What do we learn from assessment?

*Developing an observation instrument to measure the use of Assessment for Learning in the classroom*

Jet Verbeeten BSc, s0193801
Educational Science & Technology,
University of Twente
Supervisors:
dr. K. Schildkamp & W.B. Kippers MSc
22-3-2017
Summary

The use of Assessment for Learning (AFL) in the classroom can lead to better education, but in order to improve instruction, it is important for teachers to know where they stand in terms of their use of AFL. By observing teachers, their use of AFL can be measured and suggestions for improvement can be made. Unfortunately, there is no instrument available that meets all the requirements to be a good observation instrument to measure AFL in the classroom, as we know of. Therefore, within this research such an instrument was developed and the following research question was answered: What are the characteristics of an observation instrument to measure Assessment for Learning in the classroom?

This thesis describes the development of the observation instrument in four phases. In the first phase, a literature study was conducted which led to information about AFL (five strategies), the development of observation instruments and a checklist of requirements for observation instruments that was used to screen instruments and questionnaires that were used as inspiration. This led to the first draft of the instrument. In the second phase, the first draft was shown to educational researchers and teachers in two focus group interviews, which led to useful comments that were used to revise the first draft of the instrument. The new version was tested in phase three via classroom observations. During three rounds of three observations in one school, the instrument was tested and revised after each round, based on the comments of the two observers and the scores calculated for Cronbach’s α. In the last, fourth, phase, classroom observations in another school took place to prevent missing out on information in the first school. Cronbach’s α (=0.731) and inter-rater reliability Cohen’s Kappa (=0.851) were calculated.

By going through these four phases, the research question was answered by delivering an observation instrument that is based on the characteristics it should contain; it meets all the requirements in the checklist (e.g. items need to be observable and there needs to be a clear distinction in scoring options) and it is built upon the five strategies of AFL (clarifying learning intentions and sharing success criteria, engineer effective classroom discussions and tasks that elicit evidence of learning, provide feedback that moves the students forward, activating students as owners of their own learning and activating students as instructional resources for one another). Implications for practice were found in using this instrument to give teachers feedback on their use of AFL, so they can adjust and improve their teaching. Implications for further research on this instrument were found in further developing this instrument, researching it in a larger sample and a different language, conducting research in which the role of the student is being observed instead of the role of the teacher, and eventually conducting research using the instrument instead of researching the instrument itself.

1 Thanks to my supervisors dr. K. Schildkamp and W.B. Kippers MSc
# Table of contents

Summary ........................................................................................................................................... 3  
Table of contents ........................................................................................................................... 4  
Word of thanks ................................................................................................................................. 5  
1. Introduction .................................................................................................................................. 6  
2. Theoretical framework .................................................................................................................. 7  
   2.1 Strategies to implement AfL in the classroom ........................................................................... 7  
   2.1.1 Conclusion .......................................................................................................................... 10  
2.2 Observation instrument .............................................................................................................. 11  
   2.2.1 Requirements for observation instruments .......................................................................... 11  
2.3 Research question ..................................................................................................................... 15  
3. Method and results ....................................................................................................................... 16  
   3.1 Developing an observation instrument .................................................................................... 16  
   3.2 Procedure .............................................................................................................................. 16  
   3.3 Phase 1: Screening .................................................................................................................. 18  
   3.3.1 Method .............................................................................................................................. 18  
   3.3.2 Results .............................................................................................................................. 18  
   3.4 Phase 2: Focus groups ............................................................................................................ 21  
   3.4.1 Method .............................................................................................................................. 21  
   3.4.2 Results .............................................................................................................................. 22  
3.5 Phase 3: Classroom observations ............................................................................................. 25  
   3.5.1 Method (round 1, 2 and 3) ................................................................................................. 25  
   3.5.2 Data analysis (round 1) ....................................................................................................... 26  
   3.5.3 Results (round 1) ............................................................................................................... 26  
   3.5.4 Data analysis (round 2) ....................................................................................................... 29  
   3.5.5 Results (round 2) ............................................................................................................... 29  
   3.5.6 Data analysis (round 3) ..................................................................................................... 31  
   3.5.7 Results (round 3) ............................................................................................................... 32  
3.6 Phase 4: Classroom observations ............................................................................................. 33  
   3.6.1 Method .............................................................................................................................. 33  
   3.6.2 Results .............................................................................................................................. 34  
4. Conclusion and discussion ............................................................................................................ 35  
   4.1 Conclusion ............................................................................................................................. 35  
   4.2 Discussion ............................................................................................................................. 36  
5. References ..................................................................................................................................... 39  
6. Appendices .................................................................................................................................... 43  
   6.1 Appendix I: Screening instruments (phase 1) ......................................................................... 43  
6.2 Appendix II: Instrument version I ............................................................................................. 47  
6.3 Appendix III: Comments on items and examples during focus group interviews (phase 2) ..... 50  
6.4 Appendix IV: Detailed information on changes made in examples during first revision ....... 53  
6.5 Appendix V: Instrument version II (revision after phase 2) ..................................................... 55  
6.6 Appendix VI: Instrument version III (revision after phase 3, round 1) ................................... 58  
6.7 Appendix VII: Instrument version IV (revision after phase 3, round 2) ................................. 61  
6.8 Appendix VIII: Instrument version V, final version (revision after phase 3, round 3) .......... 66
Word of thanks

It took me some time, but I did it! After about a year of hard work, I finished my graduation project for my masters in Educational Science and Technology. Though graduating was not always fun (I believe no one graduates without any bumps in the road) it has been a pleasure to work on my project and to see the results of what I had in mind in real life. Working on this amazing project would not have been possible without the help of some people, who I really want to thank.

First, my supervisors Kim Schildkamp and Wilma Kippers. It has been a pleasure working with you. In all of our meetings, I felt taken seriously and that boosted my self-confidence. I really wanted to prove myself, which led to writing a thesis I am proud of. I also want to thank you both for your constructive and clear feedback. Sometimes a question mark says more than a thousand words. And please do not be surprised if you have more students that ask for your supervision this year; I might have told a few of my fellow students you two were the best supervisors they could wish for.

Second, I would like to thank all participants in the focus group interviews. You gave me plenty of insightful comments that led to a completely changed instrument after the first revision. An instrument I could take to schools to test. Therefore, I would like to say thanks to all the teachers at Metameer Stevensbeek and Zwijsen College Veghel that gave me the opportunity to test my instrument in their classroom. This testing would not have been possible without my second observer Jaap. Thanks for waking up so early on your day off to visit these schools with me and thank you so much for your useful comments and patience when we had to wait a while or when I was nervous about the scores for Cronbach’s α or Cohen’s Kappa.

Third, I would like to thank Carolien for reading my thesis quite some times and helping me with all the bits of statistics I did not understand and methodological questions I had. Also I want to thank Petra and Linda for reading my thesis, performing a last spelling- and grammar check and helping me with the last questions I had.

Last, but definitely not least, I would like to thank all my friends and family who were there for me when I needed them to, especially mum and dad. Thanks for letting me move back home and giving me the time and space to work on this graduation project. Thanks for my old room and the kitchen table, where I spent hours writing my thesis. Thanks for always supporting me, despite the doubts you surely have had about my graduation project. Last special thanks I would like to say to my grandma, who passed away during my graduation project. Thanks for always being proud of me and believing in me. I wish you could have been there for this special moment, but we weren’t that lucky. I hope you look down from above, so you can still see me graduate.

And to all others who have helped me in any other way the past year: Thank you!
1. Introduction

‘I think this test is for our teacher, so she knows what she should explain better next time’, said a 10-year old (Omdenken, 2013). This citation addresses the heart of formative assessment. Assessment is not desired to solely be used to assess students by giving grades; it can be used for other purposes as well, for example assessing the progress of the students and the instruction of the teacher (Bennett, 2011). Formative assessment can be defined as assessment that is used to support student learning (Bennett, 2011; Black & Wiliam, 2009; Kippers, Schildkamp, Poortman & Visscher, submitted; Van der Kleij et al., 2015). It may be used as an umbrella term that covers diverse approaches of assessment that all have the goal to support learning, but have different underlying theories (Briggs, Ruiz-Primo, Furtak, Shepard & Yin, 2012; Van der Kleij et al., 2015). One of these approaches is Assessment for Learning (AfL). AfL can be seen as the more daily practice of formative assessment (Klenowski, 2009): it can be defined as minute-to-minute and day-to-day assessments, initiated by teachers and students, with the goal to enhance learning (Thompson & Goe, 2009). Within this definition of AfL, assessments can be defined as all the manners in which evidence about the progress of student learning is collected (Kippers et al., submitted; Van der Kleij et al., 2015). Examples of assessments are paper-and-pencil tests, classroom discussions, homework assignments or practical tasks. AfL informs students about their own learning and process, and it informs teachers about their students’ learning process and about their instruction (Cauley & McMillan, 2010).

AfL is of great importance to improve students’ learning and teachers’ instruction and thereby the quality of education and learning (Bennett, 2011; Kippers et al., submitted). Research has shown that in classes where AfL was implemented, students achieved learning gains in six to seven months what otherwise would have taken a year (Black & Wiliam, 1998). These results were found in various countries, ages and areas (Black & Wiliam, 1998). These gains can be extended over periods of time (Leahy, Lyon, Thompson & Wiliam, 2005) and held up in measurements with externally mandated standardized tests in the USA (Black, Harrison, Lee, Marshall & Wiliam, 2004).

The use of AfL has been studied in the past, but this is mostly done by using data collection methods that are based on the perception of respondents, such as questionnaires, interviews and checklists (e.g. Kippers et al., submitted; Lysaght & O’Leary, 2013; O’Leary et al., 2013; Wiliam, 2011). These studies give insight in the extent to which and how AfL is implemented in the classroom, but these results can be influenced by the perception of the respondents. It can occur that teachers are certain they use AfL, but observations may show that they do not.

So, in order to know where teachers stand in terms of implementing AfL in their classroom, it is necessary to observe them. An observation instrument that provides criteria against which teachers will be tested can be a good instrument to determine the extent to which teachers are using AfL in their classroom (Van Tassel-Baska, Quek & Feng, 2006) and can as well form a basis for feedback towards the teachers on their use of AfL.

The researched literature has shown there is only one observation instrument that measures AfL in the classroom (Oswalt, 2013) available as we know of, but this instrument does need improvement based on findings of the researcher (Oswalt, 2013) and based on the literature study conducted within this study. This led to the decision to develop an observation instrument to measure AfL in the classroom, building on the one Oswalt (2013) developed. In the future, this instrument can be used to determine where teachers stand in terms of their use of AfL in the classroom, so they know what can be improved and eventually how this may lead to better education (Leahy et al., 2005; Wiliam, 2011).

This thesis describes the process of developing an observation instrument to measure AfL in the classroom. Chapter two will provide the reader with theoretical background on AfL and (the development of) observation instruments and will present the research question. In chapter three, the research method and results will be elaborated based on the procedure followed to conduct this study. The last chapter of this thesis, chapter five, concludes and discusses this research and development project.
2. Theoretical framework

In this chapter, the main concepts of this research will be elaborated: five strategies of Assessment for Learning (AfL), and observation instruments, definitions and requirements.

2.1 Strategies to implement AfL in the classroom

Assessment for Learning, as defined in the introduction, can be divided into five strategies that can be helpful for teachers to implement AfL in the classroom (Leahy et al., 2005; Wiliam, 2011). These strategies are:

- clarifying learning intentions and sharing success criteria;
- engineering effective classroom discussions and tasks that elicit evidence of learning;
- provide feedback that moves the student forward;
- activating students as owners of their own learning, and
- activating students as instructional resources for one another.

Wiliam and Thompson (2007) formulated these strategies in the form of a framework, based on three key instructional processes: establishing where the students are in their learning, establishing where they are going, and establishing what needs to be done to get them to reach their goals and succeed (Ramaprasad, 1983, in: Wiliam & Thompson, 2007). This framework, shown in figure 1, gives a complete overview of the strategies, learning processes and actors.

<table>
<thead>
<tr>
<th>Where the student is going</th>
<th>Where the student is right now</th>
<th>How to get there</th>
</tr>
</thead>
<tbody>
<tr>
<td>Teacher</td>
<td>Clarifying learning intentions and sharing criteria for success</td>
<td>Engineering effective classroom discussions and tasks that elicit evidence of learning</td>
</tr>
<tr>
<td>Peer</td>
<td>Understanding and sharing learning intentions and criteria for success</td>
<td>Activating students as instructional resources for one another</td>
</tr>
<tr>
<td>Student</td>
<td>Understanding learning intentions and criteria for success</td>
<td>Activating student as the owners of their own learning</td>
</tr>
</tbody>
</table>

Figure 1. Framework Relating Strategies of Assessment for Learning to Instructional Processes (Wiliam & Thompson, 2007, p.63)

While traditionally these processes are all under the full responsibility of the teacher, when it comes to AfL, students need to have a role in this learning process as well. Therefore, the definition of AfL states: ‘…assessments, initiated by teachers and students…’ (Thompson & Goe, 2009). To make this distinction clear, Wiliam and Thompson (2007) made not only a distinction in instructional processes, but also in actors: teacher, peer and student.

In terms of where the student is going, the teacher has the responsibility to clarify the learning intentions and share the success criteria, peers have the responsibility to share the learning intentions, and students have the individual responsibility to make sure they understand them. Because in AfL these processes are not always under the full responsibility of the teacher, it can occur that success criteria will be set by teachers and students and not only by teachers. This can help students to be owner of their own learning.

When it comes to monitoring the students (where the student is right now), the teacher needs to engineer effective classroom discussions and tasks that elicit evidence of learning, the peers need to use each other as instructional resources (ask questions to their peers and help one another before asking the teacher), and the students need to be owners of their own learning (know what they are doing for what purpose and assessing their own work). For peers and students, these strategies (in the column where the student is right now) are also used to see what needs to be done to reach the goals.

In order to help students reach their goals and succeed (how to get where they should be going), feedback that moves the student forward is of great importance (Wiliam & Thompson, 2007). To provide students with this feedback, the teacher can use the evidence of learning gathered using the previous strategy. Not only does the teacher give feedback, in order for students to reach their goals, they can also take other instructional actions, such as explaining subject matter once again. Moreover, students and peers can provide each other with feedback.
The strategies by Leahy et al. (2005) and Wiliam (2011) and shown in the framework of Wiliam & Thompson (2007) will form the basis for the observation instrument that will be developed during this research and therefore will be explained more elaborately.

**Clarifying learning intentions and sharing success criteria**

This first strategy is about making students aware of what is expected from them. Learning intentions are statements that the teacher creates and which describe what the teacher wants the students to know, understand and be able to do after a (series of) lesson(s) (NCCA, 2015). Learning intentions are similar to learning objectives, but with learning intentions the emphasis is more on the process rather than on the end product (NCCA, 2015). Success criteria are developed by the teacher or by the teacher and students together, and describe what success looks like in the context of the learning intention (NCCA, 2015). For example, for the learning intention: the student can write a short essay in proper English, success criteria might be: the student uses the right amount of words, the student does not make grammar mistakes and the student elaborates the topic well.

Teachers need to clarify the learning intentions in a way students can understand them, because low achievement can be caused by students not knowing what is expected from them (Black & Wiliam, 1998; Wiliam, 2011). Both teachers and students need to understand how success is defined, because clearly formulated success criteria will help reaching the learning intentions and might eventually lead to education students can benefit from the most (Oswalt, 2013).

