of a student that differ in difficulty, content and number of items can be compared to previous test results of the same student. Second, the position of a student on the scale can be compared to that of other students of their grade, year or national. Third, the position on the scale provides information about the degree of mastering a particular subject. With all this information the teacher can indicate whether and in what domain learning problems occur.

Another advantage of IRT is the possibility of analyzing incomplete designs (Glas & Geerlings, 2009; Kamphuis & Moelands, n.d.). SMS make use of an item bank. An item bank is "a pool of test items where the item parameters are known through pretesting" (Glas & Geerlings, 2009, p. 85). Because of the size of the item pool that is needed in SMS, it is practically impossible to calibrate this pool by administering each student every item (Glas & Geerlings, 2009). With the possibility of analyzing incomplete designs the administration of items to persons is such that different groups of persons have responded to different sets of items (Glas & Geerlings, 2009). In these different tests, there are some common items. The design is linked to obtain parameter estimates on a common scale (Glas & Geerlings, 2009). Figure 3 depicts the principle of the linking pattern.

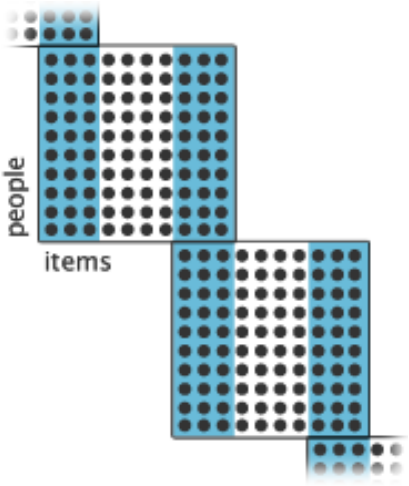


Figure 3. Diagnostic Linking Pattern. From: "Beating standardized tests with their own magic" by A. Tombari, 2009, retrieved from: http://grockit.com/blog/main/2009/07/08/beating-standardized-tests-with-their-own-magic/

IRT has also the advantage of implementing computer adaptive testing (CAT). CAT is administered by computer and the optimal test for an individual student is selected (Glas & Geerlings, 2009). The next test item to be administered to the student, depends on the previous responses of the student (Glas & Geerlings, 2009). "If the previous responses indicate that the student has a high ability, a difficult item will be administered. If, on the other hand, the previous responses seem to indicate a low ability, an easier item will be administered" (Glas & Geerlings, 2009, p. 83). This advantage is also demonstrated in the test called WISCAT. The WISCAT test is a math test for first year students from the teacher training college in the Netherlands. The test is developed by CITO. The test uses an item bank to compose the CAT.

Within IRT there are some models that are commonly applied in progress testing: the one parameter logistic model (1PLM), the two parameter logistic model (2PLM) and the three parameter logistic model (3PLM). When choosing an appropriate IRT model some considerations play a role, which are described by Glas and Geerlings (2009). The model should fit the data because only then the model is valid. Furthermore, the goal of the test and characteristics of the items in the test may determine the choice. For example, the 3PLM with guessing parameter is a good option when a test consists of multiple choice items. The SMS from CITO uses the 2PLM because this is less complex than the 3PLM and it provides the necessary and correct information.

There are also a few disadvantages to applying IRT. IRT is a complex method and it requires lots of time to construct and calibrate the results (Koornneef, 1991). Also the result of the test is hard

to estimate because of the need for large test samples (Koornneef, 1991). Furthermore, IRT is only applicable in subject matter areas which are homogeneous (Koornneef, 1991). For some subjects it is therefore necessary, that different domains be distinguished to meet this requirement. For example: CITO distinguishes in the subject matter area language in primary education different domains like technical reading, reading comprehension, and spelling.

2.1.5 Pros and cons

SMS have a lot of pros but also some cons. In this paragraph the advantages, disadvantages and side effects of using an SMS will be explained.

The longitudinal character of an SMS provides a unique opportunity for feedback. The growth of the students can be shown throughout their education programme (Boshuizen, van der Vleuten, Schmidt & Machiels-Bongaerts, 1997; Glas & Geerlings, 2009; Gillijns, 1991; van der Drift, 1995; van der Vleuten, Verwijnen & Wijnen, 1996; Vlug, 1997; Wrigley, van der Vleuten, Freeman & Muijtjens, 2012). In this way the progress of a student can be monitored. An SMS and its techniques provides an enormous source of information for feedback for students, teachers and the school for internal and external evaluation (Berkel, 1990; Glas & Geerlings, 2009; Gillijns, 1991; McHarg et al., 2005; van der Drift, 1995; van der Vleuten et al., 1996; Vlug, 1997). The standardized tests and IRT make it possible to compare students in different ways (Glas & Geerlings, 2009; van der Drift, 1995). Students can be compared studying previous results, comparing results to those of their fellow students or within different schools. What students already know can also be detected as well as what is unknown (Gillijns, 1991; van der Vleuten et al., 1996). The information of an SMS can be used for diagnosis and remedial teaching (Gillijns, 1991; Kamphuis & Moelands, n.d.; van der Drift, 1995; Vlug, 1997; Wrigley et al., 2012). Low and high achievers can be early detected (Kamphuis & Moelands, n.d.; van der Vleuten et al., 1996; Vlug, 1997). As a result of this informative feedback, teaching methods can be adjust and education can be improved (Kamphuis & Moelands, n.d.; Wrigley et al, 2012). Furthermore, the data can be used as an evaluation to the extent of which a school meets its curriculum objectives (Gillijns, 1991; van der Drift, 1995; van der Vleuten et al., 1996; Wrigley et al., 2012; Vlug, 1997). All this information and feedback provides also strong research potential at different levels (van der Vleuten et al., 1996).

An SMS is curriculum independent and breaks the relationship between the taught programme and assessment (Berkel, 1990; McHarg et al., 2005; Schuwirth & van der Vleuten, 2012). Course assessment within higher education, which is now mostly the case, can unconsciously stimulate cramming or learning for the test (van der Vleuten et al., 1996). With progress testing study behavior should change because it is difficult to prepare specifically for the progress test (van der Vleuten et al., 1996; Wade et al., 2012). Students are stimulate to learn on an on-going basis and develop high quality learning strategies required for their professional development (Berkel, 1990; Hérnandez, 2012; McHarg et al., 2005; Schuwirth & van der Vleuten, 2012; van der Vleuten et al., 1996; Wade et al., 2012). Continuously learning is rewarded (Berkel, 1990). A progress test can be based on the end terms of the curriculum. Therefore, it also helps students to become familiar with the level of knowledge expected at the end of the programme (McHarg et al., 2005). The character of the progress test ensures that the content of the whole curriculum is repeated (van der Vleuten et al., 1996). The characteristics of an SMS, repetitive and longitudinal, emphasizes long-term and functional knowledge (Schuwirth & van der Vleuten, 2012; van der Vleuten et al., 1996).

Longitudinal testing contributes to the reliability of making important decisions (Schuwirth & van der Vleuten, 2012). It is logic to assume that results of multiple measurements of continuous learning are more reliable than a one-shot method (Schuwirth & van der Vleuten, 2012; Wrigley et al., 2012). An SMS provides reliable information on the progress of individual students and groups of students (Glas & Geerlings, 2009).

An SMS consists of several tests a year. One bad test result over the year cannot undo a series of good results (Schuwirth & van der Vleuten, 2012). As a result, students will experience less examination stress (Schuwirth & van der Vleuten, 2012).

Yet, the importance and advantages of a SMS in education has been highlighted. However, there are also a few disadvantages of working with an SMS.

Researchers despite the amount of time that educators need to invest in an SMS (Gillijns, 1991; McHarg et al., 2005; van der Drift, 1995; Vlug, 1997). Teachers already have a busy job. The implementation of an SMS cost extra valuable time. The HBO council advised test development in a bottom-up manner (HBO Raad, 2012). This means that educators in higher education need to develop the tests. Furthermore, the upkeep of a question bank with administration, data analysis, review process, etc. is costly (McHarg et al., 2005). Implementing an SMS will cost lots of time, effort and money.

A danger of (progress) testing is the so called 'teaching to the test' (Gillijns, 1991; van der Drift, 1995; Vlug, 1997). That is that teachers adjust their lesson material to the content of the test to obtain higher test results. This could lead to impoverishment of the curriculum (van der Drift, 1995; Vlug, 1997).

Another limitation of the SMS is the content which is tested. Only basic subject knowledge content can be tested (Gillijns, 1991). The SMS reveals nothing about the characteristics of a student with regard to the motivation, effort, etc. Progress tests do not measure the competences and performance in practice of students (Muijtjens, Schuwirth, Cohen-Schotanus, Thoben & van der Vleuten, 2008). Therefore, the progress test should not be the only assessment tool in education. It should be complemented with other assessment forms. Furthermore, regarding the content, a precondition for developing an SMS is longitudinal learning progression (Vlug, 1997). Higher education should be thoughtful about this and develop or secure learning progression and concrete end- and sub-targets for different subjects or sub-topics. Moreover, "only a limited part of the test reflects the content of the first year curriculum. Therefore, the progress test may be less sensitive for achievement in the freshman year" (van der Vleuten et al., 1996, p. 106). Educators should also be aware for of potential demoralization for freshman (McHarg et al., 2005). The tests cover the entire curriculum so first year students are exposed to very difficult questions which can cause anxiety.

Last, teachers should be aware of the performance anxiety of their students (Gillijns, 1991; van der Drift, 1996). This stress could lead to undesired results. Although, as mentioned earlier, one bad test result over the year cannot undo a series of good results in an SMS, so students might experience less stress from multiple progress tests a year.

2.2 Examples of Student Monitoring Systems

Using SMS and longitudinal testing in education in the Netherlands is known since decades. The best known is the SMS for primary education from CITO in the Netherlands. Almost all primary schools use this SMS but it is not mandatory. Schools can also use other systems or data to monitor their students. Furthermore, CITO also developed SMS for secondary education in the Netherlands but these are less used. The Interuniversity Progress Test Medicine (IPTM) is another well-known SMS in the Netherlands at university level. These cases will be explained in the next paragraphs.

2.2.1 Student Monitoring System for primary education from CITO

In the late eighties CITO started developing an SMS for primary education. The motivation was to make use of standardized tests as they did in the USA (van der Drift, 1995). The monitoring and evaluation system of CITO is developed to monitor students development in a number of meaningful ways: in relation to both personal and peer development, at given moments and over time during their primary school career (CITO, n.d.). The system consists of standardized tests, a registration system and material for analysis and remedial teaching of low achievers (Vlug, 1997). The practical tool was developed to help teachers obtain reliable data from their pupils progress in a systematic way (Vlug, 1997). If the data indicate that the learner is not performing well, further analyses have to be done and remedial actions will have to be taken. The SMS provides teachers and schools with answers to their everyday questions as: how much progress do my students make? What can I do to improve the results of my students? Which students need extra help? Do I have to adjust my teaching methods? So, the system allows teachers to determine if their students educational progress is satisfactory and to adapt education to the needs of the individual learner (CITO, n.d.).

During all grades of primary school students are tested twice a year with the SMS of CITO: half way the school year and at the end of the school year. The SMS contains different subject matter

areas such as mathematics and language. The content is based on the national curriculum. The results of the test can be processed manually or with help of an integrated computer program. With the data extensive and advanced reports of the progress can be generated. Reports can be made for each individual learner, for each grade, for the school as a whole and reports across schools. So, the system can be used to provide information on learning outcomes to all stakeholders.

The SMS from CITO does not require a computer although there are many software programs available to register the results of the tests. In addition, it offers advantages for teachers and schools. The software can mark tests and create surveys and reports. The system organizes the data which adds to the effectiveness of working (Vlug, 1997). CITO has already started implementing digital test, which may also be CAT (CITO, n.d.). This has several advantages. It means that testing time is shorter, the test is related to the students ability and for the teacher the workload is reduced because the data can be analyzed and registered by the system (CITO, n.d.).

The SMS of CITO consists of three phases: identification, analysis and action (van der Drift, 1995; Vlug, 1997). The identification phase consists of administering the test and processing and registering the data (van der Drift, 1995; Vlug, 1997). On basis of the test results the teacher selects students. The analysis phase is to explain the shortfalls and gather more data on these selected students (van der Drift, 1995; Vlug, 1997). The teacher makes a remedial plan for the students and implements it: action phase (van der Drift, 1995; Vlug, 1997).

2.2.2 Feedback from the Student Monitoring System for primary education from CITO

The main purpose of an SMS is to monitor the development of students. The results of the test can indicate if a student makes progress or falls behind. This can be shown in several surveys and reports that CITO uses as feedback. CITO uses IRT to analyze the data from the students. This has several advantages for the information and feedback that can be provided about the learning outcomes. Advanced reports on micro, meso and macro (pupil, class and school) level can be made. Computer software can automatically generate different surveys and reports like pupil reports, group surveys, cross sections, etc.

It should be noted that interpreting the test result is not an easy task for teachers. Users of the SMS have difficulties in interpreting the test results (Meijer, Ledoux & Elshof, 2011). This sometimes leads to making incorrect decisions. Teachers often do not have enough knowledge and skills to interpret the results correctly. The SMS from CITO is a reliable and valid tool to measure student ability. However, when users draw incorrect conclusions, the validity is negatively affected.

2.2.2.1 Feedback at micro level

The aim of the SMS is to monitor the development of students. Therefore, the feedback from the results of the test is aimed to depict this progress. The SMS of CITO offers the opportunity to depict the ability score of the pupil on a scale and to depict the progress in a pupil report.

IRT makes it possible to bring the ability of a student and the item characteristics on one scale. Because this can be depicted on the same fixed scale it is easy to conclude the degree of mastering subject matter of a student. The test score has to be converted into an ability score by looking it up in the table from the SMS. The ability score on the ability scale indicates the level of competence. By using the ability scale the results from different test of the same subject over the years can be compared. In this way the progress of a pupil can be monitored over a number of years. Figure 4 shows a part of the ability scale for mathematics. For example: at the end of grade 4 Thomas has an ability score of 80 which means he has the ability to subtract numbers under 10. He has not yet the ability to subtract numbers above 10.

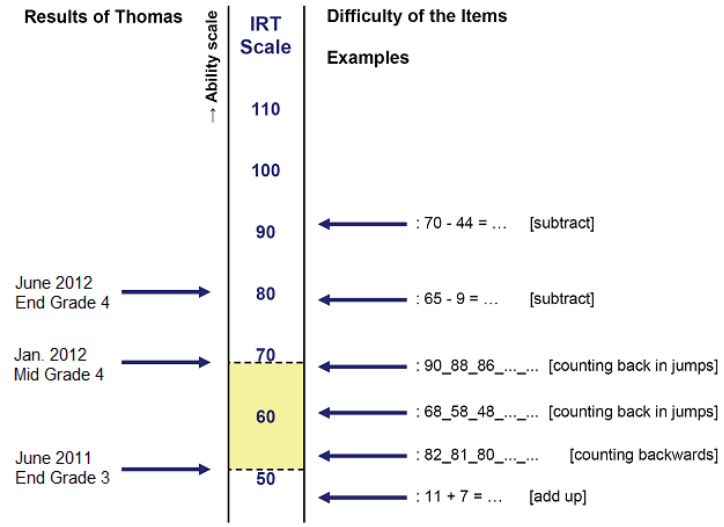


Figure 4. Example of an ability scale.

The pupil report is a graph in which the ability score of the student over the years can be compared to the national norm. The pupil report gives the opportunity to see how a student has developed over the years. The student's result can be interpreted even better when compared to the national norm. The national norm is distinguished into five levels:
Level A: the best 25% of the students
Level B: the 25% students who score above average
Level C: the 25% students who score below the average
Level D: the 15% students who score far below average
Level E: the lowest 10% of the students

Teachers often interpreted level C as average, while, in fact, this should be below average. Because the interpretation of the levels A – E was often wrongly interpreted, CITO recently came up with levels I – V. Levels I – V are classified by 20% for each level. In this case level III is indeed the average. The levels are depicted in figure 5.

Figure 6 shows the pupil report from a pupil from grade 4 to grade 8 for mathematics. The raw test score is converted into an ability score. This ability score is expand in the graph. The line refers to the level of the student over time. This can be compared to the national norm. The white line is the national average.

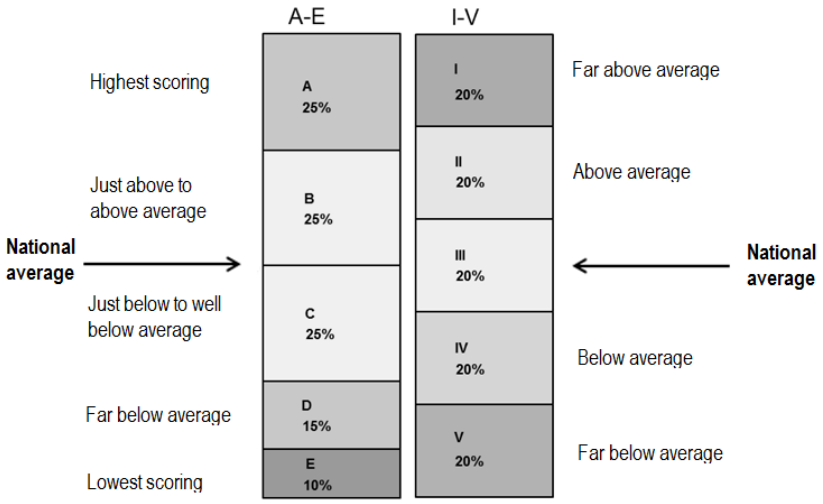
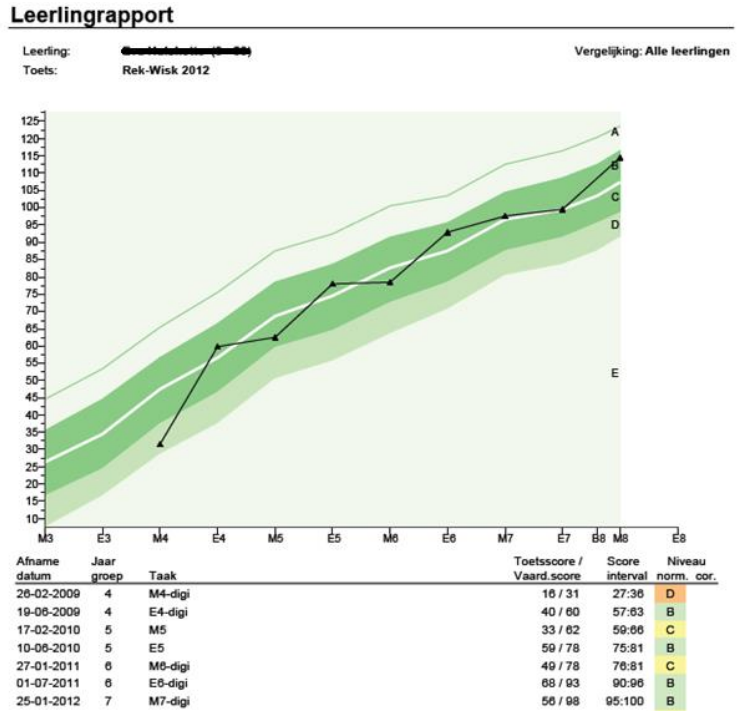


Figure 5. Level indicators.



Leerlingrapport

Leerling: [redacted]
Toets: Rek-Wisk 2012
Vergelijking: Alle leerlingen

Afname datum	Jaar groep	Taak	Toetsscore / Vaard.score	Score interval	Niveau norm.	cor.
26-02-2009	4	M4-digi	16 / 31	27:36	D	
19-06-2009	4	E4-digi	40 / 60	57:63	B	
17-02-2010	5	M5	33 / 62	59:66	C	
10-06-2010	5	E5	59 / 78	75:81	B	
27-01-2011	6	M6-digi	49 / 78	76:81	C	
01-07-2011	6	E6-digi	68 / 93	90:96	B	
25-01-2012	7	M7-digi	56 / 98	95:100	B	

Figure 6. Example of a pupil report. Retrieved from the SMS from CITO.

2.2.2.2 Feedback at meso level

As a teacher it is important to know the level of individual students but also of the group as a whole. The feedback from CITO offers the opportunity to gain also group surveys, group reports and an ability growth report.

A group survey shows the results of the students of one group. It can show the results of different subject test in one year or the results from one subject matter over a number of years. Figure 7 shows a group survey of grade 6 for three tests at one moment: reading comprehension, mathematics and spelling. It shows the ability score and the level for every student.

Groepsoverzicht - afnamemoment

Afnamemoment: Medio 2011-2012 Vergelijking: Alle leerlingen
Groep: 6 - 6B

	Begrijpend lezen 2012		Rek-Wisk 2012		Spelling 2012	
	Score	Niveau	Score	Niveau	Score	Niveau
[redacted] Berns	54	I+	77	IV	137	II
[redacted] Cortenraad	32	III	66	V	141	I
[redacted] Gaertner	33	III	85	III	138	II
[redacted] Haas	33	III	80	III	139	I
[redacted] Hardin	18	V	69	III → V	128	IV
[redacted] Hashad	42	II	76	IV	129	IV
[redacted] Hoemoez	9	V-	70	IV → IV	132	III
[redacted] Huisman	45	I	70	IV	144	I+
[redacted] Jetten	38	II	90	II	132	III
[redacted] Kallen			119	I+	139	I
[redacted] Wiertz	39	II	100	I	135	III
Aantal leerlingen	10		11		11	
Gemiddeld	34,3	III	82,2	III	136,0	I+

Figure 7. Example of a group survey. Retrieved from the SMS from CITO.

An ability growth report shows the growth in ability between two tests from one subject. This can be seen in figure 8. For each moment and subject matter area the expected growth can differ. In this example the expected growth in ability is 21 for the national average.

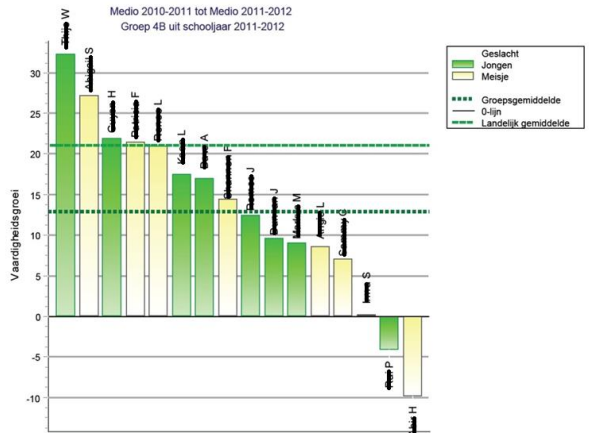


Figure 8. Example of an ability growth report. Retrieved from the SMS from CITO.

2.2.2.3 Feedback at macro level

CITO also offers feedback at school level to evaluate the results and education of a school. This is presented by a cross section and a trend analysis.

The cross section shows the distribution of pupils over the different levels (A-E) for the tested grades of one moment. Figure 9 shows an example of a cross section for mathematics for grade 3 to grade 8. The 0% line indicates the national mean. In this example the grade 3 pupils are performing above the national mean, because about 75% of the pupils in this grade have an A or B level instead of the national reference group of 50%.

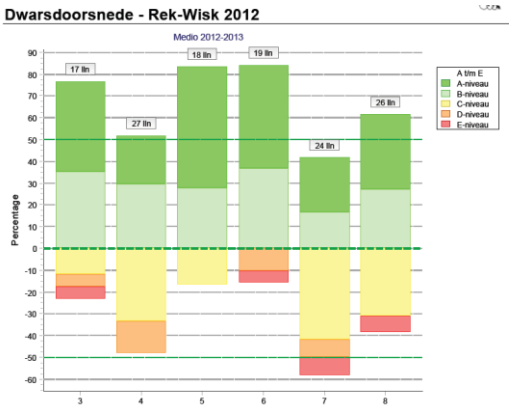


Figure 9. Example of a cross section. Retrieved from the SMS from CITO.

Trend analysis can be compiled for cohorts or grades. The trend analysis shows the test results followed over the years. This feedback can be used by the school to get an impression on the effectiveness of the school. Figure 10 shows an example of a trend analysis of cohorts.

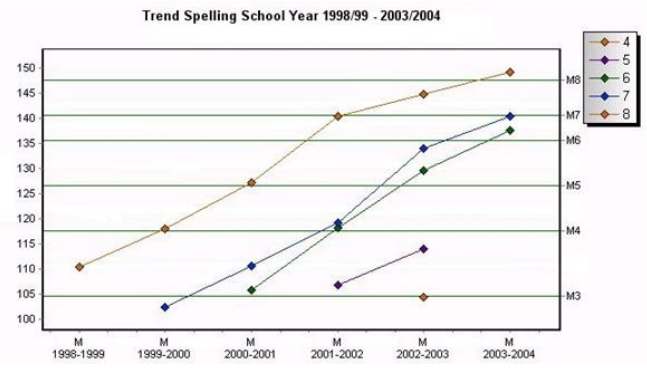


Figure 10. Example of a trend analysis. Retrieved from the SMS from CITO.

2.2.3 Interuniversity Progress Test Medicine at University Level

Another well-known SMS in the Netherlands is the Interuniversity Progress Test Medicine (IPTM). The IPTM was developed in the eighties to stimulate self-directed, enquiry based learning to enable students to develop the high quality learning strategies required for continuing professional development and revalidation (Van der Vleuten et al., 1996; Wade et al., 2012). According to McHarg et al. (2005), progress testing is a form of longitudinal examination which samples at regular intervals from the complete curriculum.

The IPTM is an instrument to measure the development of medical students' knowledge in the Netherlands. The measurement instrument consists of four progression tests a year, because only then progress can be monitored. In the Netherlands, every test is administered by circa 8000 students from four different universities at one point in time. The tests assess students' knowledge at graduation level. Students from all years take the test. Students should score higher every time they take the test. The test consists of approximately 200 true/false items with a 'don't know' option. Nowadays, some question are also multiple-choice questions with three or four options to answer (T. Eggen, personal communication, June 4, 2013). The scores of the progress test lead to a qualification as: insufficient, sufficient or good. It is up to the university to decide to award the students with credits for passing the test. However, the off-score is the same at all universities (HBO Raad, 2012).