Proof that clarifying learning intentions and sharing success criteria can lead to increasing the understanding of students can be found in the research of Rust, Price and O’Donovan (2003). In their research, they found that students who were aware of the assessment criteria and assessment results had significantly increased achievement compared to students who were not aware of the assessment criteria and assessment results. White and Frederiksen (1998, in: Fletcher-Wood, 2003) found that students who were introduced to the assessment criteria scored significantly higher than students who did not. Sharing learning intentions and success criteria might therefore lead to better learning outcomes and even more effective education (Oswalt, 2013).

**Engineer effective classroom discussion and tasks that elicit evidence of learning**

This second strategy is about monitoring students: seeking to elicit evidence of learning by discussions and tasks in the classroom (Leahy et al., 2005; Oswalt, 2013). This is not so much about the discussions itself, but more about finding out what the actual knowledge of the students is (what they already know) and not what teachers assume students know (Wiliam 2011, in: Galileo.org educational network 2014). By asking the right questions, during classroom discussions or in assignments, teachers can see what students know or have learned (Leahy et al., 2005). An example of a good question can be ‘Why are 7 and 17 prime numbers?’ instead of ‘Are 7 and 17 prime numbers?’ The first question does not only show if the student has the knowledge to answer the question, but it also reveals the student’s thinking in answering the question. Next to effective classroom discussions, there are other tasks that can elicit evidence of learning, for example (homework) assignments or games played in the classroom that focus on what should have been learned, so the teacher can see how far along the students are in their learning process.

Asking the right questions will help monitoring the students on a day-to-day and minute-to-minute base, which can provide the teacher with information about student learning (Oswalt, 2013; Ruiz-Primo & Furtak, 2007; Wiliam 2011, in: Galileo.org educational network, 2014). When the teacher masters this strategy, it can increase the teachers’ ability to diagnose the state of student learning on a daily basis (Oswalt, 2013).

A study by Ruiz-Primo and Furtak (2007) in which they explored the informal formative assessment practices of teachers in three middle school science lessons, showed that teachers who used questioning and discussion according to this strategy gained significantly higher scores among their students than teacher who did not use this strategy. Though there were only three teachers participating in this research and therefore the generalizability is limited, these results do support the idea that effective classroom discussions and questioning may lead to improved student performances (Ruiz-Primo & Furtak, 2007).
Provide feedback that moves students forward

The third strategy is about providing students with feedback that makes them think about their work (Leahy et al., 2005). It can be seen as a response to the monitoring of student learning (Oswalt, 2013). Feedback can be defined as ‘information provided by an agent (e.g. teacher, peer, book, parent, self, experience) regarding aspects of one’s performance or understanding’ (Hattie & Timperley, 2007, p. 81). Assessment results, teachers and/or students can provide meaningful feedback to bridge the gap between students’ current and desired situation (Bennett, 2011; Black & Wiliam, 2009; Cauley & McMillan, 2010; Hattie & Timperley, 2007; Kippers et al., submitted; Sadler, 1989; Van der Kleij et al., 2015). This can for example lead to the teacher adjusting instruction based on what was seen in assessment results, or students adjusting their learning strategies based on the comments the teacher or other students made.

Feedback is a key element in improving student achievement (Hattie, 2009 in: Oswalt, 2013). It can be very effective and useful, but only when applied right. Feedback on the process level (how students achieve the goal and what ideas/strategies are used to achieve the goal, for example: you used the strategies that were discussed in class very well in your own assignment) and on the self-regulation level (the way students monitor, direct and regulate actions towards the learning goals, which can effect self-efficacy, self-regulatory skills and students’ own beliefs about him/her as a student, for example: I am impressed that you checked your answer in the solutions book, found out you were wrong and tried to adjust your answer) is considered powerful and thereby can contribute to reaching learning goals and to improved education (Hattie & Timperley, 2007). Also, the type of feedback (e.g. negative or positive) and the context (e.g. timing of feedback) are factors that influence the efficacy of feedback. Feedback is most effective when it is specific, descriptive, direct and it focuses on the work of the student instead of personal characteristics (Chappuis & Stiggins, 2002).

A study by Lyster & Siato (2010) on oral corrective feedback in second language acquisition showed that corrective feedback had a positive effect on the students. Students who received corrective feedback showed larger effect sizes compared to their performance on the pre-test than students who did not get corrective feedback. Though this research focused on second language acquisition, it supports the idea that feedback can improve student learning (Oswalt, 2013). A meta-analysis by Kluger and DeNisi (1996) showed that feedback interventions had on average a moderate positive effect on performance and a meta-analysis by Hattie and Timperley (2007) showed that the average effect of schooling is 0.40 and the effect of feedback on achievement is 0.79, almost twice the average effect. Thereby, the effect of feedback is in the top five of effects on achievement, which shows that feedback can be powerful (Hattie & Timperley, 2007).

Activating students as owners of their own learning

The fourth strategy is about self-assessment and making students aware of their level of understanding (Leahy et al., 2005; Oswalt, 2013). By being aware of their level of understanding, students are aware of where they are in their learning and how they can reach their learning goals. Self-assessment can be defined as students making judgment about the extent to which they have met the learning objectives and success criteria (Boud, 1991 in: Boud, 1995). It is more than just let students grade their own work, it is about involving students in the process of determining when their work is good in any given situation (Boud, 1995). Then, students and teachers share the responsibility for learning (Leahy et al., 2005). In order to let students use self-assessment and make them owners of their own learning, they have to be able to regulate their own learning (Wiliam, 2011). Self-regulated learning is about students taking control of their own learning by monitoring, directing and regulating actions toward the learning goals (Paris & Paris, 2001). This helps students to be aware of where they are in their learning and how they can reach their learning goals. This can be a first step in self-assessment; when a student is able to self-regulate their learning, assessing their own work and perhaps even giving feedback to themselves can be the next step.

Students who use self-assessment tend to score higher on tests, reflect more on their own work, take more responsibility of their own learning, and their understanding of problem-solving increases (Dochy, Segers & Sluijsmans, 1999).

A study of Harward, Allred and Sudweeks (1994) showed that for primary school students, scores on spelling of words increased when they immediately corrected themselves and Ross,
Hogaboam-Gray and Rolheiser (2002) showed that fifth and sixth grade students performed better on mathematics when the teacher implemented self-assessment strategies in the classroom (Brown & Harris, 2013).

**Activating students as instructional resources for one another**

The fifth and last strategy is about peer-assessment and peer-feedback (Leahy et al., 2005; Oswalt, 2013; Wiliam, 2011). Peer-assessment can be defined as ‘the process through which groups of individuals rate their peers’ (Falchikov, 1995 in: Dochy et al., 1999). Peer-feedback is about giving advice to peers about their work and how to improve it (Education Services Australia, 2016). Peer-assessment and peer-feedback are part of a learning process in which skills are developed. Giving feedback to peers or assessing the work of peers requires social skills and skills in assessing and giving feedback. Students have to be able to explain to one another why they assessed the work in a certain way or why they give certain feedback. This has to be done in a way both the receiver and giver benefit from it; that is where the social skills are important. Self-assessment can be quite difficult for students, but assessing the work of their peers is easier (students are more likely to find errors in the work of others) and both the assessor as well as the assessed student can benefit from it (Leahy et al., 2005).

One advantage of using peer-feedback is that students who assess and give feedback are forced to understand the assessing method (for example a rubric) and the work of the peer, which can give the student who gives feedback other insights on the subject and help him/her to understand it better. Another advantage of using peer-feedback is that communication between students is more efficient than communication between a teacher and a student; among students there are less communication barriers because they usually use the same language and way of communication. Also, students tend to be more engaged when the feedback is given by a peer (Leahy et al., 2005; Wiliam, 2011).

A study by Rust et al. (2003) among college students showed that students who are engaged in peer processes, which were designed to increase understanding of grading criteria, significantly increased achievement. They found that socialization processes are essential for implicit knowledge transfer to occur. Thereby, looking at younger students, it has been noted that they are much better in detecting errors in the work of their peers rather than in their own work, which can make peer-assessment and –feedback an important part of education (Leahy et al., 2005).

### 2.1.1 Conclusion

All the strategies described above can be used by teachers to implement AfL in their classroom, but not all teachers will implement them in the same way, caused by a difference in subjects, students and the way they teach. However, these strategies are the basics that define AfL and therefore are important to implement in any classroom (Leahy et al., 2005). Implementing these strategies gives teachers the opportunity to adapt education very fast, which can lead to better learning outcomes. Waiting for test results to come back and acting on them a week after the test has been taken may decrease the learning effect, because there is too much time gone by. Adapting education on the spot works better for both teachers and students (Leahy et al., 2005).

This does not necessarily mean that a teacher who implements not all strategies is a bad teacher when it comes to AfL. It is important that a teacher who wants to use AfL in the classroom implements at least one strategy for all of the instructional processes (where the student is going, where the student is right now and how to get them there). By implementing at least one strategy, all instructional processes are addressed and a start in using AfL has been made. In order to use AfL to its full extent, it is important to use all of the five strategies.

When a teacher is implementing these strategies, it can be helpful to know to what extent they are implemented and where improvements can be made towards implementing the strategies of AfL (Leahy et al., 2005). In order to gain this knowledge, an observation instrument to measure the extent to which AfL is implemented in the classroom will be developed. This instrument can give teachers insight in their instruction according to the strategies of AfL and so can lead to feedback that helps implementing these strategies better.
2.2 Observation instrument

Classroom observation can be described as ‘a performance-based assessment of the teacher within the context of the learning environment’ (Van Tassel-Baska, Quek & Feng, 2006, p.85) and is the most direct way to measure what is going on in a classroom (Womack, 2011). By using classroom observations, the behaviour (both conscious and unconscious) of the teacher can be measured (Baarda et al., 2013). Also, observations show directly what is happening during a lesson (Baarda et al., 2013; Womack, 2011). However, it needs to be taken into account that while observing people, there is always a chance that the observed ones behave slightly different than they would have done when they were not being observed. This makes it important to plan the observations in the most non-intrusive way possible, to get the most reliable results (Leff et al., 2011).

Stecher, Hamilton, Ryan, Robyn and Lockwood (2006) noted that it is easier to incorporate quality in observational ratings rather than in other methods, because the measurements are not dependent on the perception of respondents. However, the measurements are dependent on the perception of the observers; they have to observe and interpret what they see to translate that to a score or comment in the instrument. In order to keep the influence of the observers as small as possible, there are certain requirements an observation instrument needs to meet. These are listed in the next paragraph.

Though observations can be a quite expensive and time consuming method, it can be very useful (Womack, 2011). Specifically in this study, observations can form an important starting point for teachers in improving their use of AfL and thereby their instruction in the classroom. An observation instrument can give them a clear insight in how they are doing for each individual strategy and for the entire concept of AfL. This can provide them with clear examples of how to improve. After some time, observations may take place again and show them their progress. By using an observation instrument to do this, the teachers are certain that what is noted is what has been seen, and there is no guessing of any kind involved.

2.2.1 Requirements for observation instruments

When developing an observation instrument, there are several requirements that have to be met in order to develop an instrument that can actually work. This study focuses on 21 requirements that need to be taken into consideration (e.g. Boehm & Weinberg, 1977; Croll, 1986; Danielson, 2012; Grossman, 2011; Harkink, 2013; Leff et al., 2011). These requirements were found during a literature study and are linked to five overarching categories: formulation of items, feasibility, scoring, quality and usability.

Formulation of items

The first category is about formulation of items. This is about the language used to enhance comprehensibility of the instrument and the formulations of items in a way it helps the observers and increases the reliability of the instrument. The more clear items are formulated, the less confusion they can cause and the more the reliability of the instrument can be influenced positively (Boehm & Weinberg, 1977; Womack, 2011). In this category, there are seven requirements to be met.

The first one is that the items must be formulated in language that can be easily understood. This can help prevent misunderstandings among observers, which can lead to better reliability in the instrument (Boehm & Weinberg, 1977; Danielson, 2012). An example of an understandable item might be the teacher lets students assess their own work instead of the teacher implements strategies of self-assessment. It might occur that the observer does not know what is exactly meant by self-assessment, so then this item is not written in understandable language, which can cause confusion and decrease the reliability of the instrument. This can be prevented by using other words which describe self-assessment in a way all observers should be able to understand it.

The second requirement is that words need to be chosen in a manner that they can exactly describe the observed behaviour (Danielson, 2012; Grimm, Kaufman & Dory, 2014; Leff et al., 2011). This has to do with the item describing what can be seen in the classroom, in order to limit the risk of interpretation errors from the observers (Grimm et al., 2014). For example the teacher writes the learning goals on the whiteboard instead of the teacher tries to make students aware of what the learning goals are. In the first item the behaviour is clearly described and observable and the second
item shows what the teacher aims to do, but does not describe observed behaviour. In the second item, there is a greater risk of interpretation errors than in the first one and so the first one is better fitted for an observation instrument (Danielson, 2012).

The third requirement is that items need to be mutual exclusive. This means that one item cannot be a prerequisite to meet another item (Boehm & Weinberg, 1977). For example, the item the student answers questions the teacher asks is not mutual exclusive. In order to score this item, the part of the item the teacher asks questions must be observed, otherwise the other part of the item cannot be scored. Better would be the teacher asks questions in language that fits the students. This way it is all about the action the teacher takes (asking questions) and there is no action (answering them) that follows, which makes the item mutual exclusive.

The fourth requirement in this category is that items cannot be multi interpretable. The items have to be formulated in a way that all observers that may use the instrument, interpret it the same way. This prevents errors that may influence the reliability of the instrument (Boehm & Weinberg, 1977; Womack, 2011). For example, the item the teacher tries his best to make student understand a concept can be interpreted in many ways. One observer may find he does try his best while another may find he does not try his best at all. Better would be to formulate the item like the teacher gives a definition of the concept.

The fifth requirement to be met is that items need to be observable (Joe, Tocci, Holtzman & Williams, 2013). Though this is a quite logical requirement for an observation instrument, it can help to be aware of this when formulating the items for the instrument. There may be some items that can provide useful information, but that cannot be observed in the situation the observation takes place in, like students inform their parents on their progress in school. This item is not observable in the classroom and therefore has no place in this observation instrument.

The sixth requirement is that items need to fit the indicators in the instrument (Harkink, 2013). Observation instruments are in many cases built upon indicators and items; the indicators being the overarching construct that needs to be observed and the items being the smaller bits of the indicators that can be scored (Harkink, 2013). In the case of this instrument, there are also examples given for each item, but these do not have to be scored. The indicators in this instrument are equal to the strategies for AfL by Leahy et al. (2005) and Wiliam (2011). The items in this instrument need to be formulated in a way they support the indicator, in this study, the strategies. This will increase the internal consistency of the instrument, measured by Cronbach’s α (Harkink, 2013). For example, the item the teachers makes clear what the learning goals are using language that fits the students fits the strategy clarifying learning intentions and sharing success criteria.

The seventh requirement is closely linked to the sixth; examples need to fit the items in the instrument (Harkink, 2013). When examples are used to elaborate the items more clearly, it is important that these examples support the items and indeed make it clearer, instead of creating misunderstandings. For example, for the item the teacher uses different methods to find out what the prior knowledge of the student is, examples can be the teacher initiates a class discussion or the teacher uses brainstorming in the classroom. These examples show what is meant by the item and can therefore be helpful in observing this item in the classroom.

Feasibility

The second category, feasibility, is about how feasible the items are in terms of the observations. This may have to do with the time, lesson, type of school or country the observation takes place in. If there are items in the instrument that cannot be observed within the specific situation the observation takes place in, the results can never be reliable.