This SMS provides longitudinal monitoring of the development of knowledge of the students. The SMS can determine whether a student reached the learning goals of that time. The content of the progress test is based on the end terms of the curriculum. Therefore, it is almost impossible to learn test-aimed. It discourages superficial and rote learning (Van der Vleuten et al., 1996; Wade et al., 2012).

The test is developed by collaborating universities: Maastricht, Groningen, Leiden and Nijmegen. They work together to construct and review items and thereby set up an item bank (IPTM, n.d.). The collaboration provides opportunities to lower the work pressure from constructing and controlling the items (McHarg et al, 2005). Furthermore, they can share knowledge, resources and ideas, assuring high quality of assessment, supporting innovative assessment formats and development of examinations.

The items are constructed by professors of the collaborating universities. The item is described with a literary reference to control the content and for the purpose of studying. The item is controlled by a committee before ending up in the item bank. The item bank consists of lots of items which can be used for the progress test. If an item is used, it will be blocked for the next three years. Items will be reassessed to keep the item bank up to date. Professors are responsible for constructing items. The committee is responsible for the selection of items in the progress test. They have to make sure that the items in the blueprint are not related. Next to this they need to make sure that the difficulty remains about the same in every test. (IPTM, n.d.)

The IPTM was analyzed by using CTT. Currently there is a pilot to investigate the use of IRT in the IPTM and to investigate the possibilities of CAT (T. Eggen, personal communication, June 4, 2013).

The IPTM provides several advantages. It is programme independent (Berkel, 1990; Schuwirth & van der Vleuten, 2012). In this way, universities can collaborate in developing the tests (Schuwirth & van der Vleuten, 2012). Also, because it is curriculum independent, merely studying for the test is undesirable and continuous learning will be promoted (Berkel, 1990; McHarg et al., 2005; Schuwirth & van der Vleuten, 2012). It has a positive influence on student learning and this is actually the reason why progress testing was originally developed (Schuwirth & van der Vleuten, 2012). Students can determine their individual position to the final level and they get familiar with the level of knowledge expected at the end of the programme (Berkel 1990; McHarg et al., 2005). Also it is assumed that progress testing reduce examination stress, because one bad result cannot undo a series of good results (Schuwirth & van der Vleuten, 2012). Another advantage of progress testing is, that it provides detailed feedback for students, mentors, lecturers and universities for internal and external evaluation (Berkel, 1990; McHarg et al., 2005; van der Vleuten et al., 1996). In this way, students can direct their own learning process (Schuwirth & van der Vleuten, 2012). Finally, according to Schuwirth and van der Vleuten (2012), progress testing contributes to the validity and reliability. Longitudinal data is more predictive of future performance than one-off measurements. The combination of results adds to the reliability of a decision.

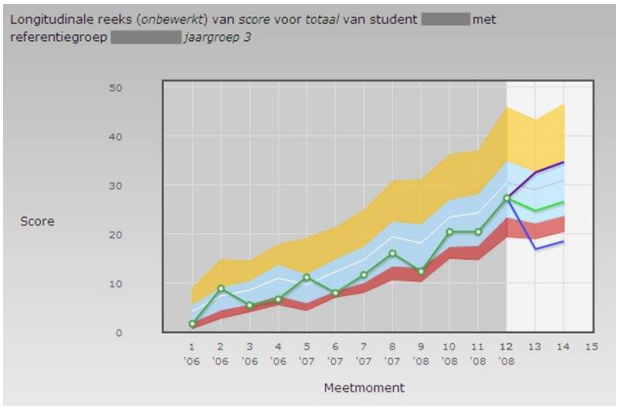
In spite of the advantages, there are also some disadvantages in progress testing. It is labour-intensive and costly (McHarg et al., 2005). The upkeep of an item bank, administration, data analysis etc. will cost a lot of time and money. Furthermore, progress testing can be a potential demoralization for freshmen (McHarg et al., 2005). Students in the first year are only able to answer a small percentage of the questions and these questions are relatively difficult for them.

2.2.4 Feedback from the Interuniversity Progress Test Medicine at University Level

Next to the purpose of an assessment instrument, the progress test is also a means for feedback (Berkel, 1990; Boshuizen et al., 1997). The feedback from the IPTM is delivered to students and staff by the web-based feedback system ProF (Muijtjens et al., 2008). The feedback consists of quantitative data on the overall test scores and subtest scores. The test is divided into different subtests or domains. The information retrieved from ProF can be momentaneous, longitudinal or predictive. Momentaneous information is the information from a specific test. Longitudinal information is the development of scores over a series of test. Predictive information is the predicted score on a future test. The system provided feedback in the form of graphs on micro, meso and macro (student, group and university) level.

2.2.4.1 Feedback at micro level

By providing feedback for the student, the student can direct the own learning process. For example they can detect their strengths and weaknesses by the information from the tests. In consultation with other students or with a mentor they can find out what causes it and how it should be improved. Figure 11 shows an example of a longitudinal series of test scores. This graph can also be made for a longitudinal series of subtopic scores (e.g. anatomy). Figure 12 shows an example of the scores on different subtopics of one test. It also shows where the student is compared to the national average.



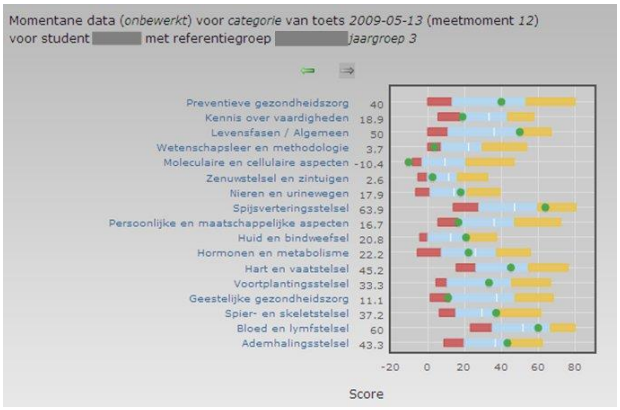
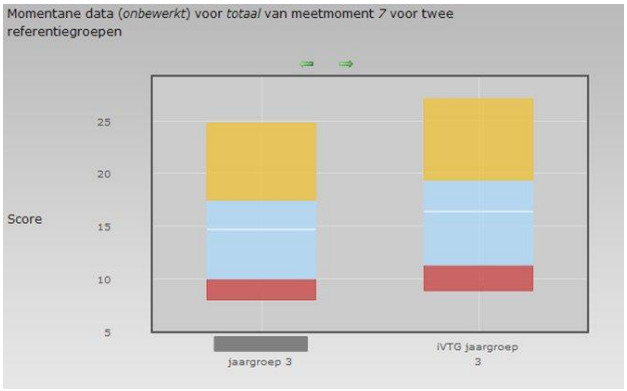


Figure 11 (left). Example of a longitudinal series of test scores from one student. Retrieved from: http://www.ivtg.nl/
Figure 12 (right). Example of the subtopic scores deposited the national average from one student. Retrieved from: http://www.ivtg.nl/

2.2.4.2 Feedback at macro and meso level

Teachers can also use the results to find out how students performed in their domains. If there are serious weaknesses in the development of scores, a teacher can decide to look for possible causes and how these might be remedied. The students of a certain university will be compared to students of another university. Figure 13 shows an example of two distributions: one university compared to the others. Figure 14 depicts a longitudinal series of tests from one university compared to the other universities.



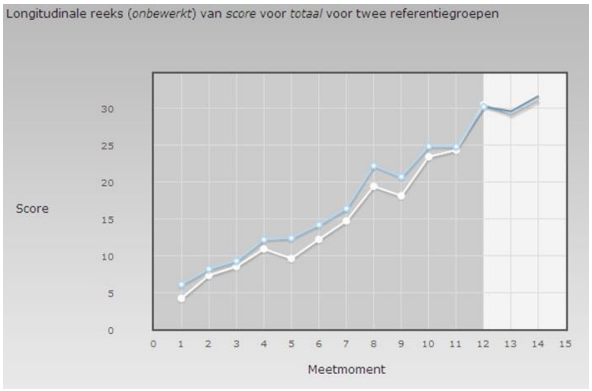


Figure 13 (left). Example of comparable distributions between two schools. Retrieved from: http://www.ivtg.nl/
Figure 14 (right). Example of longitudinal series of tests compared between two schools. Retrieved from: http://www.ivtg.nl/

2.3 Feedback in the Learning Process of Students

One reason to assess students can be to provide students and their teachers with feedback (Brown, 1999; van Berkel & Bax, 2006). Providing feedback to students is an important role of teachers in higher education (Irons, 2008). This feedback can be used in the learning process (van Berkel & Bax, 2006). According to Gibbs (1999), assessment is the most powerful tool for teachers to influence student learning.

Not all data is information and feedback. In an SMS data has to be interpreted to be transformed into information (Davenport & Prusak, 1998). Only by correct interpretation this information can be used as feedback. Feedback is "information provided by an agent regarding aspects of one's performance or understanding" (Hattie & Timperley, 2007, p.81). The purpose of feedback is to close the gap between a student's current performance and the intended learning outcomes (Hattie & Timperley, 2007; Irons, 2008). Evans (2013) distinguished assessment feedback and defines it as "all feedback exchanges generated within assessment design, occurring within and beyond the immediate learning context, being overt or covert (actively and/or passively sought and/or received), and importantly, drawing from a range of sources (Evans, 2013, p. 71). Reviews from Black and William (1998) and Hattie and Timperley (2007) stated that feedback is arguably the most critical element in facilitating students' learning.

In higher education, assessments mostly take place at the end of a course. After the course is passed, the knowledge from this course is usually not tested again. Students often only receive feedback on how well they have mastered the subject matter in the form of a grade. This has the consequence that students probably are focused on passing the test but not in a way that will help them remember the knowledge for a long time (Schaap, Schmidt & Verkoeijen, 2012). Students use undesirable learning strategies as 'learning for the test' or 'cramming' (Van der Vleuten et al., 1996). Course assessment makes it also difficult to use the feedback from assessment for future learning (Johnsson, 2012). Continuous assessment, such as implementing an SMS, provides opportunities to get feedback on learning and to adjust learning and teaching methods (Hérnandez, 2012; Isaksson, 2007; van Berkel & Bax, 2006). Then, the feedback has to be specific and detailed (van Berkel & Bax, 2006). It also can be seen as a feed forward function: progress tests can be used to influence student learning (van Berkel & Bax, 2006). They can be made aware of the topics and the intended level (van Berkel & Bax, 2006).

The feedback deriving from assessment, should not focus on grading. Grading is a less effective form of feedback than providing qualitative feedback (Blok, Otter & Roeleveld, 2002; Irons, 2008). Grading can cause a conflict between students' preferences of receiving informative feedback and their actual use of feedback (Johnsson, 2012). However, students claim to appreciate grades (Johnsson, 2012), grading can have a negative effective on the self-esteem of low performing students.

On the other hand, high performing students do not look at the feedback when they are satisfied with the grade awarded (Johnsson, 2012).

Feedback from assessment should not focus on grading but it has to be informative feedback. Students can learn from assessment if it provides feedback on how and what they understand and misunderstand (Blok, Otter & Roeleveld, 2002; Hattie & Timperley, 2007). Hattie and Timperley (2007) argue, that effective feedback must answer three major questions: Where am I going? (feed up), how am I going? (feedback) and where to go next? (feed forward). The feedback from an SMS should provide information on the weak and strong points of a student. For example: the feedback from IPTM provides information on the weak and strong domains of a test. Only then students and their teachers can close the gap between the current performance and the desired performance.

Feedback needs to be useful to students. Students prefer critical, detailed and meaningful feedback that can be used for future learning (Havnes, Smith, Dysthe & Ludvigsen, 2012; Johnsson, 2012). From a research of Huxham (2007) regarding students perspective to the feedback in higher education, students complain most about the technicalities of feedback, including content, organization of assessment activities, timing, and lack of clarity about requirements. Students appreciate the personal communication with teachers about their learning (Havnes et al., 2012). The teacher-student relationship can influence how students relate to feedback (Havnes et al., 2012).

Feedback must be used by the students to be effective (Johnsson, 2012). Students can value receiving the feedback, but they also have to act on feedback to adjust and improve their learning (Johnsson, 2012). After students accepted the feedback they have to use it carefully to improve learning (Evans, 2013; Hattie & Timperley, 2007). Self-regulated students can use the feedback to adjust their learning. Self-regulated learning involves complex learning skills like using cognitive strategies; make use of metacognitive strategies and showing engagement (Boekaerts, 1999). Self-regulated learners can apply effective strategies as increase effort or seek better learning strategies (Hattie & Timperley, 2007). Ideally, students would increase effort to reach goals or set more challenging goals when the goals are already reached. However, students also use less effective strategies to reduce the gap between the current performance and the desired performance and are less effective: abandon goal which can lead to non-engagement, lowering goals, accepting lower performance, etc. (Hattie & Timperley, 2007). Also low-qualitative feedback can be an explanation to why students do not use the feedback they receive (Johnsson, 2012). Furthermore, students may not know how to apply the feedback in their future learning (Johnsson, 2012). Students lack strategies for productive use of feedback or they may have problems understanding the meaning of the feedback (Johnsson, 2012). Students need to be open to the feedback and know what to do with it (Johnsson, 2012). They need self-regulating skills in order to react upon the feedback (Evans, 2013; Johnsson, 2012).Therefore, to adjust and improve learning with the received feedback requires a lot from students (and their teachers). Notwithstanding its complexity, feedback is a first step in helping students improve on their learning.

Blok, Otter and Roeleveld (2002) pointed out that many teachers underestimate the efficacy of providing their students with feedback. Teachers do not believe that their students really want to receive informative feedback (Havnes et al., 2012). They think that students are only interested in grades. They find their feedback useful but are worried about students not making use of or acting upon feedback (Havnes et al., 2012; Huxham, 2007). Furthermore, assessment feedback is, by teachers, too often seen as making statements about students and not about their teaching (Irons, 2008; Timperley & Wiseman, as cited in Hattie & Timperley, 2007). Teachers can also use feedback results to improve their teaching methods (Irons, 2008). They can consider adapting teaching practice or consider what needs to be taught differently (Irons, 2008).

In higher education feedback is crucial to facilitate students development as independent learners who are able to monitor, evaluate, and regulate their own learning (Evans, 2013). Teachers in higher education want their students to expand their knowledge during their study and even afterwards (Schaap et al., 2012). This professional development requires high quality learning strategies (Wade et al., 2012). Wade et al. (2012) describe the origination from progress testing in medical education. For continuing professional development in medical education high quality learning strategies are required. This must be encouraged by self-directed, enquiry based learning. Along with this change in medical education also the assessment programmes needed to support this approach. Superficial last minute learning had to be avoided (Van der Vleuten et al., 1996). Longitudinal examinations were

developed to support learning on an on-going basis. With progress testing you stimulate students to learn in a different way (Hérnandez, 2012). Students believe that the progress test and the feedback from it helps them to motivate and self-direct their learning (Hérnandez, 2012; Wade et al., 2012).

It is too easy stated that feedback can improve test results. Feedback alone does not improve results (Evans, 2013). Education and learning are complex issues and it is not easily stated which factors improve learning (Evans, 2013; Johnsson, 2012). The factors influencing the power of feedback are very complex and unpredictable (Johnsson, 2012). Nevertheless, providing qualitative feedback can help students and teachers to adjust learning and teaching and this can improve learning outcomes.

3. Context

In this chapter the context will be described. The context and views of stakeholders will be described at macro, meso and micro level. The content of this chapter was realized by document analysis, internet search and field-based investigation. Collaboration with practitioners helps to better understand the context and involving stakeholders can be positive for the change (McKenney & Reeves, 2012; Fullan, 2007)

3.1 Context at macro level

At macro level the context will be described for the Ministry of Education, Culture and Science even as the HBO Council. They reported important aspects of the improvement of quality of higher education in reports.

3.1.1 Strategic Agenda of the Ministry of Education, Culture and Science

The Ministry of Education, Culture and Science of the Netherlands reported in 2011 the strategic agenda. The main subject of this agenda is quality improvement of universities of applied sciences because experts think that the current situation is not future-proof (Ministerie OCW, 2011). The quality has to be improved in several ways. For example: students have to be challenged more by offering honors programmes.

Society should have no doubts regarding the diploma of students from universities of applied sciences (Ministerie OCW, 2011). Unfortunately, the inspectorate and the NVAO ('Nederlands Vlaams Accreditatie Organisatie') did some research and concluded, that in some educational programmes the examination situation is alarming (Ministerie OCW, 2011). The NVAO is an organization which secures the quality of higher education. To prevent wrong examination three measurements are described in the strategic agenda: better control on quality by accreditation and inspection, external validation of assessment and examination and better governance.

The universities should improve the knowledge level and assessment and examination should be validated externally (Minsterie OCW, 2011). Their advice is to develop national progress tests (Ministerie OCW, 2011). They refer to the HBO Council who complete this measurement by an investigation and report.

The colleges have great autonomy. The Ministry thinks this should be maintained but with this autonomy also comes accountability (Ministerie OCW, 2011). The colleges have to take responsibility for the quality of their education. The improvement of quality should be developed from inside the organizational institutes and so it requires effort of all stakeholders (Ministerie OCW, 2011).

3.1.2 HBO Council

As mentioned in the origins of the research (paragraph 1.1) the HBO Council set up committee to investigate external validation in higher vocational education. Commissioned by the Ministry they advised about various options of external validation. To monitor the theoretical knowledge level of students in higher education, the committee suggested to develop and implement longitudinal tests. The main goal of the measure is to validate assessment and to strengthen the theoretical knowledge level of students (HBO Raad, 2012).

The SMS has to be developed bottom-up (HBO Raad, 2012). They want schools to jointly develop the tests. Top-down testing, in which the test will be developed by an external organization, is not recommended (HBO Raad, 2012). It will not contribute to the expertise of testing by teachers (HBO Raad, 2012). Furthermore, it causes a lack of support by stakeholders and it is very costly to develop the tests in a top-down manner (HBO Raad, 2012). So, the bottom-up development has to make sure that several programmes of different colleges would jointly develop national progress tests like in the example of the IPTM. The HBO Council (2012) recommend to collaborate with three or more universities of applied sciences where possible.

The SMS can be developed next to the existing tests. They could remain the same so that every study programme has still the possibility of setting its own specialism. However, existing test could also disappear when there is an overlap (HBO Raad, 2012). There should and probably will be

enough common trunk to test nationally. The colleges already should have developed a national educational curriculum. Every college can set its own profile within it (HBO Raad, 2012). Physiotherapy studies developed a national transcript in which the core subjects and competences are described (HBO Raad, 2012).

The HBO Council (2012) thinks the progress test should be held one to four times a study year. The test would be taken by all students of all years of all colleges at the same time. They advise no retakes. Every next progress test is also the retake exam.

The HBO Council (2012) is discussing about awarding credits to students for the progress test. They leave it up to the schools to decide about the dividing of credits. However, they describe that in the implementation phase, schools have to be careful with awarding credits, with regard to the limited experience and adaptation period. On the other hand, experience tells us, that students and teachers have to be motivated to invest time in the progress test (HBO Raad, 2012). As a guideline, the HBO Council (2012) advises to award students with at least 4 credits for progress testing at each study year.

3.2 Context at meso level

At meso level the context will be described for Saxion: university of applied sciences. This paragraph also involves information on physiotherapeutic organizations and physiotherapy studies. These organizations monitor the quality of the profession and the quality of the physiotherapy studies.

3.2.1 Saxion

Saxion is an university of applied sciences and is located in three cities: Enschede, Deventer and Apeldoorn. Saxion provides over 60 different studies. These studies are divided in the 12 academies Saxion offers. Examples of academies are: school of health, hospitality business school and school of governance and law. The institution has about 25,000 students.

Saxion wants to provide high quality, innovative and practical education. They want their students to be well prepared for their job. Next to the education Saxion offers, they also are active in conducting research. Different research groups provide new knowledge and try to process this in education methods.

The motto of the Strategic Agenda Saxion 2012-2016 is ‘vision, focus, action’ (Saxion, 2012a). The strategic agenda of Saxion is established by processing the strategic agenda of the Ministry of Education, Culture and Science and the report of the HBO Council.

Saxion suggests that the quality of education is their highest priority (Saxion, 2012a). They want to invest in the quality of the content of their education. This involves the strengthening the knowledge components of the programmes (Saxion, 2012a).

Furthermore, Saxion thinks stakeholders should be involved in the improvement. “Employees are vital for increasing quality, as we ask them to translate the professional space into professional responsibility” (Saxion, 2012a, p. 17).

Saxion wants their students to become active learners (Saxion, 2012b). Assessment should focus on judgement on theoretical knowledge, skills and attitude (Saxion, 2012b). They also find it important that students can give and receive feedback (Saxion, 2012b). The feedback should focus on the result as well as on the learning process (Saxion, 2012b).

3.2.2 Physiotherapeutic organizations

Physiotherapy is a paramedic health care profession. Physiotherapist are paramedic professionals who are experts in movement and function of the body. A physiotherapist is able to help patients with problems in the mobility, functional ability or with injuries. They can screen, diagnose and make therapeutic treatment plans for the patient. As a physiotherapist you help people of all ages and you treat acute injuries as well as chronic conditions.

Two physiotherapy organizations to monitor the quality of physiotherapy in the Netherlands are KNGF and SROF. KNGF (‘Koninklijk Nederlands Genootschap voor Fysiotherapie’) is an organization for physiotherapy in the Netherlands. KNGF represents the interests of about 22,000 associated physiotherapists at professional, social and economic level. The organization exists since

1889. SROF ('StudieRichtingsOverleg Fysiotherapie') is a consultative body of the 11 physiotherapy studies in the Netherlands. SROF accomplish a facilitating and coordinating role in the collaboration. SROF also collaborates with KNGF to monitor the quality of physiotherapists and physiotherapy studies. Every physiotherapy study sends a delegate to consulate in SROF.

KNGF and SROF collaborate in setting up inter alia the professional profile and national transcript. The professional profile serves the purpose of creating consensus about the profession and to compare the situation in the Netherlands with international developments (KNGF, 2006). It describes the competences students and physiotherapist need to work in the field of physiotherapy (KNGF, 2006). The purpose of the national transcript is to provide information to improve international transparency and the fair academic and professional recognition of qualifications (SROF & KNGF, 2008). The document will be revised for the study year 2013-2014. KNGF is responsible for the substantive vision on the profession. SROF is responsible for putting this vision into education.

The strategic agenda of SROF 2010-2013 only mention that there are initiatives for developing national progress test to monitor the theoretical knowledge level of students (SROF, 2010). The report is written in 2009-2010 so back then it was a relative new initiative. The new strategic agenda would probably focus more on this new development.

3.2.3 Physiotherapy studies

To become a physiotherapist in the Netherlands you have to study for four years at an university of applied sciences. To enter this study you have to passed 'HAVO' or' VWO' at secondary education. The study has a numerus fixus, so there is an enrollment quota for this study. You can study physiotherapy at 11 schools in the Netherlands: Saxion Enschede, HAN Nijmegen, Hogeschool van Amsterdam, Fontys Eindhoven, Hogeschool Rotterdam, Hogeschool Leiden, Hogeschool Utrecht, Heerlen Zuyd, Avans Breda, Hanze Groningen and Nieuwegein.

After graduating you can work as a physiotherapist at lots of different places; private practice, hospital, retirement home, etc. Many physiotherapist are doing a master study to specialize. There are several specializations, for example: sports physiotherapist, child physiotherapist, etc.

3.3 Context at micro level

In this paragraph the context at micro level will be described. This involves physiotherapy studies at Saxion and the assessment of this study.

3.3.1 Physiotherapy studies at Saxion Enschede

At Saxion Enschede you can study to become a physiotherapist. The study lasts for four years. The major programme exists of 210 credits. Each study year covers 60 credits. 30 credits can be earned by a minor of the students own choice.

The first year is the propaedeutic phase. In this year students learn the basic theory and skills. Students will have courses to learn a lot about anatomy and physiology. In this year the students also having their first internship. The goal of this internship is to introduce the students to the practice. The second study year exists of two phases which each will last a half year. The two phases are neurology/internal and orthopedics. In this year students will also do internship which is one day a week. In the last two years of the study students will do two times half a year internship. Furthermore they spend half a year to a minor of their own choice. Also a half year is spent on a project in which the student will focus on the roles manager and developer of the profession. Students can choose their own sequence in the last two years. The international programme differs from the regular programme. Saxion Enschede has inter alia German and Norwegian students. (Saxion, 2011)

The curriculum of physiotherapy studies from Saxion Enschede (2011) explains that the study programme is concentric. That means that knowledge will be repeated and deepened as the study continues.

3.3.2 Assessment at physiotherapy studies at Saxion Enschede

Assessment at physiotherapy studies are aimed at competences (Saxion, 2012c). These competences are divided into three professional roles: care provider, manager and developer of the profession (Saxion, 2012c). Physiotherapy studies at Saxion conducts four types of assessment: knowledge assessment, product assessment, develop assessment and performance assessment (Saxion, 2012b; Saxion, 2012c). In this paragraph only knowledge assessment will be described because this is relevant for the research.

Assessment serves several purposes at physiotherapies studies at Saxion. In the propaedeutic phase assessment is focused on selection and therefore the tests are mostly summative (Saxion, 2012b). Formative tests are emphasized in the post propaedeutic phase (Saxion, 2012b). It is thought that in this phase the students can direct their own learning process (Saxion, 2012b).

The main goal of knowledge assessment is to test whether students possess sufficient knowledge (Saxion, 2012b). Assessment takes place at the end of an educational period. The test consists of three subtests. Each subtest consists of 40 items. The assessment plan describes that there are multiple choice items in various forms (Saxion, 2012c). However, in practice currently the items have only two answer options: true/false. The items are developed by teachers and are based on the learning targets (Saxion, 2012c). The student has to answer 70% of the items correct to pass the test (Saxion, 2012b). The students get a grade on a scale from 1 to 10 on the test.