This category consists of one requirement: observers have to be able to observe the items in the time given for the observation (Joe et al., 2013). If there is only one lesson that can be observed, items that can only be observed in more lessons cannot be included in the instrument. For example: All the lessons have the same structure. Though this might provide useful information, it cannot be observed in one lesson and therefore is not going to be included in the instrument.


**Scoring**

The third category, *scoring*, is about the scale and scoring used in the instrument. Scale and scoring are the base on which you observe and score the teachers, so this has to be good considered the instrument, otherwise it can cause misinterpretations. This may lead to advices that are not applicable to the situation or decreasing reliability of the instrument. This category has five requirements that need to be met.

The first one is that the length of the scale, the number of scoring options, has to be optimal for the purpose of the instrument (Croll, 1986). This requirement is hard to assess, but it can be done by researching options for the length of the scale. For example, looking at other observation instruments and their scales, and study why this length of the scale has been chosen and whether it worked or not.

The second requirement in this category is that the explanation on the scoring must be formulated in language that can be easily understood. It is important that the spoken and written explanation of the scoring is clear for all observers to prevent misunderstandings and interpretation errors that may lead to decreasing reliability of the instrument (Croll, 1986).

The third requirement is that scoring needs to be described in a qualitative rather than a quantitative manner (Danielson, 2012). Instead of using descriptions like *never, sometimes* and *always*, it is better to use descriptions like *strong, more strong than weak, more weak than strong and weak*. When using quantitative descriptions, chances are that only the score *sometimes* will be used: it is impossible for an observer to conclude that a teacher does a certain thing always or never when only part of the teacher’s instruction is being observed. Therefore, it is better to choose qualitative descriptions: an observer can say, based on a part of the teacher’s instruction that the shown behaviour is weak, strong or somewhere in between (Danielson, 2012).

The fourth requirement is that responses must be scored unambiguously, by using the same scale and scoring (Leff, et al., 2011). This can be handled by using the same instrument, which automatically contains the same scale and scoring, for every observation. In case the observed teachers need to be compared, the unambiguously scoring helps to compare them all on the same grounds (Leff et al., 2011).

The fifth and last requirement in this category is that there needs to be a clear distinction in scoring options. This helps the observers to choose what option is fitted for the observed behaviour (Boehm & Weinberg, 1977; Danielson, 2012; Womack, 2011). Unclear distinctions may lead to wrong interpretations, which may lead to decreasing reliability of the instrument.

**Quality**

The fourth category, *quality*, is about the quality of the instrument. Although all requirements contribute to the quality of the instrument, the requirements below are not fitted in any other category, but are important for the quality of the instrument. Only when the quality of the instrument is good enough, the instrument will be of added value. Within this category, there are five requirements that need to be met.

The first one is that the instrument needs to provide an objective view to assess the teacher. This means the items and scoring must be formulated in a way observers can score it as objective as possible. Clearly described scoring options and items formulated according to the requirements in the category *formulation of items* will help meeting this requirement. This way, the reliability of the instrument is being kept as high as possible (Grossman, 2011).

The second requirement is that the instrument must allow to rate teachers across different types of lessons (Grossman, 2011). Though the instrument developed in this research is focused on AfL, there are still different types of lessons that can be observed, for example giving instruction (classical or individually), working individually, testing and giving feedback. The instrument needs to be developed in a way it can assess more than one of these types of lessons.

The third requirement is that the instrument must be considered valid (Kimberlin & Winterstein, 2008). Validity, and more specifically content validity, is about the extent to which the instrument measures what it is supposed to measure (Dooley, 2001). In this case: the extent to which the instrument measures the use of AfL in the classroom.
The fourth requirement is that the instrument must be considered reliable (Kimberlin & Winterstein, 2008). Reliability in this study is divided in inter-rater reliability and reliability in consistency among items within the instrument. A way to establish the inter-rater reliability is to calculate Cohen’s Kappa. This is a number between 0 and 1 which shows to what extent the two observers agree (Stemler, 2001). Cohen’s Kappa needs to be between 0.61 and 0.80 to be considered substantial and above 0.81 to be considered nearly perfect (Landis & Koch, 1977 in: Stemler, 2001). Another form of reliability is about consistency among items within the instrument (Dooley, 2001). In order to establish that kind of reliability, Cronbach’s α can be used. Cronbach’s α is about the extent to what items measure the underlying construct. This coefficient gives information about the length of the instrument (should it be longer to be more reliable?) and the inter-item correlation (will adding more reliable items result in a higher coefficient?) (Dooley, 2001).

The fifth requirement is that all factors that can influence the observation need to be written down in the instrument (Stuhlmans, Hamre, Downer & Pianta, 2010, in: Harkink, 2013). This way, the researcher can see if any irregularities might be due to the circumstances the observation was taken under, for example the first lesson after a break, the last lesson before the weekend starts or a lesson just before or after a test; this can all have its influence on the instruction of the teacher and thus on the scoring in the instrument. By adding general information at the start of the instrument, in which the observer has to fill in the name of the school and teacher, date and time, class, subject and type of lesson and number of students and his/her own name, this requirement can be met easily.

**Usability**

The fifth category, usability, is about how easy to use the instrument is. These requirements are all about making the use of the instrument as easy as possible. The easier to use the instrument is, the more time the observers have to look around and actually observe. In this category, three requirements need to be met.

The first requirement is that the instrument must allow comparing teaching across classrooms (Grossman, 2011). One teacher may act in another way in different classrooms; therefore the instrument must be developed in a way it can be used in more than one classroom.

The second requirement in this category is the instrument needs to be designed in a way that results can be interpreted directly and feedback based on the results can be given to the teacher (Harkink, 2013). It may occur that the observed teacher wants to know what has been observed and how he/she can improve instruction based on the findings. In this case, it is advisable to develop the instrument in a way it gives a direct overview, this can been done by using clear scoring options, on which feedback can be based.

The third requirement is that recording must take place in a feasible and non-intrusive way using a paper-and-pencil format (Leff et al., 2011). The paper-and-pencil format helps observing as structured as possible, which helps to perform the observations the same way every time an observation takes place. This makes it easier to compare observations. It is important to perform the observations in a non-intrusive way, because it gives the most reliable and veracious results. By sitting quietly in the back of the classroom or by videotaping the lesson and observing it later on, observation can be done as non-intrusive as possible.

All the requirements together lead to the checklist of requirements that is shown below.

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Checklist of requirements</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Category</th>
<th>#</th>
<th>Requirement</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Formulation of items</td>
<td>1a</td>
<td>The items must be formulated in language that can be easily understood.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1b</td>
<td>Words need to be chosen in a manner that they can exactly describe the observed behaviour.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1c</td>
<td>The items have to be mutual exclusive.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1d</td>
<td>The items cannot be multi interpretable.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1e</td>
<td>Items need to be observable.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1f</td>
<td>Items need to fit the indicators in the instrument.</td>
<td></td>
</tr>
</tbody>
</table>
When all the requirements mentioned above are met, the instrument should be ready to use to observe AfL in the classroom. In order to develop an observation instrument using the five strategies for AfL (Leahy et al., 2005; Wiliam, 2011) that meets all the requirements, this research will be conducted guided by the research question below.

### 2.3 Research question

The question that will be endeavoured to answer in this research is as follows:

*What are the characteristics of an observation instrument to measure Assessment for Learning in the classroom?*


3. Method and results

This chapter describes the research procedure and methods used to conduct this research. The conducted study can be defined as a design based study. Because all phases build upon the results that are gathered in the prior phase (e.g. phase two builds upon the results of phase one), the results will be presented in this chapter as well.

3.1 Developing an observation instrument

In order to develop an observation instrument to measure AfL in the classroom, it is important to first look at what instruments are already available. This search resulted in one instrument that measures AfL: the instrument of Oswalt (2013). In this instrument, the five strategies of AfL are used as categories. Within these categories, three to five items were used to score the use of AfL on a five-point-scale ranging from 1: Not observed at all/ Not demonstrated at all to 5: Observed to a great extent/ Demonstrated to a great extent. There were no examples given. According to Oswalt (2013), his instrument needs improvement, since it focused mainly on the question if AfL is observable in the classroom and not so much how it should be observed in the classroom. Looking at the requirements that are listed above, the instrument of Oswalt (2013) does not meet all the requirements for observation instruments and therefore needs improvement as well. Mainly the requirements in the category scoring are not met, so in order to develop an instrument that has proper scoring, another instrument to draw inspiration from has to be found. The items used in the instrument of Oswalt (2013) met most of the requirements and can therefore form an inspiration for the new instrument.

The second instrument that will be used to draw inspiration from is the ICALT instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012). This instrument is developed to assess didactical skills of teacher and can therefore not form an inspiration when it comes to the content of items, but it can be used when it comes to scoring and the instrument itself. The ICALT-instrument has been tested widely: 845 mathematics lessons were observed (Van der Grift, 2007) and in later research, 1319 teachers from various European countries have been observed using this instrument (Van der Grift & Van der Wal, 2012). The results of both studies show that the ICALT-instrument as it is available these days is a valid and reliable instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012). Though the ICALT-instrument serves another purpose than the instrument developed in this research, both instruments are focused on observing teachers in one lesson and therefore it is safe to assume that the structure used in the ICALT-instrument will be working in the new instrument as well. The use of indicators and examples is expected to work for a complex concept like AfL, because this gives a more clear insight and examples that can actually be found in practice. This will make it easier to observe, because the observers are given clear pointers on which they can focus.

With the observation instruments described above, a start is made to develop the new instrument. However, there is one instrument that may be very helpful in developing the new instrument that has not been mentioned yet: a questionnaire on AfL developed by Kippers et al. (submitted) and Wolterinck, Kippers, Schildkamp and Poortman (2016). Because this instrument is a questionnaire and not an observation instrument, the scoring and usability requirements will not apply, but the content of items can form a good inspiration for the new instrument.

The three instruments as described above and the found literature on the strategies of AfL and the requirements of observation instruments will form the main source of inspiration to develop the new instrument. In order to develop the first draft of the instrument, these instruments will all be screened using a checklist of requirements (table 1). The procedure that has been used to develop this instrument is elaborated in the next paragraph, followed by a detailed methodological description of all the phases and the results gathered during this research.

3.2 Procedure

The procedure to develop an observation instrument to measure AfL in the classroom is based on the procedure Oswalt (2013) followed developing his instrument. The procedure for this research is shown in the figure below and consists of four phases.
*Observations will take place during the subjects Dutch, English and mathematics. The same lesson will be observed for 3 weeks. These same subjects will be observed in school 2.

**Figure 2.** Research procedure to develop the observation instrument, based on Oswalt (2013).

The first phase in this procedure was conducting a literature study (chapter 2), studying the strategies of AfL and requirements observation instruments must meet. The literature study covered 47 articles/books about both AfL and observation instruments. Among these articles were the articles about the (observation) instruments used to draw inspiration from (Kippers et al., submitted; Oswalt, 2013; Van der Grift, 2007; Van der Grift & Van der Wal, 2012; Wolterinck et al., 2016). The information about observation instruments and about the characteristics of AfL derived in this literature study led to the checklist of requirements. The three instruments were screened using the checklist and this resulted in the first draft of the instrument.
The second phase in this procedure was to evaluate the first version of the instrument. This was done by conducting two focus group interviews; one in which educational researchers participated and one in which secondary education teachers participated.

According to Oswalt’s (2013) procedure, the next step is evaluating the instrument in the setting it was developed for: a classroom (phase 3 and 4). In contrary to Oswalt’s study (2013), who first observed eight lessons and in a later stadium 16 teachers in three schools, due to limited time, this instrument was tested in only two schools with a total of six different teachers. Of those six teachers, three were observed three times (phase 3), to prevent having one lucky shot (Oswalt, 2013). The other teachers (phase 4) were observed once, to make sure there are no items missing or redundant that could not be concluded based on the observations in the first school. All observations were conducted by the same two observers.

The phases in this procedure as well as the respondents and data analysis will be elaborated in the next paragraphs.

### 3.3 Phase 1: Screening

Developing an observation instrument to measure AfL in the classroom started by conducting a literature study. During this literature study, information on the characteristics of AfL, on observation instruments (e.g. Oswalt, 2013 and Van der Grift & Van der Wal, 2012) and on a questionnaire about AfL (Kippers et al., submitted; Wolterinck et al., 2016) were found. In order to use this information to develop the new instrument, a checklist of requirements has been used (table 1). This checklist has formed the basis for the screening of the instruments which led to the first draft of the observation instrument.

This checklist of requirements is conducted based on requirements that were found in the literature and consists of requirements like *items must be formulated in language that can be easily understood, items need to be observable and scoring needs to be described in a qualitative rather than a quantitative manner*. These requirements are categorized in five categories: *formulation of items, feasibility, scoring, quality and usability* and are described more elaborately in the theoretical framework (paragraph 2.2.1).

#### 3.3.1 Method

**Procedure and data analysis**

In order to develop the first draft of the observation instrument, the observation instrument of Oswalt (2013) and the ICALT- instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012) were screened using the checklist of requirements. Screening these instruments revealed the strong points and the points that needed improvement in a systematic way: screening has been done the same for all the instruments. The instruments were screened by working through the entire checklist for each instrument and note the requirements the instrument did not met, so they could be adjusted in developing the new instrument.

To generate more items for the instrument, the questionnaires on AfL (Kippers et al., submitted; Wolterinck et al., 2016) were screened for the requirements as well. Because these questionnaires were only used to draw inspiration from the items, only the items were screened. All items have been written down and the requirements that were not met were noted. For the items in the instrument of Oswalt (2013), the same has been done.

The results of the screening were used to develop the first draft of the instrument. This draft was then checked again using the checklist of requirements to make sure all requirements were met.

#### 3.3.2 Results

**The ICALT-instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012)**

The first instrument that was screened was the ICALT instrument (Van der Grift & Van der Wal, 2012). Because this instrument does not focus on AfL, the content of the items was not useful, so only the categories *scoring, quality and usability* were used.
In the category *scoring*, no remarks were made. In the categories *quality* and *usability* there were some requirements (5a, 5b, 5c, 5d and 6a) that could not be screened. These requirements are very closely related to the content of the instrument and since the content of the ICALT-instrument differs from the new instrument, assessing these requirements was not of added value. The other requirements in these categories were all met. This led to the choice to use the scoring and layout of the ICALT instrument as most important inspiration for the new instrument.

The instrument of Oswalt (2013)
The second instrument that was screened was the instrument of Oswalt (2013). This instrument focuses on AfL and so all categories were used to screen this instrument.

For the category *formulation of items*, comments were mostly on items not formulated in language that can easily be understood (1a) and items not being observable (1e). For the category *feasibility*, comments were about items not able to observe it in the given time (2a). These items related to two categories were either removed or changed. For example, the item *Does the teacher utilize the results of peer activities to strengthen ongoing assessment of student learning?* has been removed, because this is not feasible, and the item *Does the teacher give students opportunities to use self-regulatory competencies, such as the ability to accurately assess their own knowledge?* was changed to *the teacher stimulates students to assess their own work in order to fit in the new instrument*. Decisions made per item are to be found in appendix I.

In the *scoring* category, conclusions were drawn that scoring was too limited and there was no explanation written down in the instrument (3b, 3c, 3e). The explanations that were found elsewhere in the document were not elaborated clearly and there was more than one description for a score (for example 4: *frequent or effective use*). Though the length might be optimal (3a), the scoring options need to be improved to be used in this instrument. This, in combination with the scoring options in the ICALT-instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012) that met all the requirements, led to the decision to not use the scoring in Oswalt’s instrument (2013).