The theoretical knowledge tests is aimed initially at the first two levels according to Bloom's taxonomy (Saxion, 2011; Saxion, 2012c). These levels are knowledge and comprehension. This means, that students should reproduce and explain knowledge in terms of description, explanation, selection, summarize, give an example etc. Other levels of Bloom are: application, analysis, synthesis and evaluation. These levels are tested using other testing forms at Saxion.

After test administration, students can see the test accompanied by teachers. Teachers can provide for information on the results and decisions. Furthermore, they can help students in their learning process. Next to the summative goal of this assessment, it can also be used formative. The test policy of Saxion claims that an e-learning environment provides example tests so students can practice. Feedback on example tests will provide information on topics which the student does not yet master.

To monitor the quality of testing, Saxion control it at several aspects. The validity is controlled by making each study year a study plan. Also, in every test is stated which competences are tested. The reliability of the theoretical knowledge test is controlled by analysis the test items. To make sure students can prepare well for the test and so monitor transparency, information is given by the e-learning program Blackboard. (Saxion, 2011; Saxion, 2012b; Saxion, 2012c)

4. Method

This research is the analysis phase in educational design research. It consists of two parts. In the first part the qualitative method of conducting interviews with the stakeholders will be described. The second part consists of describing the secondary data analysis method with IRT.

4.1 Method part 1

Educational design research starts with an analysis to explore the gap between the current and desired situation (McKenney & Reeves, 2012). In the first part of this study we explore the current and desired views of stakeholders on different levels: macro, meso and micro level. On macro and meso level the views were described by document analysis which can be found in chapter 3. On micro level the views of mentors, teachers and students from physiotherapy at Saxion Enschede were analyzed by conducting interviews.

4.1.1 Respondents

In this research convenience sampling is used to select participants for the interviews. Convenience sampling is a technique in which individuals or groups are selected on availability and willingness to participate (Dooley, 2001; Onwuegbuzie & Leech, 2007). Different types of stakeholders participated in the interviews. Seven students participated in the interview. Five first year students of which three women and two men. Also two female students from the second and third year, were willing to be interviewed. Three teachers of which two men and one woman, participated in a group interview. Furthermore, four mentors of which three women and one man also participated in a group interview.

4.1.2 Instrumentation

Stakeholders were interviewed to understand the current and desired situation of theoretical knowledge assessment and feedback at physiotherapy studies at Saxion Enschede and to understand their opinion regarding progress testing and the feedback possibilities from an SMS. The interviews were, if possible, group interviews. The interviews lasted approximately 45 minutes.

In a group interview the researcher can obtain information from a certain group respondents at different topics. There is a small difference between group interview and focus group. Focus group aims to obtain enriching data on one topic. The advantage of a group interview is the opportunity to stimulate each other to think about the topic (Emans, 2002). It can be motivating to participate in a group interview. The disadvantage of a group interview is that it requires a good moderator (Baarda, de Goede & Teunissen, 2009; Emans, 2002). Also it is more difficult to transcript the interview (Baarda et al., 2009; Emans, 2002). The interviews were recorded so that transcribing was possible. Another disadvantage of group interviews is group pressure, in which participants may not feel safe to give their opinion or are influenced by the group members (Baarda et al., 2009).

A semi-structured interview offers flexibility (Baarda et al., 2009). The topics and some questions are fixed but the interviewer can change the sequence and has the opportunity to ask further questions (Baarda et al., 2009). An interview schedule was constructed to guide the interviewer. After a pilot and some minor adjustments in the schedule, it was used for the interviews.

4.1.3 Procedure

In qualitative research it is important to use a combination of different methods or so called triangulation to monitor the reliability and validity (Baarda et al., 2009). Triangulation is important to the internal validity and reliability in qualitative research (Baarda et al., 2009; Maso & Smaling, 1998; Miles & Huberman, 1994). Triangulation can take place at different levels. In this research triangulation took place at respondent level (Miles & Huberman, 1994). Different stakeholders participated in the interviews: students, teachers and mentors. Furthermore a pilot interview was held to construct a decent and relevant interview schedule.

4.1.4 Data analysis

The data retrieved from the interviews was analyzed according to the theory of Boeije (2005) and Miles & Huberman (1994). The interviews were recorded (with permission of the interviewees) and transcribed. After that the interviews were encoded. Encoding offers the opportunity to decrease and organize an amount of data (Boeije, 2005). A code in the interview is a part of the text in which the meaning of the fragment is expressed (Boeije, 2005). These codes can be assigned to parts of text of different size: words, sentences, paragraphs, etc. (Miles & Huberman, 1994). In encoding the interviews, three phases were used: open encoding, axial encoding and selective encoding. In this process of encoding, an encoding schedule was processed to offer structure and to analyze the data in a structured manner (Boeije, 2005; Miles & Huberman, 1994). To ease the process of encoding, a computer program was used. In this research the software 'NVivo.10' was used.

4.2 Method part 2

In the second part of this study a secondary data analysis was conducted. Existing test data was analyzed by IRT to look for test items for future item banking.

4.2.1 Data analysis

The test data which was used for secondary data analysis consisted of four theoretical knowledge assessments. The first test was a subtest anatomy of 40 items taken by 196 first year students in the school year 2011-2012. The second test was also (but another) subtest anatomy of 40 items taken by 155 first year students in the school year 2012-2013. The third test was a subtest neurology of 40 items taken by 135 second year students in the school year 2012-2013. The last data consisted of a subtest orthopedics of 40 items taken by 122 second year students in the school year 2012-2013.

A preselection of items was done by analyzing the test data with TestVision. TestVision is the testing service system used at Saxion Enschede. This system used by Saxion is based on CTT. Items not meeting the quality criteria were removed to make it possible to analyze the data with IRT. Items were removed based on the Rit value. The Rit value shows the distinctiveness of a question; the score of the test of the participant is compared to the score on a particular item (van Berkel & Bax, 2006). The Rit value shows the correlation between the item score and the test score: when a specific item is answered correctly, this participant is likely to score well on the whole test as well (van Berkel & Bax, 2006). When the Rit value is ≥ 0.35, it is considered very good (van Berkel & Bax, 2006). Items with Rit value ≤ 0.20 were removed, because these are considered insufficient (van Berkel & Bax, 2006). The removal of items continued until the alpha of the test was of sufficient value with a maximum of a removal of 20 items. Alpha estimates the reliability of the test (van Berkel & Bax, 2006). This value should be at least between 0.7 and 0.8 to be sufficient and to be useful for formative tests next to other assessments (van Berkel & Bax, 2006).

After the pre-selection of the items, the data was analyzed with IRT. The data was analyzed by the software 'R'. Also in this analysis, some items had to be removed to get the best possible model fit. The 2PLM was applied because this fitted better than 3PLM.

5. Results

In this chapter the results of part one and part two will be described. The first part describes the results of the interviews. The second part consists of the IRT analysis.

5.1 Results part 1

This paragraph describes the results of the interviews. After analysis the results are divided into a few subtopics: theoretical knowledge test, feedback, test development and progress testing.

5.1.1 Theoretical knowledge test

Since a few years the test is divided into three subtests (e.g. anatomy, physiology, etc.). The interviewed students mention to like the subtests, because they think the chance to pass is larger because they do not have to pass everything at once. Furthermore, the subtests are a kind of feedback for them because they can see which part is good or insufficient.

There is a discussion going on among the stakeholders about the question type. A subtest consists of 40 true/false items. Most students, teachers and mentors criticize these items. They think the high chance of guessing is unfair. Two students know that other students are flipping a coin at the test to answer the question. They get demotivated seeing this:

"I think the theoretical knowledge test has no added value. There are students who study a lot and some student who do not study at all, because they think the changes to pass are 50%. I have seen some students flipping a coin at the test. And also there are students, like me, who have studied very hard and they do not pass the test while they are sure that they have the ability."

"Some students take a coin to the test. Yes, seriously. I do not know those students, but I have heard it several times that they threw a coin at the test and that they have passed. Well, then I do not want to study at this school. I think the quality has to be improved."

The true/false item also often causes big confusion, because students think an item can be answered both ways depending on the reasoning. Words like 'often' or 'sometimes' in an item causes a lot of confusion. It also causes confusion by teachers who have to develop the items. Teachers mention it is difficult to formulate a correct sentence in developing the items. Also, teachers wonder whether the test tests knowledge or the verbal skills of students. So, they hesitate the validity of the test.

At last, teachers mention that the question type is too restricted. It only tests knowledge at recognition level:

"The true/false item is a too restricted item form to test knowledge. You can only test a part of the knowledge then. Another part of the knowledge would be better tested with other items, like multiple choice or match items."

The second and third year students agree with this. They would like to see more questions on application level.

According to a teacher, the test is also a communication tool towards students. He thinks that the school is communicating to students that the recognition of knowledge is enough.

The first year students are in general more positive experienced with the true/false items. They think it is easier to answer such questions.

5.1.2 Feedback at the theoretical knowledge test

Students differ in their opinion about the feedback after the theoretical knowledge test. Some think it is enough to get a grade. Some other students would like to receive more feedback. One student says:

"I think feedback is really important. For my own learning process in becoming a good physiotherapist. I would like to receive feedback from professionals."

All stakeholders agree, that the exam inspection is not well organized. Students are provided the opportunity to gain feedback from a teacher after the test. There is only one teacher available for feedback for a lot of students. This teacher does not know all questions and their answers. The teachers confirmed this problem. Two mentors think that students fear to ask feedback from the teachers. A student disagrees:

"The relation between the teachers and students is good. There is no real distance so you can ask them questions. There is time in or after the lesson to ask some questions. For example about the test. I think there are opportunities as well as time for that."

Physiotherapy studies at Saxion spends little time and attention to the feedback. In this way, the students in general, cannot use the feedback for their learning process. The feedback is insufficient to do so.

"Little. Because I think we receive too little feedback to use it. Sometimes I study harder for a retake or I just study the same materials as the first time."

A few students mention that they write down some information in the exam inspection:

"I write down some things I did wrong and I can repeat that to pass the retake."

With this information they can go to the teachers another time and ask for feedback. However, four students and one mentor mention that there is too little time during the exam inspection.

Students, teachers and mentors would like to improve the feedback after the theoretical knowledge tests. Most students would like to receive more feedback on good and bad aspects and why some items are good or wrong. The interviewees have some options to improve the exam inspection. First, one student suggest to provide a short explanation with each item. Second, a mentor suggests an option to organize a meeting after the test with all teachers and students. Students can ask feedback from the teacher who wrote the test item. However, one teacher has bad experience with this idea. He says that students will not come to a certain meeting, because they cannot get credits for it. The third suggestion is to evaluate the test in the lesson. At last, one teacher suggests to get more feedback out of the test self. He wants to get a good look at the development of the student by longitudinal testing:

"I think it will be good to test a kind of development. For example: to test the second year students with more difficult items than the first year students, but also repeat some information from the first year. So, like a progress test."

At last, some students, teachers and mentors suggest to implement example tests to get feedback before the final test. Students would like this to study, practice and learn from it.

The teachers do not claim to use the test results for their own evaluation. They do have other methods for this: ask feedback in the lesson and they get evaluated by the student through the school. One teacher is afraid that with the implementation of an SMS, Saxion will be held accountable for the results:

"I have objections to hold schools accountable. Everyone knows that the quality of primary schools cannot be compared just from CITO scores. It depends also on the population." ……. "It was never a purpose of CITO to make it a test to measure quality." …… "Whatever the purpose is, I doubt whether it stays that way."

5.1.3 Test development

As mentioned before, teachers find it hard to develop test items. The wording in the item has to be clear and unambiguous because of the true/false items. The items often causes a lack of clarity by students and therefore some items have to be removed after the test administration.

Teachers and mentors hesitate regarding the validity of the test. The true/false item causes this doubt. They wonder, whether at that point, it is testing knowledge or verbal skills of students. One teacher and mentor also mentions, that there are a lot of items that go in detail, so she doubts whether the test is a good reflection of the theoretical knowledge.

Teachers and a few students think the test has a good reliability. After the test, the teachers look at the test service system from Saxion (TestVision) to check the quality of the items. They look at the Alpha and Rit value. Students seem divided about the reliability. A first year male student thinks the reliability is sufficient:

"When I have studied enough, the result of the test confirmed it."

A second year female student does not agree. She had some good results on subtests she did not study for and vice versa. Also there is a doubt in reliability in the first year because some students already know the items before the start of the test. They were passed through by students who made the test earlier.

Instead of the true/false item, some interviewees suggest to offer the possibility to add some explanation to the item or use multiple choice questions, open questions or a mix of both. However, one teacher has logistic concerns with this because of the big amount of students. He also mention that the question type depends on the level of the students. He would not like to use open questions in the first and maybe second year. One student and one mentor suggest to use pictures in items to clarify the item. In this way the item can be more on application level. Also, adding pictures to items can be fair for more visual rated students. At last, the teachers would also like to see more items in a test.