In the category *quality*, it was seen that the instrument does provide an objective view to assess the teacher, but this view might be too limited (4a, 4b), since there are only items and no examples given. More items or examples need to be added to widen this view to the desired extent. Validity and reliability are sufficient, according to the research (4d, 4e), though this does not necessarily translate to the new instrument. The factors that may influence the observation are not noted on the form (4e).

In the last category *usability*, it was seen that the instrument does allow comparing teachers across classrooms (5a), that the instrument is designed in a way the results can be interpreted directly (5b), and it is a paper-and-pencil format (5c), which all together makes it easy to use.

From the 20 items in Oswalt’s instrument (2013), four were removed. All other items were rewritten into sentences that started with *The teacher...* and six were rewritten because of formulation problems (e.g. used language was too difficult, items were multi interpretable). Though the instrument is usable and the quality is considered good according to most requirements, the scoring does lack clarity. These findings did not add valuable information that was not already captured in the scoring, quality and usability of the ICALT- instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012), so the layout was not adapted to the instrument of Oswalt (2013).

Questionnaires on AfL (Kippers et al., submitted; Wolterinck et al., 2016)
The third instrument is a questionnaire and not an observation instrument, and therefore only the first category, *formulation of items*, was assessed. All comments on the items were about the requirements regarding items have to be formulated in language that can easily be understood (1a) and items have to be observable (1e). An example of an item that was changed because of its comprehensibility is *Diagnostic information from tests is used to determine strong and weak points in instruction* that has been changed to as *The teacher knows the weak points of the students and adjusts instruction to them*. Though this instrument is a questionnaire and not an observation instrument, most items were formulated in a way they were observable. The five items (out of a total of 31 items) that were not observable (for example *students are formally concerned with the information about their learning progress towards their parents by showing them portfolios at home*) were removed. All the 26 items
that were useable, the ones that did meet all the requirements or could be rewritten into an item that met all the requirements, have been used to develop the new instrument. Decisions made per item are shown in appendix I.

**The first draft of the new instrument**

The information retrieved by screening the instruments mentioned above led to the development of the first draft of the instrument. Layout and scoring were derived from the ICALT-instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012): the items are scored on a four-point-scale (1: predominantly weak, 2: more weak than strong, 3: more strong than weak, 4: predominantly strong) and examples were scored on a two-point-scale (1: not observed, 2: observed). Observers first have to score the examples before they score the items. All the factors that may influence the observation (date, school, observed group, teacher and subject and observer) were written down at the top of the instrument. The instruction on how to score is written below these factors and after all this information is filled in and read, the instrument itself is shown.

Items were derived from the instrument of Oswalt (2013) and the questionnaires on AfL (Kippers et al., submitted; Wolterinck et al., 2016) and were rewritten to fit the new instrument. The strategies to implement AfL in the classroom by Leahy et al. (2005) and William (2011) were used as indicators.

The new instrument consists of three levels: indicators (based on the strategies), items and examples. For each indicator, one or more items are composed and for each item, two or more examples that elaborate the item more clearly are composed. Because the items found in the instrument of Oswalt (2013) and the questionnaire on AfL (Kippers et al., submitted; Wolterinck et al., 2016) were not split in items and examples, this distinction is made by the researcher based on which items were more overarching and which were more specific. For example, the item *the teacher makes clear the learning intentions for this lesson are* is overarching and so it is an item in the instrument. The item *the teacher makes clear the learning intentions in understandable language for the students* is specific and thus an example for the earlier mentioned item. Some of the examples in the instrument were added by the researcher in order to give a broader view of the corresponding item. In order to complete the first draft of the instrument, experts in the field were consulted to give their opinion, based on which the instrument was adjusted towards the first draft that is used as starting point for developing the final instrument.

**Instrument: version I**

The first version of the instrument is spread over three pages and consists of five indicators (similar to the strategies of Leahy et al. (2005) and William (2011)), ten items and 57 examples.

On top of the first page, general information like observation date, school, observed class, subject and teacher, number of students and the observer’s name has to be noted. Afterwards, a quick reminder of how the scoring works is written down; examples are scored on a two-point-scale (0: not observed, 1: observed) and items are scored on a four-point-scale (1: predominantly weak, 2: more weak than strong, 3: more strong than weak, 4: predominantly strong).

For the first indicator (clarifying learning intentions and sharing success criteria) there are two items with respectively seven and eight examples, for example the item *the teacher makes clear what the learning intentions for this lesson are* with the example *the teacher makes clear the learning intentions before the lesson starts.* For indicator two (engineer effective classroom discussion and tasks that elicit evidence of learning) there are three items with respectively five, five and eight examples, for example the item *the teacher uses different methods to find out what the students have learned/already know* with the example *the teacher initiates a classroom discussion.* Indicator three (provide feedback that moves the student forward) consists of three items with respectively four, six and two examples, for example the item *the teacher gives feedback on what needs to be learned* with the example *the teacher gives feedback based on criteria.* The fourth indicator (activating students as owners of their own learning) consists of one item with seven examples, the item being *the teacher stimulates students to monitor their own progress* with one of the examples being *the teacher stimulates students to assess their own work.* The fifth indicator (activating students as instructional resources for one another) consists of one item with five examples, the item being *the teacher...*
stimulates the students to assess the work of peers and give feedback to peers with one of the examples being the teacher reserves time during the lesson the assess the work of peers. The first draft of the instrument is to be found in appendix II.

3.4 Phase 2: Focus groups

The first draft of the observation instrument has been developed by screening other instruments and questionnaires, as described in phase 1. In order to evaluate the instrument and determine validity, two focus group interviews took place. A focus group is a group interview that focuses on a specific subject and that will be led by a moderator (Sim, 1998). One of the main aspects of focus groups is the discussion among respondents (Morgan, 1996). By letting educational researchers and secondary education teachers (both in a separate group) evaluate the instrument in a group, discussion can lead to new insights and more information than when it is evaluated in a one-on-one interview. In one focus group, six to ten respondents participate (Kitzinger, 1995). In this research, the respondents in these focus groups have evaluated the instrument.

3.4.1 Method

Respondents

The focus group interviews were organized in two groups: one with educational researchers and one with secondary education teachers. These groups were chosen so the instrument could be evaluated from both a scientific and a practical point of view.

The participants in both focus groups were chosen based on a convenience sample: within the larger group of potential participants that fit the criteria (being an educational researcher with expertise in either observation instruments or AfL, or being a secondary education teacher), these were available and willing to participate in this research (Onwuegbuzie & Leech, 2007).

The six educational researchers who participated in the focus groups are from various educational institutions in the Netherlands (University of Twente, the Dutch foundation for curriculum development SLO, the Dutch central institution for test development CITO and the Dutch board of examinations CvTE) and have been selected based on their expertise with either AfL or with observations as data collection method. The teachers who participated in the focus groups are six secondary school teachers (two teacher in general secondary education and four in the subjects science, chemistry, German and history) from four different schools and school types (three comprehensive schools and one school for lower secondary vocational education) in the Netherlands. Because observations in phase 3 and 4 are taking place in secondary education, the participants in the focus group interviews were secondary education teachers. With this group of teachers, that varied in experience from two years up to more than 30 years, in different schools and subjects, there are six different views on the instrument given.

Both focus groups are composed heterogenic. This has been done to incorporate different perspectives and expertise within each group. An advantage of heterogenic groups is that you get opinions from respondents from more than one point of view, assuming that a point of view could be formed by a job or organization. The hierarchy problem that could have occurred is solved by composing two separate focus groups with both respondents who have more or less the same profession (Kitzinger, 1995).

Procedure

Both focus groups were organized the same way. The first group (educational researchers) took one and a half hour to complete and the second group (teachers) was done within an hour. After giving the respondents some insight in the research they are participating in and elaborating the definitions they need to know to participate in this focus group, a card sort exercise, to establish the validity of the instrument, started. In this 30-minute exercise, all respondents (in duos) received a stock of cards with all the aspects of the instrument written on them. Green cards for the indicators, yellow cards for the items, white cards for the examples and one orange card to represent the other category. It was their task to compose an observation instrument by grouping the items in one of the five indicators and the examples in one of the items. The items and examples that did not fit any indicator or item could be
categorized in the category other. The respondents were invited to make notes on the cards regarding confusion or lack of clarity within the items. After this exercise was done, the researcher photographed all the composed instruments and collected all the cards, so later on the notes could be read.

After the card sort exercise, the respondents got to see the instrument (up to this point only the individual items have been showed). They got about 10 minutes to take a look at it and compare it to the instrument they had composed in the card sort exercise. Afterwards the discussion part of the focus group interview started. The respondents discussed the instrument with each other, led by the questions the researcher asked. These questions were asked based on the checklist of requirements. For example, it was asked what the respondents thought of the formulation of certain items, how they evaluated the scale and scoring, and what they would like to see changed. As a last question, the researcher asked if there were any remarks or things that were missing in the instrument. This discussion took about 40 minutes.

Data analysis
Analyzing the focus group interviews was divided in three parts.

First, the card sort exercise. The results from the card sort exercise were photographed and the cards were collected by the researcher. In order to analyze the results of this exercise, the researcher combined all the information from the three pairs (six respondents per focus group) in each focus group in one document. This gave an overview of which items were put where during the exercise and this document was used to revise the instrument.

During the card sort exercise, the respondents were invited to write comments down on the cards. These comments were written down in a separate document (appendix III) and were analyzed by coding the comments using the checklist of requirements. Comments on individual items or examples that were made during the discussion were added to this document as well. Codes used were formulation of items, feasibility, scoring, quality, usability and other. By creating a separate document for these comments, the researcher had a clear overview of what needed to be revised in the instrument.

The third part of the analysis of the focus group interviews was the discussion part. In order to analyze this part, the discussion was videotaped and transcribed. This transcription was analyzed by coding it, using the checklist of requirements. Codes used were formulation of items, feasibility, scoring, quality, usability and other.

After all the comments (written and in the discussion part of the interview) were coded, all comments per code were written down. For each code the researcher read the comments again and made a decision to make a change based on the comment or ignore the comment. This has been done for all codes.

3.4.2 Results

Comments in focus group interviews
Based on the codes from the focus group interviews, the instrument was revised. This has been done by reading the transcription carefully and paying attention to the comments the participants made that could be helpful in revising the instrument.

For the code formulation of items, most remarks were made on items (not) fitting the indicator or examples (not) fitting the item. These comments were made based on the card sort exercise and were dealt with in the processing of the card sort exercise.

Comments on scoring were mostly on the scoring of the examples. One of the respondents in the educational researchers group was trained in using the ICALT-instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012). This respondent told that during training they learned not to score the examples, but just score the items. Other respondents in this focus group could agree on this. In the teacher focus group, this issue was addressed as well. All the teachers agreed that observing the examples would be very hard, because they do not occur in the classroom that often. However, the examples would be great to give some practical implications. This, together with the amount of examples that was considered too big by both focus groups, led to the decision to remove the scoring of the examples, but keep a smaller amount of examples to illustrate practical implications of the
items. During ICALT-training, the respondent also noticed that the scale of four scoring options was interpreted as two options: the options 1 and 4 were rarely used and therefore, a yes/no scale would be sufficient. The teachers however found the 4-point scale to be very useful. This led to the decision to keep the 4-point scale and see how scoring is done in the next phase. If it does turn out to be used as a yes/no-scale, this adjustment could have been made later on.

The code usability was used most in the teacher group. With the remark that there were less examples needed and that examples did not have to be scored, the teachers thought the instrument was easy to use and easy to draw information from with regard to their own instruction. Educational researchers thought the instrument would be usable with these adjustments as well.

The code quality was used only once in the educational researchers group. This was during a discussion about the use of the ICALT-instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2011) and regarded a comment about the inter-rater reliability. This comment was not of any importance in revising this instrument in this phase of the research.

The code other has been used often, because the discussion was not only about the requirements, but also about the situation in the classroom, how to talk to teachers about results and other boundary conditions that are of great importance when using the instrument, but cannot be taken into account when designing the instrument itself. These remarks come in handy when it comes to teaching observers how to use the instrument and how to give feedback to teachers based on the instrument.

After the focus group interviews took place, there was a suggestion made by an educational researcher (who was invited to participate in the focus groups, but could not be there) that for the first strategy, the focus was mainly on the teacher and not on the students and peers, which could cause confusion and could lead to teachers thinking they are doing well in terms of AfL when in fact they are not. This led to adding two items in the instrument that will be elaborated below.

**Revision of the instrument**

With this first design decisions being made, the revision of the first draft started by comparing the original instrument to outcomes of the card sort exercise for both groups. Based on these outcomes, changes were made. All changes that are made in order to improve the instrument are displayed in the table below.

**Table 2**

*Changes made in the instrument during first revision*

<table>
<thead>
<tr>
<th>Item</th>
<th>Formulation of items</th>
<th>Feasibility</th>
<th>Scoring</th>
<th>Quality</th>
<th>Usability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clarifying learning intentions and sharing success criteria</td>
<td>Item 1 and 2 were rephrased</td>
<td>Item 1 and 2 were rephrased</td>
<td>Scoring for examples has been removed, description of scoring has changed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Two items were added</td>
<td>Two items were added</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Examples were rephrased, combined, replaced or removed</td>
<td>Examples were rephrased, combined, replaced or removed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Engineering effective classroom discussions and tasks that elicit evidence of learning</td>
<td>Item 4 has been removed</td>
<td>Item 4 has been removed</td>
<td>Scoring for examples has been removed, description of scoring has changed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Examples were rephrased, combined, moved or removed</td>
<td>Examples were rephrased, combined, moved or removed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Provide feedback that moves the student forward</td>
<td>Item 6 was rephrased</td>
<td>Item 6 was rephrased</td>
<td>Scoring for examples has been removed, description of scoring has changed</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item 8 was moved to strategy 4</td>
<td>Item 8 was moved to strategy 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Examples were replaced or removed</td>
<td>Examples were replaced or removed</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Activating students as owners of their own learning

- Item 8 was added
- Examples were replaced
- Scoring for examples has been removed, description of scoring has changed

5. Activating students as instructional resources for one another

- Examples were removed
- Scoring for examples has been removed, description of scoring has changed

Lay out of the instrument

- Font size of examples changed (smaller)

Because of the decision to remove scoring for the examples, only the items will be scored in the new instrument. Therefore it has become even more important to fine-tune the items as far as possible at this point. Specific actions taken in this fine-tuning are:

- **Formulation of items:**
  - Rephrasing item 1 and 2 (was: *the teacher makes clear what the learning intentions are before the start of the lesson*, is: *the teacher makes clear what the learning intentions are during the lesson*). This way, it is clearer the learning intentions apply to the entire lesson.
  - Adding two items (and fitting examples) in the first category: *The students show they understand the learning intentions by working with them* and *the students show they understand the success criteria by adjusting their work to them*. These items were added after a suggestion was made that the student- and peer point of view was missing in this category. With these items added, it is aimed to get this category more complete.
  - Removing item 4, *the teacher constantly keeps track of the progress of the students*. Half of the respondents did not see the item fit where it was placed, but did not know where it would fit. ‘The other items within this indicator are specific enough to capture all’ and ‘this item is too much like item 7’ were comments made. The examples for this item were replaced or removed, based on the outcomes of the card sort exercise.
  - Rephrasing item 6 (was: *the teacher gives the student feedback on what has been learned*, is: *the teacher gives the student feedback on what is learned*). This change of tense makes it clearer that the feedback applies to the current lesson and not so much lessons in the future.
  - Replacing item 8, *the teacher asks the students for feedback on his/her teaching*. This item was part of the indicator *provide feedback that moves the student forward*, but has been moved to the fourth indicator *activating students as owners of their own learning*. More than half of the respondents did not think this item was categorized in the correct category and suggested to move it to the fourth indicator, arguing that giving feedback on the teacher is a big part of owning your own learning.