5.1.4 Progress testing

Most interviewees are enthusiastic about the idea of implementing an SMS. Students think it is nice to compare their own results with the national average.

"Yes. I think national tests are good. Then you can compare it with other physiotherapy studies and you can conclude whether the programme in Enschede is good."

Two students are critical about the study load and think the implementation of an SMS would cause more stress.

One teachers knows that a precondition for the implementation of an SMS are longitudinal learning progression with learning targets. He knows that these are available because SROF developed it. However, the teachers claim they do not get a good overview of the learning progression. Furthermore, one teacher thinks you also need the same literature in all physiotherapy studies.

The stakeholders have different opinions about the implementation of the SMS regarding to use the test as an extra formative test or that it should replace the current knowledge assessment tests. Some stakeholders think the study load or work load can become too large when there is an extra test. Two students seem frightened about the idea of implementing an extra test. A reason to use it as an extra formative test, according to a teacher and a student, can be to test the own interpretation of the curriculum with the theoretical knowledge assessment. In these assessment the school can choose its own accents. Furthermore, stakeholders also think it depends on the quality of the progress test if it can replace the current theoretical knowledge test.

Students do not think their learning approach would differ if an SMS would be implemented. They say to already repeat learning materials. A few students think it is good to get the pressure of learning on an on-going basis. A third year student thinks the opposite. She thinks it is good to learn more targeted so she can focus on the learning goals of a certain educational period. One teacher, who has experienced the IPTM himself, mentioned that when the progress test is combined with summative tests, the learning process will still be aimed at summative tests.

The stakeholders would not like the idea of testing the whole curriculum in one test. They think it would scare the freshmen and that it is too difficult. Also most students do not like the idea that they cannot learn targeted for the test. The stakeholders were more positive about adaptive tests. Student think they would be motivated to perform as good as they can in adaptive tests.

Students would like to use the results of the SMS to compare themselves to the national average. Mentors think they would use the results of the SMS to confront students with their results and to use the results for conversation about the study progress.

The opinion on the consequences of the results of the progress test are clear. Most stakeholders think there should be a consequence to motivate students for the test. However, it is not clear what this consequence should be.

During the interviews also some examples of graphs of the IPTM were showed to the stakeholders. All stakeholders would really appreciate it if something alike would be implemented in their education. They think the graphs are clear and simple to understand. Adjustments in the graphs could be made by adding a legend and some information (e.g. how many mistakes, percentages, etc.). Especially the graph where the results are divided into subtopics was favorable received. Students claim to use it for their learning process. They would like to receive the graph digital with an option to print it. One student says the following about an example of a graph with scores on the subtopics deposited the national average:

"I think it is really important to receive this kind of feedback. At some subtopics you really have to improve, other subtopics you do not have to study so hard anymore because you already manage them. I think that would be real good and interesting."

5.2 Results part 2

In this paragraph the results of the IRT analysis will be given. There will be a short explanation of the test statistics followed by interpretation of these test statistics. For each test the following test statistics will be given per item: degrees of freedom (df), chi-square distribution (X^2), p-value (p), discrimination parameter (αi) and difficulty parameter (βi).

The statistic chi-square (X^2) can be used to check whether the model fits for the item. If this statistic is much larger than the degrees of freedom, a significant result can be reported (p-value) and there can be concluded that the model does not fit. Items that do not fit the model can be removed or rewritten. The misfit can be due to poor item quality, for example an unambiguous item. When a lot of items do not fit the model, the construct validity of the test needs to be reconsidered.

The discrimination parameter (αi) refers to the extent to which an item differentiates among participants on basis of their ability. It is also referred to as the slope. A high value for αi indicates a steep slope in the middle of the ICC. A high value for αi means that the item discriminates well between high and low ability students: persons with high ability have a much greater chance of correctly responding than persons of lower ability. Items with high discrimination power contribute in information greatly but over a narrow range. An ICC with a small slope has a low discrimination parameter and the probability of a correct answer changes slowly over the whole ability range. A low value for αi means that the item does not discriminate well between high and low ability students: a student with low ability has almost the same chance of correctly responding than a student with high ability. Items with low discrimination power provide less information but over a wide range. It is preferred to have items with a high discrimination power, so a high value of αi. For this scale, items with $\alpha i < 0.3$ will not contribute much to the test regarding discrimination (B. Veldkamp, personal communication, June 19, 2013). These items can be included in the item bank, but there has to be an awareness that items with $\alpha i < 0.3$ might not discriminate well enough between high and low ability students.

The difficulty parameter (βi) refers to the difficulty level or location of an item which defines the amount of the latent trait needed to have a probability of 0.5 of endorsing the item. βi ranges from -3 till +3. The higher the difficulty, the higher on the trait level a participant needs to be in order to endorse the item. In other words: if βi increases, the ability of a participant should be higher to answer the item right. In the test, there should be a variation in the difficulty of the items.

For the protection of the items in the current testing bank, the content of the items are not stated here, only the item numbers are.

5.2.1 Anatomy 2011-2012

Table 1 provides the test statistics for the remaining 19 items from the anatomy test school year 2011-2012. Figure 15 shows the ICCs of these items. In the table can be found that item 12 does not fit the model well (X^2=9.825, df=3, p=0.020). However, when item 12 was removed, the model fit changed in a negative way. So this item was included in the test data to get the best model fit for the whole data set. The rest of the items fit this 2PLM with a significance level of 0.05. Items 9, 10, 13, 17, 25, 26, 30, 32, 33, 34, 35 and 40 have $\alpha i < 0.3$. So, these items do not discriminate well between high and low ability students. Item 12 did not fit the model, but has the highest discrimination power in this data set ($\alpha i = 0.687$). There is little variation for the difficulty parameter. The difficulty is centered around the average level. There could be more easy and more difficult items in this data set. Items with difficulty > 0 could also function as an relatively easy question for the second year students (all items excluding items 8, 34 and 40).

Table 1

Test statistics. Degrees of freedom (df), chi-square distribution (X^2), p-value (p), discrimination parameter (αi) and difficulty parameter (βi) of 19 items from the sub test anatomy 2011-2012 with *n*=196.

Item	df	X^2	p	αi	βi
7	9	6.609	0.678	0.419	0.657
8	7	1.116	0.993	0.353	-0.224
9	8	7.343	0.500	0.229	0.192
10	8	5.848	0.664	0.127	0.658
12	3	9.825	0.020	0.687	0.427
13	5	6.772	0.238	0.215	0.249
15	5	5.934	0.313	0.337	0.405
17	5	5.968	0.309	0.285	0.466
25	5	4.893	0.429	0.203	0.220
26	6	4.171	0.654	0.220	0.302
28	4	3.443	0.487	0.572	0.385
30	7	10.436	0.165	0.212	0.456
32	8	13.570	0.094	0.082	0.340
33	8	9.703	0.287	0.231	0.729
34	8	8.549	0.382	0.072	-0.268
35	8	9.110	0.333	0.140	0.133
36	8	6.118	0.634	0.537	0.322
37	9	9.095	0.429	0.376	0.020
40	8	10.043	0.262	0.229	-0.486

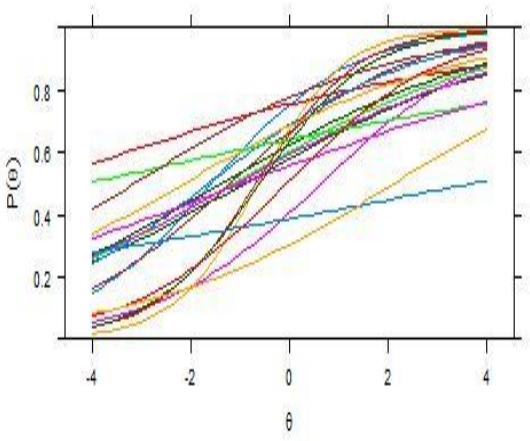


Figure 15. ICCs of 19 items from the anatomy test 2011-2012.

5.2.2 Anatomy 2012-2013

Table 2 provides the test statistics for the remaining 21 items from the anatomy test school year 2011-2012. Figure 16 shows the ICCs of these items. All items fit the 2PLM that was used with a significance level of 0.05. The discrimination parameter of items 1, 19, 24 and 25 is < 0.3, so these items discriminate not well. There is a tolerable variation in the difficulty level of the test. Difficult items ($\beta i > 0$) could also be used for the second year students.

The two subtests anatomy have one overlapping item. Items 12 in the anatomy test 2011-2012 corresponds with item 21 in the anatomy test 2012-2013. These items have exactly the same formulation in the test, but the test statistics for these items differ a lot. The explanation for this can be, that the two tests have different populations. Furthermore, the sample is small. For the anatomy test 2011/2012 n = 196 and n = 155 for the subtest anatomy 2012-2013. For IRT analysis with the 2PLM it is suggested that n is around 500 to estimate the parameters.

Table 2

Test statistics. Degrees of freedom (df), chi-square distribution (X^2), p-value (p), discrimination parameter (αi) and difficulty parameter (βi) of 21 items from the sub test anatomy 2012-2013 with n=155.

Item	df	X^2	p	αi	βi
1	10	9.199	0.513	0.212	0.378
5	9	8.224	0.512	0.316	0.436
6	7	3.293	0.857	0.498	0.091
10	7	10.563	0.159	0.451	0.546
11	1	1.342	0.247	0.510	1.662
13	6	4.992	0.545	0.349	-0.076
16	5	7.643	0.177	0.420	1.074
17	4	2.158	0.707	0.497	0.832
18	5	6.920	0.227	0.637	0.274
19	6	8.324	0.215	0.255	-0.403
21	4	3.860	0.425	0.416	0.086
23	5	8.634	0.125	0.521	-0.349
24	8	5.286	0.727	0.112	-0.157
25	7	8.924	0.258	0.186	0.329
29	5	5.447	0.364	0.872	0.675
33	6	4.816	0.568	0.337	1.410
34	8	5.339	0.721	0.528	1.165
35	8	14.086	0.080	0.687	1.057
36	5	5.355	0.374	1.081	2.110
39	7	3.954	0.785	0.740	1.045
40	8	4.416	0.818	0.606	0.021

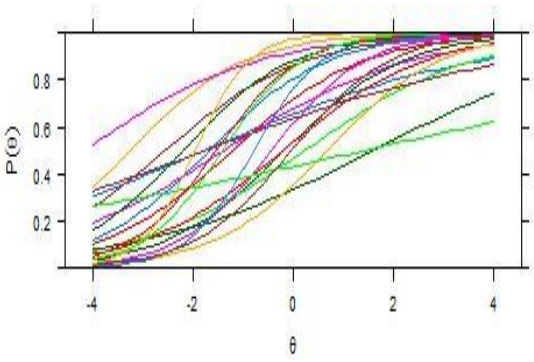


Figure 16. ICCs of 21 items from the anatomy test 2012-2013.

5.2.3 Neurology 2012-2013

Table 3 provides the test statistics for the remaining 17 items from the subtest neurology. Figure 17 shows the ICCs of these items. All items fit the 2PLM because there is no significant value for p (< 0.05). Assuming that items with $\alpha i < 0.3$ do not contribute to the test, items 2, 8, 9, 15, 19, 25, 30, 36 and 37 are not discriminating enough and using these items in future progress test should be done with care. There is some variation in difficulty in this test. Items that have $\beta i < 0$ can also be used as items for the first year test; items: 8, 37 and 38. More difficult items, for example item 22 with $\beta i = 1.579$ could be used for a third year progress test.

Table 3

Test statistics. Degrees of freedom (df), chi-square distribution (X^2), p-value (p), discrimination parameter (αi) and difficulty parameter (βi) of 17 items from the sub test neurology 2012-2013 with n=135.

Item	df	X^2	p	αi	βi
2	8	11.759	0.162	0.130	0.225
3	5	5.907	0.315	0.536	0.209
4	3	2.952	0.399	0.627	0.811
8	4	5.645	0.227	0.229	-0.215
9	4	2.489	0.647	0.157	0.553
15	4	5.582	0.233	0.057	0.075
19	4	5.205	0.267	0.270	0.677
22	3	5.103	0.164	0.823	1.579
25	5	5.096	0.404	0.208	0.107
26	5	2.237	0.816	0.450	0.986
30	6	9.775	0.134	0.214	0.814
32	5	10.203	0.070	0.416	0.301
36	7	10.998	0.139	0.121	0.727
37	7	8.565	0.286	0.132	-0.133
38	6	6.801	0.340	0.439	-0.109
39	7	7.098	0.419	0.632	0.470
40	7	7.900	0.342	0.308	0.446

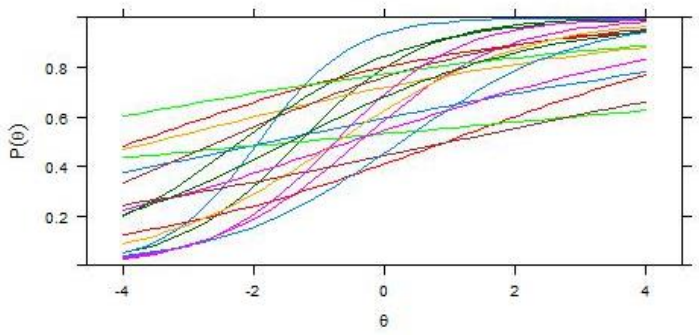


Figure 17. ICCs of 17 items from the neurology test 2012-2013.