With these changes made, the items and indicators in this instrument meet all the requirements in the checklist. In order to make the complete instrument meet the requirements, the examples needed some changes too. Most changes consisted of rephrasing, replacing or removing items, based on the outcomes of the card sort exercise. The exact changes for each example are to be found in appendix IV.

**Instrument: version II**

The instrument as it is now is spread over three pages and consists of five indicators (similar to the strategies of Leahy et al. (2005) and Wiliam (2011)), 12 items and 61 examples.
On top of the first page, general information like observation date, school, observed class, lesson and teacher, number of students and the observer’s name has to be noted. Afterwards, a quick reminder of how the scoring works is written down; items are scored on a four-point-scale (1= predominantly weak, 2= more weak than strong, 3= more strong than weak, 4= predominantly strong). Examples are not being scored.

For the first indicator (clarifying learning intentions and sharing success criteria) there are five items with respectively six, four, five, four and three examples, for example the item **the students demonstrate they understand the learning intentions by working on them** with the example **the students wrote down the learning intentions**. For indicator two (engineer effective classroom discussion and tasks that elicit evidence of learning) there are two items with respectively four and seven examples, for example the item **the teacher uses different strategies for effective questioning** with the example **the teacher applies an adequate waiting time with regard to the answering of students**. Indicator three (provide feedback that moves the student forward) consists of two items with respectively seven and six examples, for example the item **the teacher uses the knowledge about student's progress to give feedback** with the example **the teacher gives students feedback on their learning process**. The fourth indicator (activating students as owners of their own learning) consists of two items with respectively seven and three examples, for example the item **the teacher asks students for feedback on his/her teaching** with the examples **the teacher asks if his/her instruction is clear**. The fifth indicator (activating students as instructional resources for one another) consists of one item with five examples, the item being **the teacher stimulates the students to assess the work of peers and give feedback to peers** with one of the examples being **the teacher stimulates students to ask questions to peers before asking the teacher**. The instrument itself can be found in appendix V.

### 3.5 Phase 3: Classroom observations

In phase three, the instrument was tested in the classroom for the first time. This was done by two observers (researcher included) who observed three lessons (the subjects Dutch, English and mathematics) in one school for secondary education. Schools for secondary education offer different subjects in various levels, which is an opportunity to see if the instrument works for different subjects and in various levels. Before starting these observations, the observers both got a clear explanation on AfL, the instrument itself and how to use it. Unfortunately, it was not possible to practice with the instrument in a real life setting, so the first observation was the first time both observers used the instrument. This led to minor misunderstandings, which will be elaborated in the results section of this phase. During observations, the observers were not allowed to deliberate their findings. After this first observation, the observers discussed their findings and the instrument has been revised based on those findings (other formulation of items, remove/add items based on what was seen in the classroom). This process was repeated twice, so a total of three rounds of observations were held in the same school with the same teachers.

#### 3.5.1 Method (round 1, 2 and 3)

**Respondents**

For the observations in this phase, the observers observed three teachers during three lessons. The respondents were one mathematics teacher, one English teacher and one Dutch teacher, who all worked in the same school. This school was selected based on a convenience sample: Within the larger group of schools that could have participated (schools for secondary education in the Netherlands) this one was available and had teachers in the three subjects that were willing to participate in this research (Onwuegbuzie & Leech, 2007). Observations were planned to take place in the same class for every round, but due to schedule changes in the school, this was not possible. For the first round, the observation of the mathematics teacher took place in a first grade, for the teachers in English and Dutch the observations took place in a fourth (final) grade. All the observed lessons in all of the rounds were in the lower general secondary education level.

For the second round, the observation for mathematics took place in a third grade and for English in a fourth (final) grade. Because of the changed schedule, the Dutch teacher did not teach that
subject anymore on the days that were available for observation. This led to the choice to observe her other subject, English. In this way, the observed teacher would be the same, which was found most important according to the researcher, because this instrument is about measuring the teachers’ use of AfL and not so much about the subject. At the start of the English lesson (the one the Dutch teacher taught), there appeared to be a problem. The English lesson was replaced with a mentor lesson due to severe problems in the classroom. The teacher proposed to observe this lesson and because there was no other option to observe a lesson Dutch or English, this mentor lesson was observed. This is not a regular lesson and it was taught by two teachers, but since the instrument is developed to observe the teacher, the researcher decided to observe this lesson and focus on the teacher that had been observed in the first round as well.

For the third round of observations, the observed mathematics lesson was the same as in the second round. The lesson English taught by the English teacher, was the same as well. For the Dutch teacher, the class was the same, only they were taught English instead of their mentor lesson in this round.

Procedure

In order to plan the observations, schools needed to be contacted. For this phase, three schools were contacted from which one immediately replied they would be interested in participating in this research. The researcher then went to the school to provide them with additional information so they could contact teachers and ask them if they would be interested in participating in this research. After two weeks, the school sent a planning and observation started one week after the autumn break, this was the last week before the school had their test week. The second and third round of observations took place in the two weeks after the test week, so it took a month to conduct the three rounds of observations.

Before the observers started their observations, they introduced themselves to the teacher and the students so they all knew for what purpose they had visitors. During the lesson, both observers sat quietly in the back of the room, without disturbing the lesson. Afterwards, the observers thanked the teacher and answered the questions, in case the teachers had questions. Outside of the classroom, both observers deliberated their findings and discussed what was seen during observation with regard to changes that could be made in the instrument. After all the observations in one round were done, the researcher improved the instrument based on the comments and the data analysis.

3.5.2 Data analysis (round 1)

Before revising the instrument based on the comments of the observers, the consistency among items within the instrument, measured by Cronbach’s α, was calculated (Dooley, 2001). For an instrument used in real life settings, this number should be at least 0.70 (Dooley, 2001).

In order to analyze the comments made by the observers in this round of testing, the comments have been coded using the codes derived from the checklist of requirements. Codes used were formulation of items, feasibility, scoring, quality, usability and other. The comments have been categorized and per code all comments were read again and a decision to make a change or ignore the comment was made. After changes based on Cronbach’s α and comments of observers have been made, the instrument was once again checked using the checklist of requirements to make sure the instrument still met all the requirements after revision.

3.5.3 Results (round 1)

Comment made by the observers

During observations, both observers have made comments on the observation instrument regarding items, examples or other features of the instrument that caused confusion. These comments were coded using the coding scheme; the results are shown in the table below.

An example of comments made by the observers is the following; for the second indicator engineering effective classroom discussions and tasks that elicit evidence of learning, it was not clear whether the items were only to be used for classical instruction or for individual instruction as well. This led to rewriting these items and/or examples. During observations, the observers noted that items
2 and 4 were not mutual exclusive, which led to removing both items. Also, it was noted that there was an item missing regarding feedback from students. It occurs quite some times that students give feedback and the teacher uses this feedback to adapt instruction, but this was not written down in the instrument.

These comments and the items on which both observers scored different frequently were leading in the changes made to the instrument after the first round of observations. These changes can be found in the table below.

**Revision of the instrument**

The found number for Cronbach’s α for the three observations together was 0.660. This is considered sufficient, but when using the instrument in real life settings, α should be at least 0.70 (Dooley, 2001). In calculating Cronbach’s α, the scores for *scale if item deleted*, which show if the coefficient α increased if a certain item is being removed, were remarkable. For none of the items, the coefficient α would increase if that item was deleted. Based on the score for Cronbach’s α, items could have been removed, but this would not lead to an increased score for Cronbach’s α and therefore no items were removed based on this score. The changes made during revision were based on the comments of the observers and the checklist of requirements.

**Table 3**

<table>
<thead>
<tr>
<th>Changes made in the instrument after the first round of observations</th>
<th>Formulation of items</th>
<th>Feasibility</th>
<th>Scoring</th>
<th>Quality</th>
<th>Usability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clarifying learning intentions and sharing success criteria</td>
<td>Item 1 and 3 rewritten</td>
<td>Item 2 and 4 removed, Item about learning intentions added</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Engineering effective classroom discussions and tasks that elicit evidence of learning</td>
<td>Examples for item 6 changed</td>
<td>Item 7 rewritten</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Provide feedback that moves the student forward</td>
<td>Item 8: one example changed</td>
<td>Item 10 added</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Activating students as owners of their own learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Activating students as instructional resources for one another</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The table above shows what actions have been taken in revising the instrument after the first round of observations. Specific actions that were taken in this revision are:

- **Formulation of items**:  
  - Rephrasing item 1 and 3: because item 2 and 4 are being removed (see next bullet), items 1 and 3 needed to be rephrased in order to make these items cover the actors of
peers and students as well. The items are rewritten to *the teacher makes clear what the learning intentions are during the lesson in a way students understand them* and *the teacher makes clear what the success criteria are during the lesson in a way students understand them*. To make this clearer, two examples were added for each item: *the students demonstrate they understand the learning intentions by working on them* and *the students demonstrate they understand the success criteria by adjusting their work to them.*

- Removing item 2 and 4: during observations the researcher noticed items 2 and 4 were not mutual exclusive and thus needed to be adjusted. Due to the rephrasing of items 1 and 3, these items were chosen to be removed.
- Adding item *learning intentions differ per student*: during observations the item *success criteria differ per student* was rarely seen, but teachers did however differentiate in learning intentions for students, so this item was added to cover that part.
- Item 6 was the only item never scored the same by the two observers during the first round. Discussing the observations, it turned out one of the observers thought this item only applied to classroom settings and not to individual instruction, while the other observer applied it to individual instruction as well. To solve this problem, there were examples added about individual instruction: *the teacher asks questions to individual students to find out what their prior knowledge is* and *the teachers gives the students assignments to find out what their prior knowledge is.*
- Though item 7 was never scored differently during this round of observations, there is some information added to this item during revision, so there is no difference between the two items in one indicator. The item *the teacher uses different strategies to question the students effectively* has now the words (classical or individual) added.
- For item 8, one example has changed. In the example *the written feedback focuses on the content of the task*, the word *written* was removed. Observation showed that feedback can also be spoken instead of written when it is about the content of the task.
- Item 10 was added. During observations students gave feedback to the teachers and in most cases the teachers used this feedback immediately to adapt instruction. However, the instrument had no item to measure this, so the item *the teacher used students’ feedback on instruction immediately to adapt instruction* was added.

- Quality:
  - In terms of quality, there is one change in the general information in the beginning of the instrument. During observations the observers found out that the time of the lesson and the kind of lesson were of great influence on the scores in the instrument. These two factors were therefore added to the general information section.

- Usability:
  - In terms of usability, both observers felt the need to write down additional notes about what was being observed, for example things the teacher could do better in a certain situation or things a teacher did that deserved to be noticed, but were not mentioned in the items or examples. Therefore, an *additional notes* section was added at the end of the instrument.

With all these changes made, the instrument is revised and ready for a second round of testing.

**Instrument: version III**

The instrument as it is now is spread over three pages and consists of five indicators (similar to the strategies of Leahy et al. (2005) and Wiliam (2011)), 12 items and 64 examples. On top of the first page, general information has to be noted. Afterwards, a quick reminder of how scoring works is written down.

For the first indicator (clarifying learning intentions and sharing success criteria) there are four items with respectively seven, three, six and three examples, for example the item *learning intentions differ per student* with the example *the teacher provides some of the students with extra work.* For indicator two (engineer effective classroom discussion and tasks that elicit evidence of learning) there
are two items with respectively six and seven examples, for example the item *the teacher uses different methods to find out what the prior knowledge of the students is* with the example *the teacher uses brainstorm techniques*. Indicator three (provide feedback that moves the student forward) consists of three items with respectively seven, six and four examples, for example the item *the teacher uses students' feedback to adjust instruction* with the example *when the students say instruction is not clear, the teacher adjusts the instruction*. The fourth indicator (activating students as owners of their own learning) consists of two items with respectively seven and three examples, for example the item *the teacher asks students for feedback on his/her teaching* with the examples *the teacher asks if students are being challenged during the lesson*. The fifth indicator (activating students as instructional resources for one another) consists of one item with five examples, the item being *the teacher stimulates the students to assess the work of peers and give feedback to peers* with one of the examples being *the teacher stimulates working in groups if an assignment is too hard to complete individually*. At the bottom of the last page, there is room for additional notes. The instrument itself can be found in appendix VI.

3.5.4 Data analysis (round 2)
The revised instrument was tested in the second round of observations in the same school with the same three teachers. Before revising the instrument based on the comments of the observers, the consistency among items within the instrument, measured by Cronbach’s $\alpha$, was calculated again (Dooley, 2001). For an instrument used in real life settings, this number should be at least 0.70 (Dooley, 2001).

In order to analyze the comments made by the observers in this round of testing, the comments have been coded using the codes derived from the checklist of requirements. Codes used were formulation of items, feasibility, scoring, quality, usability and other. The comments have been categorized and per code all comments were read again and a decision to make a change or ignore the comment has been made. After changes based on Cronbach’s $\alpha$ and comments of observers have been made, the instrument was once again checked using the checklist of requirements to make sure the instrument still met all the requirements after revision.

3.5.5 Results (round 2)

*Comments made by observers*
During observations both observers have made comments on the observation instrument regarding items, examples or other features of the instrument that caused confusion.

One of the comments made was that during the observed lessons, teachers did not pay attention to learning intentions, but they did mention the goals for the coming lesson(s). Since the first strategy of Leahy et al. (2005) and Wiliam (2011) clearly is about learning intention and not lesson goals, this was not changed in the instrument. The difference between learning intentions and lesson goals will be elaborated in the general pointers for using the instrument that will be added. Another comment made within this first strategy is about the learning intentions and success criteria being fit for just one lesson or for long term as well. Especially in the final grade, teachers are talking about learning intentions and success criteria towards the exams and less about them within this one lesson.

The last comment made by the observers was about the new item 2 and 4. During observations the observers discovered both items were not mutual exclusive, so they have to be changed during revision. All the changes made are shown in the table and explained below.

*Revision of the instrument*
The found number for Cronbach’s $\alpha$ was 0.207, which is too low to be considered sufficient (Dooley, 2001) and lower than the score for $\alpha$ in the previous round. This might had to do with the fact that items were added based on what was seen in practice and not based on literature. These added items did measure what was going on in the classroom, but did not fit the strategies for AfL. In calculating Cronbach’s $\alpha$, the scores for *scale if item deleted*, which show if the coefficient $\alpha$ increased if a certain item is being removed, showed that Cronbach’s $\alpha$ would be 0.609 when item 9 was removed from the instrument. Item 9 is one of the items added based on what was seen in practice and not based on
literature. This and the increasing score for Cronbach’s α led to the decision to remove item 9. Other changes were made based on the comments of the observers and are described in the table below.