5.2.4 Orthopedics 2012-2013

The last results are described for the orthopedics test. Results are shown in table 4: the test statistics for the 25 remaining items. Figure 18 shows the ICCs of these items. There is one item that does not fit the model: item 8 (X^2=16.752, df=8, p=0.033). The rest of the items fit this 2PLM with a significance level of 0.05. Items 8, 11, 13, 19, 23, 26, 32 and 38 have $\alpha i < 0.3$ so these items do not discriminate well. There is little variation in the difficulty level of the test, but most items are centered around the average. Again, difficult items could be used for a third year test. Easy items could be used as a difficult item in a first year test.

Table 4

Test statistics. Degrees of freedom (df), chi-square distribution (X^2), p-value (p), discrimination parameter (αi) and difficulty parameter (βi) of 25 items from the sub test orthopedics 2012-2013 with *n*=122.

Item	df	X^2	p	αi	βi
3	7	3.228	0.863	1.011	0.395
4	10	7.421	0.685	0.351	-0.070
5	9	8.324	0.502	0.539	0.190
6	6	6.088	0.413	0.438	0.939
8	8	16.752	0.033	0.272	0.125
9	2	2.605	0.272	0.948	1.424
11	8	5.999	0.647	0.219	0.306
12	4	6.621	0.157	0.439	1.080
13	8	5.721	0.678	0.052	-0.050
16	6	4.298	0.637	0.571	0.260
18	2	2.698	0.260	0.654	1.096
19	6	7.154	0.307	0.232	0.554
20	3	3.542	0.315	1.031	0.544
21	6	4.641	0.591	0.641	0.421
23	7	5.367	0.615	0.281	0.513
25	5	4.850	0.435	0.817	0.249
26	9	13.666	0.135	0.145	-0.005
27	8	6.374	0.605	0.450	0.726
28	8	8.150	0.419	0.581	0.295
32	9	9.257	0.414	0.159	0.390
34	10	13.730	0.186	0.319	0.464
36	6	5.502	0.481	0.905	1.543
38	11	11.101	0.435	0.173	0.406
39	9	6.725	0.666	0.451	0.725
40	9	9.173	0.421	0.742	-0.423

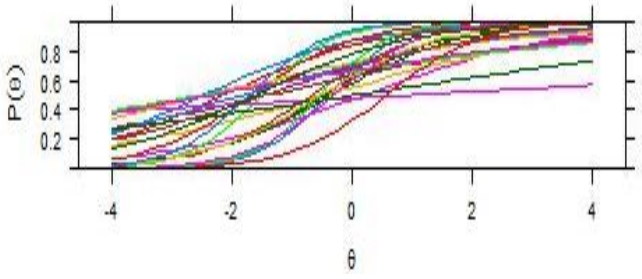


Figure 18. ICCs of 25 items from the orthopedics test 2012-2013.

6. Conclusion & Discussion

In this chapter, the sub questions and research question will be answered. Recommendations for further research will be given and some limitations of this study are described.

Through description of the context and interviews with stakeholders the first sub question will be answered: *What is the current situation in assessing theoretical knowledge and the feedback at physiotherapy studies at Saxion Enschede?*

Physiotherapy studies at Saxion Enschede utilizes different types of assessments of which theoretical knowledge assessment is one. The aim of this summative test is to test whether students possess sufficient theoretical knowledge at the end of a certain educational period. The test is aimed at the knowledge and comprehension level from Bloom's taxonomy. The three subtests consists of 40 true/false items. Items are developed by teachers and are based on learning targets.

Students, teachers and mentors of physiotherapy studies at Saxion discuss the question type in theoretical knowledge assessment. First, some students get demotivated by the true/false items because there is a high probability of guessing. However, first year students like this question type because they think it is easy to answer such items. Second, this question type often causes big confusion in interpreting the item. For teachers it is difficult to develop the items, because it is hard to formulate the true/false item. Van Berkel and Bax (2006) confirm these statements. They also mention the guess probability in true/false items of 50% as a disadvantage of true/false items. Also, the item has to be 100% true or false to prevent wrong interpretation (van Berkel & Bax, 2006).

Furthermore, teachers mentioned that the true/false items are too restricted because it only tests knowledge at recognition level. However, the policy of Saxion state to test at Blooms knowledge and comprehension level in the theoretical knowledge assessment (Saxion, 2011; Saxion, 2012c). So, the items should focus on recognition level as teachers mention this as disadvantage. Other levels of Bloom are tested using other assessment forms. The true/false items are testing at knowledge and comprehension level and this could also be a reason why first year students like the true/false items and state that it is 'easy' for them to answer such questions.

In contradiction of Saxion's policy, the school does not pay enough time and attention to feedback. The policy state that feedback should focus on result as well as on the learning process (Saxion, 2012c). However, through interviews, the stakeholders stated that there is too little time and attention for feedback. As feedback students get a grade for the test and they can see the test afterwards accompanied by a teacher. All stakeholders agree, that the exam inspection is not an optimal form to get feedback after the test. Only one teacher, who does not know all questions and the answers, is available for feedback for a lot of students. According to most interviewed students, the feedback is insufficient to use for their learning process. Some students claim to ask extra information from teachers but this requires a good self-regulation from students. In self-regulated learning complex metacognitive skills and strategies are aimed to improve the learning process and thereby the learning results (Boekaerts, 1999). Self-regulation requires a lot effort and complex skills of students (Boekaerts, 1999).

Saxion's policy also claims to provide example tests for students (Saxion, 2012c). However, in the interviews the stakeholders mention that these example tests are not available in practice. They would like example tests to be available so they can practice before the test.

At last, the stakeholders thought the reliability of the test was in general acceptable. However, analysis of TestVision shows alpha values of 0.39, 0.51, 0.41 and 0.57 for respectively the subtests: anatomy 2011-2012, anatomy 2012-2013, neurology 2012-2013 and orthopedics 2012-2013. These values for reliability are far below the accepted level of 0.70 (van Berkel & Bax, 2006). So, the theoretical knowledge tests of Saxion are not very reliable.

The second sub question will also be answered by results of the interview and description of the context: *What is the desired situation in assessing theoretical knowledge and the feedback for physiotherapy students at Saxion Enschede?*

Stakeholders on all levels agree that assessment should be improved and thereby improve education. According to the Dutch Ministry of Education, Culture and Science, the quality has to be improved by external validation of assessment (Ministerie OCW, 2011). The HBO Council advised in

their report to implement national progress test that are bottom-up developed (HBO Raad, 2012). This will be further described in answering the research question.

Almost all stakeholders from physiotherapy studies at Saxion agree that they would like to see other types of question in the theoretical knowledge test. Instead of true/false items the best option seems to be multiple choice items. This question type is suitable for a large amount of students (van Berkel & Bax, 2006). Through more answer options the item has a lower guess probability (van Berkel & Bax, 2006). Multiple choice questions test in general reproduction knowledge (van Berkel & Bax, 2006). Some teachers and some students would like to see more items on application level instead of knowledge and comprehension. However, and already mentioned before, this is not in accordance with Saxion's policy. Also open answer questions could be a possibility but this has logistic concerns because of the large amount of students. Therefore, multiple choice questions seems a plausible solution for adjusting the question type. There are a lot of different types of multiple choice questions. For example: the conventional multiple choice, match items or complex multiple choice items. For different forms of multiple choice items, their explanation and examples see for example van Berkel and Bax (2006) and Haladyna, Downing and Rodriguez (2002).

The interviewed stakeholders from physiotherapy studies at Saxion Enschede agree that the feedback from the theoretical knowledge assessment should be improved. They would like more information regarding the item, the correct answer and a short explanation with the item. Students would like to receive feedback on good and bad aspects of their test result. There were already some options mentioned for improving the feedback: (1) provide a short explanation with each item, (2) a meeting with all teachers and students after the test and (3) evaluate the test in the lesson. Saxion could consider to use these options for improvement of the feedback after the test.

Furthermore, one interviewed teacher suggested to get feedback out of the test itself. He would like to see longitudinal testing to monitor the study progress of students. This also offers more detailed feedback. It is also what students want: detailed feedback on good and bad aspects of their test.

At last, the feedback could be used for improvement of the education by teachers. The interviewed teachers said to not use the results of the test systematically to improve their own courses. Physiotherapy studies at Saxion could consider to make more use of the test results for internal evaluation to improve their education.

Also the last sub question will be answered: *How can IRT be used in developing an SMS for physiotherapy studies at Saxion Enschede?*

IRT provides a framework for constructing measurements, establishing validity measurements, estimating item and test characteristics, estimating abilities of individuals and spread of abilities in populations as well as providing a framework for interpreting test results (Kamphuis & Moelands, n.d.; Vlug, 1997). Separating the ability of a student and the item characteristic and put them on one scale provides the opportunity to compare results of students to previous measurements.

Because IRT is not population dependent, the item parameters can be estimated. In this study, the item parameters are estimated for existing test data. These item parameters can be stored in the item bank and be used for future progress tests. An item bank is "a pool of test items where the item parameters are known through pretesting" (Glas & Geerlings, 2009, p. 85).

In assessing students you want to discriminate between students in knowledge and ability. As described in the results, items with a low discrimination parameter ($\alpha i < 0.3$) are not discriminating well between low and high ability students. These items can be used, but there should be an awareness of the low discrimination power.

In a test, the difficulty level of the test has to be spread. There should be some difficult, average and easy items present in the test. When constructing a test, items of different difficulty levels should be represented. As described in the result, difficult items from the first year test could also be used as easy items in the second year test. Easy items from the second year test can be used as difficult items for the first year test. However, a remark should be made. Physiotherapy studies at Saxion Enschede has chosen to split the second year in educational periods of specialism neurology and orthopedics. So, even if items are very easy for second year students, there might be a chance that they are far too difficult for first year students because they have not studied these subjects yet.

Analysis of true/false items with the 3PLM requires big amount of data to get good estimation of the item parameters. The 2PLM can also be used but then some distortion takes place. The items are

estimated slightly easier than actual is the case. From a technical point of view it would be better to use items with more answer options to get good estimation of the parameter with limited amount of data.

So, IRT can be used as a measuring technique for developing an SMS. Data analysis in this study already provided items for the item bank. Furthermore, IRT could be used in developing CAT.

With answering the sub research questions, it is possible to give answer to the main research question: *How should an SMS for theoretical knowledge level for physiotherapy studies at Saxion Enschede be developed?*

The development and implementation of a national SMS for physiotherapy studies is inevitable. Currently there were doubts about the quality of the diplomas from higher education. There were some cases know in which students wrongly received their diploma. The Ministry of Education, Culture and Science thinks this is an alarming case and serious actions have to be taken. Therefore the Ministry and the HBO council suggest to develop national longitudinal progress testing in higher education to improve the quality of education by external validation. Universities of applied sciences have to collaborate to invest in developing such tests.

The first precondition of developing an SMS is, that the system should involve the entire education curriculum. The IPTM takes this literally and so the tests consists of items based on the end terms of the curriculum. The vision of the IPTM was to stimulate students to learn on an on-going basis. Therefore, the whole curriculum is included in the progress tests to not let students learn targeted for the test. However, testing the whole curriculum can cause demoralization for freshmen. The HBO Council also suggested to take a representative sample of the entire curriculum in the progress test from students of all levels. However, the results of the interview conclude the opposite. The interviewed stakeholders would not like this kind of testing. They think it would cause stress and demoralization. Also, the main goal of the SMS for higher education is external validation and monitoring progression. In this sense, it can be a good idea to not test the whole curriculum but implement adaptive tests. Stakeholders were also positive about adaptive testing. They think this would be more motivating and realistic. Adaptive testing seems a fair alternative. It can test the entire curriculum but the optimal level of items are selected for the individual student. This would involve digital testing. However, an SMS cannot depend on software, it is a useful and unavoidable tool in developing a good and adaptable SMS.

The second precondition of developing an SMS is to have learning progression with end- and sub-targets. Physiotherapy studies already did a good job in developing the curriculum. The national transcript is developed by SROF and explains the physiotherapy curriculum. The document has to be revised for next school year. The interviewed teachers know these learning progression exists but they claim to not have a good overview of the learning progression. So, SROF should not only revise the document but also raise awareness and attention among stakeholders. One of the teachers from Saxion mentioned that the literature among all physiotherapy studies should be the same. This is not necessary. If the learning progression with end- and sub-targets are known, each school have its own autonomy to create the courses. This is also the case in primary education in the Netherlands. The CITO tests are aimed at the national curriculum. Each school uses its own learning methods which are tuned at this curriculum. A remark has to be made: there has to be carefully dealt with specializations in the study programme. So, it is necessary to tune the physiotherapy curriculum but not the shape of educational methods to fulfill this curriculum.