### Table 4
Changes made in the instrument after the second round of observations

<table>
<thead>
<tr>
<th>Formulation of items</th>
<th>Feasibility</th>
<th>Scoring</th>
<th>Quality</th>
<th>Usability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clarifying learning intentions and sharing success criteria</td>
<td>Item 1 and 3 rewritten and examples changed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Item 2 and 4 replaced</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Engineering effective classroom discussions and tasks that elicit evidence of learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Provide feedback that moves the student forward</td>
<td>Item 9 removed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Activating students as owners of their own learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Activating students as instructional resources for one another</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lay out of the instrument</td>
<td>Adding a document with general pointer for the use of the instrument</td>
<td>Adding a document with general pointer for the use of the instrument</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As shown in the table above, there are some changes made in the instrument after the second round of observations:

- **Formulation of items:**
  - Item 1 and 3 were rewritten and examples were added. Because teachers do not only think about learning intentions and success criteria for one lesson, but also for the rest of the chapter or in some cases the rest of the year, these items were rewritten. The words *during the lesson* were removed, which lead to the new items *the teacher makes clear what the learning intention are in a way the students understand them* and *the teacher makes clear what the success criteria are in a way the students understand them*. In order to make sure observers know this is about both the lesson and long term, examples like *the teacher makes clear what the learning intentions for this lesson are* and *the teacher makes clear what the long term success criteria are* were added.
  - Item 2 and 4 were found not to be mutual exclusive; in order to score these items, items 1 and 3 always needed to be present. This led to changing these items into the following ones: *the teacher adjusts teaching based on the differences in pace of the students* and *the teacher adjusts teaching based on the differences in level of the students*. For item 2, there were examples added like *some student get less work* or *some student can go earlier after all their work is done*. For item 4, the example *some students get extra work* is deleted.
Item 9 was removed based on the score for Cronbach’s $\alpha$. The scale if item deleted calculation showed that the coefficient $\alpha$ would be increased from 0.207 to 0.609 if item 9 would be removed. Since this item was added based on what was seen during observations and was not added based on the literature study, the decision was made to remove this item.

Quality and usability:
- Even during the second round of observations, there were still misunderstanding regarding the instrument. For the next round, a document will be added that consists of general pointers for the observers, like all items can be scored in the classical situation as well as for one-on-one (individual) instruction. With all these changes made, the instrument is revised and ready to use in the third round of testing, in which the same teachers will be observed again.

Instrument: version IV

The instrument as it is now is spread over four pages and consists of five indicators (similar to the strategies of Leahy et al. (2005) and Wiliam (2011)), 11 items and 65 examples. On the first page, all important information the observer needs to know is written down.

On top of the second page, general information has to be noted. Afterwards, a quick reminder of how scoring works is written down.

For the first indicator (clarifying learning intentions and sharing success criteria) there are four items with respectively nine, four, eight and three examples, for example the item the teacher adjusts teaching based on the differences in level of the students with the example the teacher provides some of the students with easier work. For indicator two (engineer effective classroom discussion and tasks that elicit evidence of learning) there are two items with respectively six and seven examples, for example the item the teacher uses different strategies of effective questioning with the example the teacher asks the students open questions. Indicator three (provide feedback that moves the student forward) consists of two items with respectively seven and six and for examples, for example the item the teacher gives feedback on what needs to be learned with the example the feedback is focused on the content of the task. The fourth indicator (activating students as owners of their own learning) consists of two items with respectively seven and three examples, for example the item the teacher stimulates students to monitor their own progress with the examples students are able to explain what they are learning at this moment. The fifth indicator (activating students as instructional resources for one another) consists of one item with five examples, the item being the teacher stimulates the students to assess the work of peers and give feedback to peers with one of the examples being the teacher uses anonymised examples to show the students frequently made mistakes. At the bottom of the last page, there is room for additional notes. The instrument itself can be found in appendix VII.

3.5.6 Data analysis (round 3)
The revised instrument was tested in the third round of observations in the same school with the same three teachers. Before revising the instrument based on the comments of the observers, the consistency among items within the instrument, measured by Cronbach’s $\alpha$, was calculated again (Dooley, 2001). For an instrument used in real life settings, this number should be at least 0.70 (Dooley, 2001).

In order to analyze the comments made by the observers in this round of testing, the comments have been coded using the codes derived from the checklist of requirements. Codes used were formulation of items, feasibility, scoring, quality, usability and other. The comments have been categorized and per code all comments were read again and a decision to make a change or ignore the comment was made. After changes based on Cronbach’s $\alpha$ and comments of observers have been made, the instrument was once again checked using the checklist of requirements to make sure the instrument still met all the requirements after revision.
3.5.7 Results (round 3)

Comments made by observers
During observations both observers have made comments on the observation instrument regarding items, examples or other features of the instrument that caused confusion.

For this last round of observations in this phase of the research, there were only a few comments made. Because the document with general pointers was added, it was no longer necessary to add comments like classical or individual (item 6) in the instrument itself, since this was mentioned in the general pointers.

Other comments that were made were about the examples, like adding one more example and rewriting another example to create more clarity.

Revision of the instrument
The found number for Cronbach’s α for all observed lessons was 0.573. Looking at the scores for scale if item deleted, it shows that Cronbach’s α increases to 0.769 if item 2 was being removed. Removing item 4 would lead Cronbach’s α to increase to 0.713. These scores let to the decision to calculate Cronbach’s α again, but without both items 2 and 4, this led to a score for Cronbach’s α of 0.819, which can be considered usable. Because both items were in the instrument based on what was seen in the classroom during previous observations and not because of the results of the literature study, both items were removed. Next to removing these items, minor changes have been made that are described in the table below.

<table>
<thead>
<tr>
<th>Formulation of items</th>
<th>Feasibility</th>
<th>Scoring</th>
<th>Quality</th>
<th>Usability</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Clarifying learning intentions and sharing success criteria</td>
<td>Item 2 and 4: removed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Engineering effective classroom discussions and tasks that elicit evidence of learning</td>
<td>Item 6: rewritten</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Provide feedback that moves the student forward</td>
<td>Item 8: example rewritten</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Activating students as owners of their own learning</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Activating students as instructional resources for one another</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

As shown in the table above, there are minor changes made in the instrument after the third round of observations:
- Formulation of items:
  - Item 2 and 4: removed because of the increasing score for Cronbach’s α when both items are being removed.
  - Item 6: the addition (classical or individual) was removed, because this is now covered in the general pointers that are added after the second round of observations.
  - Item 8: the example the teachers gives feedback based on tests the students have made, was rewritten into the teacher gives feedback based on assignments/tests the students have made.

With these changes made, the instrument is revised for the fourth time. Apart from the removal of items 2 and 4, no major changes have been made.
**Instrument: version V, final version**
The instrument as it is now is spread over four pages and consists of five indicators (similar to the strategies of Leahy et al. (2005) and Wiliam (2011)), 9 items and 58 examples. On the first page, all important information the observer needs to know is written down. On top of the second page, general information has to be noted. Afterwards, a quick reminder of how scoring works is written down.

For the first indicator (clarifying learning intentions and sharing success criteria) there are two items with respectively nine and eight examples, for example the item *the teacher makes clear what the success criteria are in a way students understand* with the example *the teacher makes clear what the long term success criteria are*. For indicator two (engineer effective classroom discussion and tasks that elicit evidence of learning) there are two items with respectively six and seven examples, for example the item *the teacher uses different methods to find out what the prior knowledge of the students is* with the example *the teacher asks individual questions to find out what the prior knowledge is*. Indicator three (provide feedback that moves the student forward) consists of two items with respectively seven and six examples, for example the item *the teacher gives feedback on what needs to be learned* with the example *the feedback describes specifics points that need improvement*. The fourth indicator (activating students as owners of their own learning) consists of two items with respectively seven and three examples, for example the item *the teacher asks students for feedback on his/her teaching* with the examples *the teacher asks students what can be done differently with regard to instruction*. The fifth indicator (activating students as instructional resources for one another) consists of one item with five examples, the item being *the teacher stimulates the students to assess the work of peers and give feedback to peers* with one of the examples being *the teacher stimulates peers to ask each other questions during the lesson*. At the bottom of the last page, there is room for additional notes. The instrument itself can be found in appendix VII.

### 3.6 Phase 4: Classroom observations

Phase four consisted of testing the instrument in another school. Because testing the instrument in just one school could result in being lucky with what has been seen, the instrument was tested in one other school, just one lesson per subject (Dutch, English and mathematics). During this observation the observers were again not allowed to deliberate their findings. The inter-rater reliability was determined by calculating Cohen’s Kappa. To determine if more items could form one scale, Cronbach’s $\alpha$ (item-scale relation) was calculated.

#### 3.6.1 Method

**Respondents**

For the observations in the final phase the observers observed three teachers, each during one lesson. The respondents were one mathematics teacher, one English teacher and one Dutch teacher, who all worked in the same school. This school was selected based on a convenience sample: within the larger group of schools that could have participated (schools for secondary education in the Netherlands) this one was available and had teachers in the three subjects that were willing to participate in this research (Onwuegbuzie & Leech, 2007).

Observations of the mathematics teacher took place in a sixth (final) grade (atheneum), for the teachers in English observation took place in a fourth grade (atheneum) and for the teacher in Dutch the observation took place in a first grade (higher general secondary education and atheneum combined).

**Procedure**

In order to plan the last round of observations, schools needed to be contacted again. For this phase, five schools were contacted from which two replied they would be interested in participating in this research. In one of these schools, all teachers that would participate were available on the same day so for practical reasons, this school was chosen to participate in the final round of observations. This final round of observations took place in the same week as the third round of observations took place; the third round observation took place on Monday and the final round observations took place on Friday.
Before the observers started their observations, they introduced themselves to the teacher and the students so they all knew for what purpose they had visitors. During the lesson, both observers sat quietly in the back of the room, without disturbing the lesson. Afterwards, the observers thanked the teacher and answered the questions, in case the teachers had questions. Outside of the classroom, both observers deliberated their findings and discussed what was seen during observation with regard to changes that could be made in the instrument. After all the observations were done, the researcher completed the data analysis and made changes if needed.

**Data analysis**

After the fourth round of testing, the inter-rater reliability (Cohen’s Kappa) and the item-scale relation (Cronbach’s α) have been calculated. Cronbach’s α has been calculated in phase three as well, but needs to be calculated in this phase to make sure the instrument as it is now is internal consistent, based on the three observations in this round. For this last round of observations, the inter-rater reliability, Cohen’s Kappa, was calculated as well. Cohen’s Kappa needs to be between 0.61 and 0.80 to be considered *substantial* and above 0.81 to be considered *nearly perfect* (Landis & Koch, 1977 in: Stemler, 2001).

**3.6.2 Results**

The score for Cronbach’s α found in this round is 0.731 and was based on the finding of the researcher only. This score indicated the instrument can be considered usable (Dooley, 2001). Looking at the *scale if item deleted* scores, it shows that Cronbach’s α would only be significantly higher (α=0.870) if item 2 would be removed. However, it is chosen to not remove this item, based on the literature study.

Item 2, *The teacher makes clear what the success criteria are in a way students understand them*, is the only item that measures the use of success criteria, which is one of the strategies of AfL (Leahy et al., 2005; Wiliam, 2011). The score for Cronbach’s α is still high enough when this item is not being removed. The inconsistency in this item might be due to the fact that for all observed lessons, this item was scored differently; one teacher did not use it at all, one just a little bit and one used it almost to its full extent. This could have caused the differences, which could lead to a lower score for Cronbach’s α.

For this last round of observations, the inter-rater reliability, Cohen’s Kappa, was calculated as well. The score found for Cohen’s Kappa was 0.851, which can be considered *nearly perfect*. This indicated that both observers agree enough to consider this instrument reliable with regard to the observer agreement for the circumstances it is tested under (three lessons with two observers).

Furthermore, there were no additional comments made by the observers during this round of observations, which means there are no changes made in the instrument after this round. The final instrument, the one that was delivered after the fourth revision, can be found in appendix VIII.
4. Conclusion and discussion

4.1 Conclusion

In this research, the question *What are the characteristics of an observation instrument to measure Assessment for Learning in the classroom?* has been answered by conducting a literature study, two focus group interviews, and four rounds of observations to test the instrument that was developed. The two characteristics this observation instrument needs to have are meeting all the requirements in the checklist and being built upon the strategies of Leahy et al. (2005) and Wiliam (2011).

The first characteristics of an observation instrument to measure AfL in the classroom is that it has to meet all the requirements in the checklist of requirements. These 21 requirements are spread over the five categories *formulation of items* (e.g. items have to be mutual exclusive, items cannot be multi interpretable and items need to be observable), *feasibility* (researchers have to be able to observe the items in the time given for the observation), *scoring* (e.g. the length of the scale has to be optimal for the propose of the instrument, the explanation on the scoring must be formulated in language that can be easily understood and there needs to be a clear distinction in scoring options), *quality* (e.g. the instrument must be considered valid, the instrument must be considered reliable and the instrument needs to provide an objective view to assess the teacher) and *usability* (e.g. the instrument must allow to compare teachers across classrooms and recording must take place in a feasible and non-intrusive way using a paper-and-pencil format). Within this instrument, all requirements are met except for the one about validity. Due to too little items in the instrument, the instrument cannot be considered valid to measure the use of AfL in the classroom. In order to make the instrument valid to measure all the strategies of AfL individually and as a whole, more items have to be added.

The second characteristic is that the instrument needs to be built on the strategies for AfL by Leahy et al. (2005) and Wiliam (2011). These strategies are clarifying learning intentions and sharing success criteria, engineer effective classroom discussion and tasks that elicit evidence of learning, provide feedback that moves the student forward, activating students as owners of their own learning and activating students as instructional resources for one another. These strategies are leading in the implementation of AfL in the classroom and therefore are one of the main characteristics in an observation instrument to measure AfL in the classroom.

After conducting a literature study the first draft of the instrument was developed, based on the instrument of Oswalt (2013), the ICALT-instrument (Van der Grift, 2007; Van der Grift & Van der Wal, 2012) and the questionnaires on AfL (Kippers et al., submitted; Wolterinck et al., 2016). This first draft was then evaluated in two focus group interviews, one with educational researchers as participants and one with secondary education teachers. This revealed that the instrument was too long: scoring the items and examples would be too much. However, the respondents did like the practical implications of the examples. After these focus group interviews, the number of examples was decreased and the scoring of the examples was removed, so the instrument did not have too many items/examples that needed to be scored. The examples were from now on only used as help, not to be scored.

During the third and fourth phase (the four rounds of observations), several changes have been made. Some items caused misunderstandings, some items were not mutual exclusive, or some items were removed in order to increase the score for Cronbach’s α. With a total of four revisions (the first after the focus group interviews and then after all rounds of observations in phase three), the final instrument was ready for a last round of observations in phase four.

This all together led to the final version of the instrument in which all important information the observer needs to know is written down on the first page of the instrument, followed by a list of factors that can influence the observation that needs to be filled in. For the first indicator (clarifying learning intentions and sharing success criteria) there are four items with respectively nine, four, eight and three examples, for example the item *the teacher adjusts teaching based on the differences in level of the students* with the example *the teacher provides some of the students with easier work.* For indicator two (engineer effective classroom discussion and tasks that elicit evidence of learning) there are two items with respectively six and seven examples, for example the item *the teacher uses
different strategies of effective questioning with the example the teacher asks the students open questions. Indicator three (provide feedback that moves the student forward) consists of two items with respectively seven and six examples, for example the item the teacher gives feedback on what needs to be learned with the example the feedback is focused on the content of the task. The fourth indicator (activating students as owners of their own learning) consists of two items with respectively seven and three examples, for example the item the teacher stimulates students to monitor their own progress with the example students are able to explain what they are learning at this moment. The fifth indicator (activating students as instructional resources for one another) consists of one item with five examples, the item being the teacher stimulates the students to assess the work of peers and give feedback to peers with one of the examples being the teacher uses anonymised examples to show the students frequently made mistakes. At the bottom of the last page, there is room for additional notes. The instrument itself can be found in appendix VIII.

4.2 Discussion

Limitations of the study

The research and development of the observation instrument to measure AfL in the classroom took about 10 months. Within this time, a literature study which formed the base for the development of the instrument, and the further research using three methods (screening, focus group interviews, and observations) were conducted. However, this was not enough time to test the instrument extensively. Therefore, the instrument has only been tested in four rounds (12 lessons) in total. During these rounds of observations, the instrument has been revised three times, which led to only three observations that could have been used to calculate inter-rater reliability (Cohen’s Kappa) and final score for consistency among items (Cronbach’s α). Based on the scores for these three observations, the instrument is considered reliable (Cohen’s Kappa=0.851) and internal consistent (Cronbach’s α=0.731), but these scores could differ when measured with a larger sample of observations. Also, the sampling method, a convenience sample, might have affected these results. Using for example a broader variety in teachers, level of students or school type could have resulted in different results.

As mentioned in the conclusion, the developed instrument consists of too little items to be considered valid to measure the use of AfL in the classroom by individual strategies or as a whole. This can be due to the fact that the instrument is limited to the strategies of Leahy et al. (2005) and Wiliam (2011) that were found to be used most in the researched literature. In order to make this instrument valid, more items need to be added within the strategies of Leahy et al. (2005) and Wiliam (2011). Another possibility could be adding other strategies such as the ones that Heritage (2007) (elicit evidence about student learning, provide teachers and students with feedback about learning, use feedback to adjust instruction and learning tactics in real time and involve students actively in their learning) and Ruiz-Primo and Furtak (2006) (the teacher elicits a question, the student responds, the teacher recognizes the student’s response and then uses the collected information to support student learning) suggest in their research. This can lead to a broader definition of AfL in which more items can be incorporated, which may lead to a more valid instrument.

Looking at the reliability of the instrument, the score for Cohen’s Kappa can be considered nearly perfect, but since all the observations were done by the same two people, the inter-rater reliability shows that those two people agree, but it does not show how this would work for others. A more generalizable score of the inter-rater reliability could be retrieved by testing the instrument with more observers in a larger sample.

Another limitation of the study is that due to scheduling problems the intended subjects mathematics, English and Dutch could not be observed in all rounds of observations. The first round took place in all subjects, but the second and third round took place in the subjects English and mathematics. In the second round, there was an additional problem, which caused one of the English lessons to be replaced with a mentoring lesson. Although these problems occurred, the observed lessons were all taught by the same teacher. On the bright side, using the instrument in a mentor lesson was a unique opportunity to see if the instrument would work in such a lesson as well and it did. Because this instrument is developed to measure the use of AfL by the teacher, this factor was most
important to keep constant. Nevertheless, results might be different when other lesson would have been observed.

Furthermore, the fact that the instrument and all its additional information were written in Dutch is a limitation. This lowers the generalizability of the instrument; the extent to which it can be used in other situations. Also, the focus group interviews and observations were in Dutch. The language and formulation problems that occurred were solved, but specifically in the Dutch language. When translating this instrument to English (or another language), new problems in language or formulation might occur, which can influence the reliability and validity of the instrument. The instrument should be tested again, which is advisable to do using the checklist of requirements that has been used in this research. This checklist provides requirements the instrument has to meet in order to be considered a good instrument to measure AfL in the classroom.

**Implications for future research**

The first implication for future research is to develop this instrument until it is considered valid by adding more items. Examples of items that could be added for indicator 1 might be about the use of learning intentions/success criteria by the students, differentiating in learning intentions/success criteria, or the role of peers in clarifying learning intentions and sharing success criteria. For indicator 4, some examples may be fitted to rewrite as an item, for example the teacher stimulates students to assess their own work or the teacher stimulates the students to develop skills in self-monitoring. For indicator 5, items can be about the role of students and peers, and the item in this instrument can be split into one about peer-assessment and one about peer-feedback. Changing the instrument based on these findings would lead to a new instrument that needs to be tested again.

Furthermore, future research might be focused on testing the instrument in a larger sample with more than two observers. Also, it would be interesting to test the instrument is primary education or higher education, since AfL is useful to implement in any kind of education. It might be interesting to test the instrument with a broader variety of subjects, types of lessons, teachers and student levels as well. Observations have shown that the instrument is fitted for different subjects (mathematics, English, Dutch and a mentor lesson) and for different levels of students (from first year lower secondary education to final year athenaeum). It was notable to see that for students further in their education, other strategies were used that for first year students. Furthermore, differences in the use of AfL were more visible in different teachers than different subjects. But this might be different when more subjects are being observed with more different teachers. During these observations, there are not that many different types of lessons observed and the variety in age of teachers was minimal (four under 35, one around 40 and one above 50 years). It would be interesting to see if in different types of lessons or different ages of teachers the use of AfL would differ.

Another implication is that future research might be focused on researching this instrument in other countries, to solve translation problems that could occur when translating the instrument. Translating the instrument to English could be a first step, but in order to use it in non-Dutch or non-English countries, it is advised to translate the instrument to the native language of the country it will be used in, to prevent language barriers during the use of the instrument. This is time consuming, but it is expected that results will be more reliable when the instrument is written in the mother tongue of the user. Translated versions needs to be researched in order to be used and considered reliable and valid.

The observations that were done during this research did not include reactions of the observed teachers. It would be worthwhile to see what the reactions of teachers would be and to what extent they would improve their teaching using this instrument. This instrument can be a good starting point in improving instruction and can be functioning as a feedback tool, but it is important that teachers are willing to improve, otherwise it would not work. Also, research in which teachers are using the instrument to measure the use of AfL of their colleagues could be interesting in the future. That way, schools can use the instrument on their own and there is no need to hire an external observer to observe the teachers.

The instrument developed during this research focuses on the role of the teacher in using AfL in the classroom. It would be interesting for future research to focus on the role of students in using AfL in the classroom, because they do have a large role in this as well. This could for example be done
by developing an instrument focusing on the role of the students instead of the role of the teacher or interviewing students about their role in AfL in the classroom.

The last implication for future research is not on the development of this instrument, but on the use of it. After developing it to the extent it can be considered reliable and valid in a larger sample and internal consistent for all indicators separately, it would be interesting to use this instrument to study teachers’ use of AfL. When doing this, it is important to work with observers who have gained knowledge on AfL; it can be an extensive concept to learn when you have no experience in or knowledge about education. Also, it is important to observe the teachers more than once in the same lesson/class, to prevent missing out on important things. All strategies of AfL are sometimes not shown in one lesson, but can be detected in a series of lessons.

**Implications for practice**

Although the instrument as it is developed during this study needs further research, there are already some practical implications that can be given.

During observations it was noted that teachers are curious about their score on the instrument. They wanted to know how they performed by the means of the instrument. The fact that teachers are eager to know this can be helpful in improving their teaching, based on the findings of the instrument. In practice, this instrument can be used as a starting point to determine where teachers stand in their use of AfL and as a feedback tool for what and how they can improve it. The teachers that participated in the focus group interviews thought the examples were very useful when the instrument would be used to give teachers feedback on their teaching regarding the use of AfL. Because this is not a normative instrument, it is suitable for this application. However it needs to be noted that when using the instrument for this purpose, it is advisable to observe more than one lesson per teacher. During observations was seen that teachers use more or less the same strategies in their lessons, no matter what grade, type of lesson or subject was taught. Observing more lessons leads to a more complete view of the teacher’s use of AfL and so feedback can be better tuned to this.

Furthermore, there seems to be an important role for teacher training colleges in the implementation of the use of AfL in schools. During observations was seen that the younger the teacher, the more accurate the strategies of AfL were implemented in the classroom. This raises the assumption that teacher training colleges have improved their education when it comes to AfL, but it is important that they keep doing it. Using an observation instrument to measure the use of AfL in teachers in training during their internship might make this concept more comprehensible and gives teachers in training more practical implications in how to implement AfL in their classrooms. When teacher training colleges educate their students on the use of AfL, a big step in implementing AfL in the classroom can be made.
5. References


Kippers, W. B., Schildkamp, K., Poortman, C. L., & Visscher, A. J. (submitted). Teachers’ views on the use of Assessment for Learning and Data-Based Decision Making in classroom practice.


6. Appendices

6.1 Appendix I: Screening instruments (phase 1)

The tables below show the results of the screening in phase 1. Table 4 shows the results of Oswalt’s instrument (2013) and table 5 shows the results for the questionnaires on AfL by Lysaght & O’Leary (2013)

**Table 1**

*Results of screening of Oswalt’s instrument (2013)*

<table>
<thead>
<tr>
<th>Strategy</th>
<th>Item</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning targets: Clarifying learning intentions and sharing criteria for success</td>
<td>Does the teacher make certain that students understand the learning intentions for the class session?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the teacher make certain that students understand the learning intentions for each activity?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the teacher provide examples of high and low quality work?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the teacher address potential misunderstandings regarding the criteria for success?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the teacher make efforts to monitor student learning on an ongoing basis (i.e., minute-to-minute &amp; day-to-day)?</td>
<td>3a</td>
</tr>
<tr>
<td>Monitoring: Engineering effective classroom discussions, questions and learning tasks that elicit evidence of learning</td>
<td>Does the teacher give students a variety of opportunities and methods (e.g., verbal, written, electronic, &amp; visual) to respond to questions?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the teacher use effective questioning strategies (e.g., adequate wait time, open-ended questions) to elicit evidence of learning?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does teacher monitoring seek to elicit evidence from students of both factual/procedural knowledge and of deeper conceptual knowledge?</td>
<td>1a</td>
</tr>
<tr>
<td></td>
<td>Does teacher monitoring seek to elicit evidence of whether students can transfer knowledge within and between disciplines/subjects?</td>
<td>3a</td>
</tr>
<tr>
<td></td>
<td>Does the teacher provide meaningful feedback (i.e., information with which a learner can confirm, add to, overwrite, tune, or restructure understanding) immediately following formal and/or informal evaluations of student progress?</td>
<td>1a</td>
</tr>
<tr>
<td>Feedback: providing feedback that moves learners forward</td>
<td>Does the teacher provide accurate feedback that assists learning?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the teacher provide feedback in reference to a criterion based standard, avoiding feedback based in comparison to other students?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does feedback describe specific areas of needed improvement and suggest alternative strategies for making that improvement?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does feedback describe specific student strengths and suggest strategies for continued learning in those areas?</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Does the teacher give students opportunities to use self regulatory competencies, such as the ability to accurately assess their own knowledge?</td>
<td></td>
</tr>
<tr>
<td>Self-assessment: activating</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Master thesis: What do we learn from assessment?

Note: all the items from this instrument have been rephrased into sentences stating ‘the teacher.’, instead of questions.

Scoring: scoring is too limited, no explanation was written down on the instrument (3b, 3c, 3e). The length might be optimal for the purpose of this instrument (3a) and when using this instrument, respondents are scored unambiguously (3d), but all in all, this scoring needs to be improved in order use it in the new instrument.

Quality and usability: The instrument does provide an objective view to assess the teacher, but this view might be too limited (4a, 4b). It does allow to compare teachers across classrooms (5a). Validity and reliability are sufficient, according to the research (4c, 4d), though this does not necessarily translate to the new instrument. This instrument is designed in a way the results can be interpreted directly (5b) and it is a paper-and-pencil format (5c). However, the factors that can influence the observation are not noted on the form (4e).

Table 2

Results of screening of the questionnaire on AfL (in Dutch) (Kippers et al., submitted; Wolterinck et al., 2016)
Er worden verschillende technieken toegepast om de voorkennis van leerlingen in kaart te brengen, b.v. mindmap.

Leerlingen laten zien dat ze leerdoelen en/of eisen waaraan het werk moet voldoen (succescriteria) gebruiken terwijl ze aan het werk zijn. b.v. ze controleren hun voortgang ten opzichte van de leerdoelen en succescriteria voor de les zoals die op het bord of flip-over te lezen zijn

### Vragen Stellen en Klassikale Discussie

<table>
<thead>
<tr>
<th>Item</th>
<th>Details</th>
<th>1a</th>
<th>2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verschillende technieken worden gebruikt om de klassikale discussie te faciliteren b.v. brainstormen</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vragen worden gebruikt om informatie over de voorkennis van leerlingen over een onderwerp.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leerlingen worden aangemoedigd om vragen aan elkaar te stellen tijdens de les. b.v. docent nodigt routinematig de leerlingen uit om vragen te stellen aan de medeleerlingen voor een bijdrage aan discussies</td>
<td></td>
<td>1a</td>
<td></td>
</tr>
<tr>
<td>Vragen stellen gaat verder dan de één-juist-antwoord-stijl (waarbij de focus vaak ligt op het proberen te raden van het antwoord dat de docent in gedachten heeft) naar het gebruiken van meer open vragen die het dieper nadenken bevorderen.</td>
<td></td>
<td>1a</td>
<td></td>
</tr>
<tr>
<td>Onjuiste antwoorden van leerlingen worden gebruikt bij het sturen van het lesgeven en het leren b.v. een leerling wordt gevraagd om uit te leggen waarom hij/zij een bepaald antwoord gaf</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leerlingen kunnen aan anderen uitleggen wat ze aan het leren zijn. b.v. als er een bezoeker in de klas komt kunnen leerlingen verwoorden wat ze aan het leren zijn in termen die aangeven welke kennis, vaardigheden, concepten en/of houdingen werden ontwikkeld</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Feedback

Leerlingen ontvangen feedback die is afgestemd op (de) oorspronkelijke leerdoel(en) en eisen waaraan het werk moet voldoen (succescriteria). b.v. "Vandaag leren we de interpunctie op een juiste manier te gebruiken in ons schrijven; jij gebruikte hoofdletters en punten op de juiste manier in jouw verhaal, goed gedaan Jan"

(verschillende) technieken worden gedurende de lessen gebruikt om de docent te helpen bepalen hoe goed de leerlingen begrijpen wat er in de les is behandeld. b.v. duim omhoog als de lesstof is begrepen en duim omlaag als de lesstof niet is begrepen als feedback voor de docent.

Geschreven feedback op het werk van leerlingen gaat verder dan alleen een cijfer en een opmerking als 'goed gedaan', om vast te stellen wat leerlingen hebben bereikt en wat ze vervolgens moeten doen.

Diagnostische informatie uit toetsen wordt gebruikt om sterke en zwakke punten in het leren van leerlingen vast te stellen. b.v. het aanwijzen van veel gemaakte fouten in het opstellen van breuken.

Diagnostische informatie uit toetsen wordt gebruikt om bij de instructie rekening te houden met sterke en zwakke punten van leerlingen. b.v. een extra lessenserie over optellen van breuken.

Diagnostische informatie uit toetsen wordt gebruikt om sterke en zwakke punten in het lesgeven vast te stellen. b.v. veelgemaakte fouten in eindexamens worden vastgesteld en gebruikt in het lesgeven of leerlingen worden gevaagd om de les te evalueren.

Leerlingen zijn formaal betrokken bij de informatievoorziening over hun leevorderingen aan ouders of verzorgers. b.v. portfolio of schriften thuis laten zien.