The third precondition of an SMS is a technical aspect. The test instrument must make longitudinal monitoring possible. This can be done by the use of IRT. This study already estimated some item parameters for item banking. IRT is an advanced measuring technique which provides a lot of options in developing and analyzing tests. The analyzed items were true/false items. They can be used for future tests, but it should be noted that a lot of stakeholders do not like the true/false items. Rather multiple choice questions should be used. Multiple choice items can appear in lots of different forms of which true/false is one. Also, multiple choice items with more answer options give better estimation item parameters in the analysis with IRT. So, the selected items in this study could be used for future progress test but in developing the future test one should take in mind to combine the true/false items with other question types.

Saxion should collaborate with other universities of applied sciences in developing an SMS. Within physiotherapy studies a collaboration already exists. SROF is an organization in which all physiotherapy studies in the Netherlands are represented. So, SROF could cooperate in the consultation of developing an SMS. In total there are 11 physiotherapy studies in the Netherlands. The HBO Council advised to work together with three or more universities of applied sciences. The collaboration can help to improve the quality and to reduce the work load of the development.

Almost all interviewed stakeholders from physiotherapy studies at Saxion where positive about implementing an SMS. In particular, they were eager on the possible feedback an SMS could deliver. Feedback examples from the IPTM were used. By the development of feedback for an SMS the feedback from CITO and IPTM could be used as an example. All stakeholders were very positive about the comparison with the national average and the feedback divided into subtopics. Stakeholders found the feedback clear and easy to understand. It could be delivered digital with an option to print. The feedback from the SMS would mainly be used by students to compare their results with the national average. In the interview, mentors expressed to use the results to confront the students with it and to use it for conversation about the study progress. Teachers did not express how to use the feedback. Saxion could stimulate to let teachers check at which aspects students at their school falls behind and to improve this. This internal evaluation from test results is currently not much used at Saxion.

An advice regarding the implementation of the SMS would be to use it at first as an additional formative test. Some stakeholders would like to reduce the work and study load by replacing the SMS for the theoretical knowledge assessment. However, in the developing and implementation phase it can be good to implement the SMS in phases and not to substitute the test because of possible problems at the start of using an SMS. Later on in the process and when the quality is of a sufficient level, the SMS can perhaps substitute the theoretical knowledge assessment.

Also, it is not yet clear what consequences should count for the results of the progress test. The HBO Council describes that each school can choose its own consequence for the tests. They recommend to give 4 credits for progress testing a school year. Stakeholders from Saxion think there should be a consequence. Otherwise they would not be motivated. They do not know which consequence is appropriate. There should be carefully acted upon the consequences from the results of the progress test in the developing and implementation phase.

A few remarks have to be made regarding this study. First, the results should be carefully interpreted. For the qualitative part, in total 14 stakeholders were interviewed. This forms a small sample but is not enough to generalize it. However, an impression for this analysis phase could be made. Furthermore, respondents were not selected random. They were selected by convenience sampling which means, that they volunteered to participate in the interview. In general, the students were active and critical students. Two of them participate in the programme committee. Their critical attitude can be regarded as positive for this study.

The results of the quantitative data should also be carefully interpreted, because the used test data was not meant to be used for this research purpose. Originally, the idea was to perform a pilot progress test. This was, unfortunately, not possible due to time limitations and the current possibilities at Saxion. Therefore existing test data was used to estimate the item parameters. In estimating item parameters for 2PLM IRT n has to be around 500 to get a good estimation. In this study n was between 122 and 196 which is too little for good estimation. Therefore, these item parameters have to be dealt with carefully.

Further research can focus on many different aspects. This was only the analysis phase of educational design research and it is the start of a long process of developing. Therefore, this research should and will continue in order to develop, implement and evaluate an SMS for physiotherapy studies. It is important to develop an SMS with an evidence-based approach. Certain choices have to be well considered and research on different subtopics can contribute to the evidence-based approach of the development of an SMS.

The next step in the educational design based research is to start with developing the tests. Help can be found in the Handbook of Test Development from Downing and Haladyna (2006), the book from van Berkel and Bax (2006) or other test development plans or books. Wrigley et al. (2012) set up a systematic framework for progress test which could be helpful in developing. This study and the ideas about aspects like question type, consequences, feedback, etc. should be included in the

design. In designing the test, this research can contribute to the understanding of the context in which it has to be designed. Also the views of stakeholders should be used in developing tests. The designer should check several times what the views of the stakeholders are to create an external as well as an internal consistency (Fullan, 2007; van Berkel & Bax, 2006). Teachers should be involved, because they are the ones designing the items and using the results. Students should be involved because they have to make the tests and receive the feedback appropriate and use it. Mentors should be involved because they have to make use of the feedback as a bridge between school and student. Furthermore, the tests should be developed bottom-up so involvement of stakeholders is desired.

After test development, the test could be piloted and analyzed by IRT to create an item bank. A suitable system that analyzes the items and create feedback have to be developed or found. In this research the items were transferred from TestVision to 'R' via Excel. This is not desired and not user friendly. An appropriate system should be developed or sought.

Further research could also focus on CAT. Research on digital testing was not taken into account in this study, but it seems the best solution for testing the entire curriculum. In a way, students will not be demoralized and frustrated. CAT can be developed by creating developing an item bank and making use of IRT. Further research should focus on the possibilities and practical and technical aspects of CAT.

References

ARBO: Adviesraad voor het Basisonderwijs (1988). *Voorrang aan achterstand: advies over een integraal beleid ter voorkoming en bestrijding van onderwijsachterstand.* Zeist, Netherlands: ARBO.

Baarda, D.B., de Goede, M.P.M., & Teunissen, J. (2009). *Basisboek kwalitatief onderzoek* (2nd ed.). Groningen/Houten: Wolters-Noordhoff.

Berkel, H. (1990). Assessment in a problem-based medical curriculum. *Higher Education, 19* (2), 123-146.

Black, P., & William, D. (1998). Inside the Black Box: Raising Standards Through Classroom Assessment. *Phi Delta Kappan, 80* (2), 139-148.

Blok, H., Otter, M.E., & Roeleveld, J. (2002). Coping with conflicting demands: Student assessment in Dutch primary schools. *Studies in Educational Evaluation, 28* (2), 177-188.

Boeije, H. (2005). Analyseren in kwalitatief onderzoek, denken en doen. Amsterdam: Boom onderwijs

Boekaerts, M. (1999). Self-regulated learning: where we are today. *International Journal of Educational Research, 31*, 445-457.

Boshuizen, H., van der Vleuten, C., Schmidt, H., & Machiels-Bongaerts, M. (1997). Measuring knowledge and clinical reasoning skills in a problem-based curriculum. *Medical Education, 31*, 115-121.

Brown, S. (1999). Institutional Strategies for Assessment. In Brown, S., & Glasner, A. (Eds.), *Assessment matters in higher education: Choosing and Using Diverse Approaches* (pp. 3 -13). Buckingham: Open University Press.

CITO (n.d.). Retrieved from http://www.cito.nl/

Davenport, T. H., & Prusak, L. (1998). Working knowledge. How organizations manage what they know. Boston: Harvard Business School.

Dooley, D. (2001). *Social Research Methods.* New Jersey: Pearson.

Downing, S.M., & Haladyna, T.M. (2006). *Handbook of Test Development*. Mahwah, NJ: Lawrence Erlbauw Associates.

Emans, B. (2002). *Interviewen.* Groningen/Houten: Wolters-Noordhoff.

Evans, C. (2013). Making Sense of Assessment Feedback in Higher Education. *Review of Educational Research, 83* (1), 70-120.

Fullan, M. (2007). *The new meaning of educational change.* New York: Teachers College.

Gillijns, P. (1991). *Leerlingvolgsysteem*. Tilburg, Netherlands: Zwijssen.

Glas, C.A.W., & Geerlings, H. (2009). Psychometric aspects of pupil monitoring systems. *Studies in Educational Evaluation, 35*, 83-88.

Haladyna, T.M., Downing, S.M., & Rodriguez, M.C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309-334.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77* (1), 81-112.

Havnes, A., Smith, K., Dysthe, O., & Ludvigsen, K. (2012). Formative assessment and feedback: Making learning visible. *Studies in Educational Evaluation, 38*, 21-27.

HBO raad, vereniging van hogescholen (2012). *Vreemde ogen dwingen. Eindrapport Commissie externe validering examenkwaliteit hoger beroepsonderwijs.* Retrieved from http://www.hbo-raad.nl/hbo-raad/publicaties/doc_view/1637-vreemde-ogen-dwingen

Hérnandez, R. (2012). Does continuous assessment in higher education support student learning? *Higher Education, 64* (4), 489-502.

Huxham, M. (2007). Fast and effective feedback: are model answers the answer? Assessment and Evaluation in Higher Education, 32, 601–611.

IPTM (n.d.). Retrieved from http://www.ivtg.nl/en/node/69

Irons, A. (2008). *Enhancing learning through formative assessment and feedback.* London: Routledge.

Isaksson, S. (2007). Assess as you go: The effect of continuous assessment on student learning during a shortcourse in archaeology. *Assessment and Evaluation in Higher Education, 33*(1), 1–7.

Johnsson, A. (2012). Facilitating productive use of feedback in higher education. *Active Learning in Higher Education, 14*(1), 63-76.
Kamphuis, F., & Moelands, F. (n.d.). A Student Monitoring System. *Educational Measurement: Issues and Practice*.
KNGF (2006). *Het beroepsprofiel van de fysiotherapeut.* Retrieved from: http://www.fysionet.nl/centraal-kwaliteitsregister/beroepsprofiel.html
Koornneef, N.H. (1991). *Leerlingvolgsysteem voortgezet onderwijs: een haalbaarheidsstudie.* Amersfoort, Netherlands: Drukkerij Bouman b.v.
Maso, I., & Smaling, A. (1998). *Kwalitatief onderzoek: praktijk en theorie.* Amsterdam: Boom.
McHarg, J., Bradley, P., Chamberlain, S., Ricketts, C., Searle, J., & McLachlan, J. (2005). Assessment of progress tests. *Medical Education, 39* (2),221-227.
McKenney, S. & Reeves, T. (2012). *Conducting educational design research*. London: Routeldge.
Miles, M. B., Huberman, A.M. (1994). Qualitative Data Analysis: an expanded sourcebook. Thousand Oaks, CA: Sage.
Ministerie van Onderwijs, Cultuur en Wetenschap (OCW) (2011). *Kwaliteit in verscheidenheid: Strategische Agenda Hoger Onderwijs, Onderzoek en Wetenschap*. Retrieved from: http://www.rijksoverheid.nl/documenten-en-publicaties/rapporten/2011/07/01/kwaliteit-in-verscheidenheid.html
Muijtjens, A., Schuwirth, L., Cohen-Schotanus, J., Thoben, A., & van der Vleuten, C. (2008). Benchmarking by cross-institutional comparison of student achievement in a progress test. *Medical Education, 42* (1), 82-88.
Onwuegbuzie, A.J., & Leech, N.L. (2007). A Call for Qualitative Power Analyses. *Quality & Quantity, 41*, 105-121.
Saxion (2011). *Curriculum document: AGZ Fysiotherapie*. Enschede: Saxion.
Saxion (2012a). *Strategic agenda 2012-2016: Vision, focus, action.* Retrieved from: http://saxion.nl/over/organisatie/visie/
Saxion (2012b). *Toetsbeleidsplan AGZ.* Enschede: Saxion
Saxion (2012c). *Toetsplan studiejaar 2012-2013 fysiotherapie.* Enschede: Saxion.
Schaap, L., Schmidt, H., & Verkoeijen, P. (2012). Assessing knowledge growth in a psychology curriculum: which students improve most? *Assessment and Evaluation in Higher Education, 37* (7), 875-887.
Schuwirth, L.W.T., & van der Vleuten, C.P.M. (2012). The use of progress testing. *Perspectives on Medical Education, 1*, 24-30.
SROF & KNGF(2008). *National diploma supplement and National Transcript Fysiotherapie*. Retrieved from: http://www.srof.nl/rapporten/beleidsdocumenten.html
SROF (2010). *Strategische beleidsagenda SROF 2010-2013*. Retrieved from: http://www.srof.nl/rapporten/beleidsdocumenten.html
Van Berkel, H., & Bax, A. (2006). *Toetsen in het hoger onderwijs*. Houten: Bohn Stafleuk van Loghum.
Van der Drift, M. (1995). *Leerlingvolgsysteem in de praktijk: een onderzoek naar het gebruik van het Leerlingvolgsysteem van het Cito.* Thesis, Nijmegen.
Van der Vleuten, C.P.M., G.M. Verwijnen, and W.H.F.W. Wijnen. 1996. Fifteen years of experience with progress testing in a PBL-curriculum. *Medical Teacher, 18*, 103–109.
Vlug, K.F.M. (1997). Because every pupil counts: the success of the pupil monitoring system in The Netherlands. *Education and Information Technologies, 2*, 287-306.
Wade, L., Harrison, C., Hollands, J., Mattick, K., Ricketts, C., & Wass, V. (2012). Student perceptions of the progress test in two settings and the implications for test deployment. *Advances in Health Sciences Education, 17* (4), 573-583.
Wrigley, W., van der Vleuten, C., Freeman, A., & Muijtjens, A. (2012). A systematic framework for the progress test: strengths, constraints and issues. *Medical teacher, 34* (9), 683-697.