Als de docent feedback geeft gaat dit verder dan het geven van het juiste antwoord aan leerlingen, en gebruikt de docent diverse aanwijzingen om leerlingen te helpen vooruitgang te boeken. b.v. ondersteunen van leerlingen door te zeggen: "Je kunt sommige van de nieuwe bijvoeglijke naamwoorden die we vorige week geleerd hebben gebruiken om..."
**Peer- and Self-Assessment**

<table>
<thead>
<tr>
<th>Description</th>
<th>1a</th>
<th>2c</th>
</tr>
</thead>
<tbody>
<tr>
<td>De docent houdt bij of leerlingen vooruit zijn gegaan ten opzichte van de vorige prestatie, en gebruikt die informatie om feedback aan leerlingen te geven. b.v. de docent houdt een logboek, checklist of prestatielijst van zijn leerlingen bij</td>
<td></td>
<td></td>
</tr>
<tr>
<td>De docent vraagt aan mij / de klas wat sterke en zwakke punten van zijn / haar lesgeven zijn.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leerlingen geven aan het begin van de les of van een lesonderdeel aan in welke mate zij zich uitgedaagd voelen door de leertaak. b.v. leerlingen beoordelen voor zichzelf of ze de leertaak: (1) al kunnen, (2) nog niet kunnen maar er wel gaan komen of (3) hulp nodig hebben.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leerlingen worden aangemoedigd om hun voortgang bij te houden. b.v. door het gebruik van een logboek</td>
<td>1a</td>
<td></td>
</tr>
<tr>
<td>Leerlingen beoordelen en reageren op elkaars werk. b.v. leerlingen wordt geleerd hoe ze succescriteria van een leertaak kunnen gebruiken om het werk van een medeleerling te kunnen beoordelen</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leerlingen worden aangemoedigd om een reeks van verschillende werkvormen (assessmenttechnieken) te gebruiken om hun eigen werk te beoordelen. b.v. rubric om inzicht te geven in de succescriteria of het maken van een toetsanalyse zodat de leerling een positief aspect van zijn/haar werk en een verbeterpunt kan benoemen</td>
<td>1a</td>
<td></td>
</tr>
<tr>
<td>Een visuele weergave van de voortgang van leerlingen wordt gebruikt om de groei van leerlingen te vieren en te tonen op welke onderdelen de leerling zich nog kan ontwikkelen. b.v. een grafiek met daarop een overzicht van de progressie over een periode</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tijdens de les is tijd gereserveerd om self- en peer-assessment mogelijk te maken.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tijdens de leerling-bespreking tussen ouder/verzorger en mentor is tijd gereserveerd om de leerling te betrekken in het gesprek door hem of haar te laten rapporteren over het leerproces. b.v. de leerling geeft aan wat goed gaat, wat minder goed gaat en wat hij / zij nodig heeft om verder te komen in het eigen leerproces</td>
<td>2c</td>
<td>Removed: not observable</td>
</tr>
</tbody>
</table>

**Note:** because most items in the questionnaires for teachers and students were the same, only written from different points of view, only the questionnaire for teachers has been screened. The only item that was not in this questionnaire but was in the ones for students has been added in the screening process. Only the questions in the sections of the five strategies have been used.

**Note:** because this is a questionnaire, the items are typically not observable, so all items have been rephrased to sentences that are observable.
### 6.2 Appendix II: Instrument version I

#### Algemene informatie
Datum observatie:  
School:  
Geobserveerde groep, les en docent:  
Aantal leerlingen:  
Naam observant:  

Observeer het gedrag van de docent gedurende de les. Geef eerst bij de voorbeelden aan of dit voorbeeld geobserveerd is (0 = niet geobserveerd, 1= geobserveerd).

Geef vervolgens de items een score van 1 t/m 4 (1= overwegend zwak, 2= meer zwak dan sterk, 3= meer sterk dan zwak, 4= overwegend sterk).

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden geobserveerd?</th>
</tr>
</thead>
</table>
| Leerdoelen en succescriteria duidelijk maken en delen | 1 | De docent maakt duidelijk wat de leerdoelen zijn voor de les. | 1 2 3 4 | De docent vertelt de leerdoelen aan het begin van de les.  
De docent herhaalt de leerdoelen gedurende de les of heeft de leerdoelen voor alle leerlingen zichtbaar in de klas hangen.  
De docent vertelt de leerdoelen in begrijpelijke taal die aansluit bij de leerlingen.  
De docent vertelt voor activiteiten tijdens de les wat de leerdoelen zijn.  
De docent formuleert de leerdoelen in woorden die kennis, vaardigheden, concepten en/of houdingen benadrukken.  
De leerdoelen zijn gefocust op wat er geleerd wordt.  
De docent herinnert de leerlingen aan het verband tussen wat er geleerd wordt en het grotere geheel (bv: geld tellen, zodat je in de supermarkt je wisselgeld kunt controleren). | 0 1 |
| | 2 | De docent maakt duidelijk wat de succescriteria voor de les zijn. | 1 2 3 4 | De docent vertelt de succescriteria voorafgaand aan de les.  
De docent herhaalt de succescriteria gedurende de les of heeft de succescriteria voor alle leerlingen zichtbaar in de klas hangen.  
De docent vertelt de succescriteria in begrijpelijke taal die aansluit bij de doelgroep.  
De docent geeft voorbeelden van kwalitatief goed werk.  
De docent geeft voorbeelden van kwalitatief slecht werk.  
De docent behandelt mogelijke misvattingen omtrent de succescriteria.  
Succescriteria verschillen per leerling (bv: sommige leerlingen krijgen extra werk).  
De docent stelt samen met de leerlingen vast wanneer er aan een succescriterium voldaan wordt. | 0 1 |
## Construeer effectieve klassegesprekken en taken die bewijs van leren ontlokken

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden geobserveerd?</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td></td>
<td>De docent gebruikt verschillende methoden om uit te vinden wat de leerlingen geleerd hebben/al weten.</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>De docent houdt de voortgang van de leerlingen gedurende de les voortdurend in de gaten</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>De docent gebruikt verschillende strategieën om effectief vragen te stellen aan de leerlingen.</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1</td>
</tr>
</tbody>
</table>

## Feedback geven die de leerling verder helpt.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden geobserveerd?</th>
</tr>
</thead>
<tbody>
<tr>
<td>6</td>
<td></td>
<td>De docent geeft leerlingen feedback op wat er geleerd moet worden.</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>De docent gebruikt de kennis over de voortgang van de leerlingen om feedback te geven.</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>De docent vraagt feedback op het lesgeven aan de leerlingen.</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1</td>
</tr>
</tbody>
</table>
## Master thesis: What do we learn from assessment?

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden geobserveerd?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leerlingen activeren als eigenaar van hun eigen leerproces + Leerlingen activeren als instructiebronnen voor andere leerlingen.</td>
<td>9</td>
<td>De docent stimuleert leerlingen hun eigen vooruitgang in de gaten te houden.</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1 0 1 0 1 0 1</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>De docent stimuleert leerlingen het werk van andere leerlingen te beoordelen en feedback te geven aan andere leerlingen.</td>
<td>1 2 3 4</td>
<td>0 1 0 1 0 1 0 1 0 1 0 1</td>
</tr>
</tbody>
</table>

- De docent geeft de leerlingen informatie over hun leerproces.
- De docent vraagt leerling in welke mate ze uitgedaagd worden in de les.
- De docent leert de leerlingen vaardigheden in zelf-monitoren te ontwikkelen.
- De docent stimuleert de leerlingen hun eigen werk te beoordelen.
- De docent stimuleert de leerlingen hun eigen voortgang bij te houden.
- De docent reserveert tijd gedurende de les om leerlingen hun eigen werk te laten nakijken.
- De docent moedigt de leerlingen aan verschillende vormen te gebruiken om hun eigen werk te beoordelen (bv. rubrics, checklists, nakijkbladen).
- De docent stimuleert leerlingen het werk van andere leerlingen te beoordelen en feedback te geven aan andere leerlingen.
- De docent stimuleert de leerlingen om vragen aan elkaar te stellen alvorens ze naar de docent stappen.
- De docent stimuleert het werken in groepen wanneer een opdracht individueel te lastig is.
- De docent gebruikt voorbeelden van leerlingen om veelgemaakte fouten aan de leerlingen te laten zien.
- De docent gebruikt peer activiteiten om de leerlingen alternatieve strategieën aan te leren.
### 6.3 Appendix III: Comments on items and examples during focus group interviews (phase 2)

The table below describes the comments that respondents made on items and examples during the focus group interviews in phase 1.

**Table 3**

Comments made by respondents on items during focus group interviews

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>Written comments respondents</th>
<th>Comments respondents during discussion</th>
</tr>
</thead>
</table>
2: Het woord ‘les’ kan verwarrend zijn als het niet maar over één les gaat. |
2: Het woord ‘les’ kan verwarrend zijn als het niet maar over één les gaat.  
5/6: In de voorbeelden staat een paar keer iets over ‘verschillen per leerling’, maar dit item zegt daar niks over. Als je dat belangrijk vind, zou ik een nieuw item daarover toevoegen. |
6: je wilt weten wat het niveau van de leerling is  
4: gaat over uitvinden waar ze staan, niet alleen feiten, maar ook hogere orde processen. |
| 4 | De docent houdt de voortgang van de leerlingen gedurende de les voortdurend in de gaten | Duo 1: Valt uit de toon naast de andere kaartjes Resp. 5/6: item 4 en 7 lijken op elkaar. Resp. 5/6: monitoren | 5: gaat over monitoren denk ik  
2: gaat over het ontlokken van bewijs  
4: andere twee items zo specifiek dat deze overbodig was  
5: valt eronder omdat het over monitoren gaat, dat is anders bij die anderen.  
2: lastig omdat het vooral over vragen en discussie gaat, terwijl het ook over het ontlokken van bewijs moet gaan. |
| 5 | De docent gebruikt verschillende strategieën om effectief vragen te stellen aan de leerlingen. | Duo 2: Kan ook bij andere indicator                                                               | 6: gaat vooral over de verschillende strategieën                                                                          |
1: gaat over inhoudelijke feedback |
| 7 | De docent gebruikt de kennis over de voortgang van                        | Duo 2: Kan ook bij andere indicator                                                               | 5: gaat over voortgang                                                                                                  |
De leerlingen om feedback te geven.  
Resp. 5/6: item 4 en 7 lijken op elkaar.  
6: zou bijna hetzelfde kunnen zijn als de vorige  
2: misschien samenvoegen  
1: aliebeu over inhoudelijke feedback  
6: deze en de vorige niet onderscheidend genoeg om ze los te koppelen

De docent vraagt feedback op het lesgeven aan de leerlingen.  
Resp. 5/6: gaat over feedback vragen, niet geven.  
6: gaat over feedback vragen en niet geven.  
Voorbeelden zijn wel duidelijk.  
4: derde voorbeeld toegevoegd.

*Item 6,7,8 (indicator 3): mate of kwaliteit van feedback (Duo 3)
Comments written in *Italic* are comments made by teachers, other comments were made by educational researchers.

**Table 4**
Comments made by respondents on examples during focus group interviews

<table>
<thead>
<tr>
<th>Item #</th>
<th>Example</th>
<th>Written comments respondents</th>
<th>Comments respondents during discussion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>De leerdoelen zijn gefocust op wat er geleerd wordt.</td>
<td>Duo 3: ?</td>
<td>Respondent 3: je wilt niet vervallen in telkens hetzelfde, dus dit doe ik niet zo vaak</td>
</tr>
<tr>
<td></td>
<td>De docent herinnert de leerlingen aan het verband tussen wat er geleerd wordt en het grotere geheel (bv: geld tellen, zodat je in de supermarkt je wisselgeld kunt controleren)</td>
<td>Duo 1: Kan bij item 1 en 2, Duo 3: Niet gebruiken</td>
<td></td>
</tr>
<tr>
<td></td>
<td>De docent geeft voorbeelden van kwalitatief slecht werk.</td>
<td>Duo 1: Kan ook bij item 3 en 4, Duo 3: Differentiatie</td>
<td>Resp 3: geen extra werk, wel moeilijker/ander werk</td>
</tr>
<tr>
<td></td>
<td>De docent stelt samen met de leerlingen vast wanneer er aan een succescriterium voldaan wordt.</td>
<td>Duo 3:</td>
<td></td>
</tr>
<tr>
<td></td>
<td>De docent maakt gebruik van brainstormen.</td>
<td>Resp. 5/6: let op voorbeelden die niet altijd in een les naar voren komen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>De docent initieert een klasendiscussie.</td>
<td>Resp. 5/6: let op voorbeelden die niet altijd in een les naar voren komen</td>
<td></td>
</tr>
<tr>
<td></td>
<td>De docent stelt vragen aan de leerlingen.</td>
<td>Resp. 5/6: is breed</td>
<td></td>
</tr>
<tr>
<td></td>
<td>De leerlingen zijn in staat uit te leggen wat ze op dat moment aan het leren zijn</td>
<td>Duo 2: Kan ook bij andere items</td>
<td>4: hoort voor mij echt bij eigenaarschap</td>
</tr>
</tbody>
</table>
De docent maakt notities tijdens de les. Resp. 5/6: is breed

De docent brengt de voorkennis van de leerlingen in kaart. Resp. 5/6: niet logisch hier

De docent gebruikt verschillende methoden gedurende de les om te kijken of leerlingen begrijpen wat er verteld is. Duo 2: Kan ook bij andere items 4: Bij item 3 geplaatst, want het is bijna hetzelfde.

De docent weet wat de sterke en zwakke punten van elke leerling zijn en speelt hierop in. Duo 3: differentiatie Resp 3: je kunt niet in het hoofd van de docent kijken

De docent stelt open vragen. 3: te breed, kan ook over het voetbal van gisteren gaan

De docent maakt gebruik van antwoordmogelijkheden waarbij meerdere leerlingen tegelijk antwoord kunnen geven (bv de leerlingen het antwoord laten opschrijven en omhoog houden). Duo 2: Kan ook bij andere items

De docent behandelt zowel feitelijke kennis als diepere kennis tijdens klassikale discussies/vragen. Duo 2: Kan ook bij andere items Resp. 3: bij item 3

De docent geeft feedback op basis van criteria. Duo 3: ‘criteria’ is vaag

De docent vermijdt feedback op basis van vergelijkingen met andere leerlingen. Duo 2: Kan ook bij andere items

De feedback van de docent beschrijft specifieke punten waar verbetering nodig is. Duo 2: Kan ook bij andere items

De docent gebruikt de voortgang van de leerlingen om feedback te geven. Duo 3: ‘de voortgang’ is vaag Resp. 5/6: lijkt erg veel op het item

De docent gebruikt voorbeelden van leerlingen om veelgemaakte fouten aan de leerlingen te laten zien. Duo 2: Kan ook bij andere items 2: bij item 2 ingedeeld, want voorbeelden maken duidelijk wat successcriteria zijn. 1: dit gaat over voorbeelden van de leerlingen zelf, bij item 2 staat nog iets over kwalitatief goed/slecht werk. Alle duos’ alleen gebruiken wanneer het volledig anoniem is, anders schaadt je het zelfvertrouwen van de leerling enorm

De docent gebruikt peer activiteiten om de leerlingen alternatieve strategieën aan te leren. Duo 3: peer?

Comments written in *Italic* are comments made by teachers, other comments were made by educational researchers.
### 6.4 Appendix IV: Detailed information on changes made in examples during first revision

The table below described the changes made and the decisions made towards these changes for the first revision of the instrument, regarding the examples.

**Table 5**

*Changes made in examples during the revision of the instrument*

<table>
<thead>
<tr>
<th>Example</th>
<th>Change made*</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>De docent herinnert de leerlingen aan het verband tussen wat er geleerd wordt en het grotere geheel (bv: geld tellen, zodat je in de supermarkt je wisselgeld kunt controleren)</td>
<td>Removed</td>
<td>This instrument focuses on secondary education and this mostly occurs in primary education, rarely in secondary education.</td>
</tr>
<tr>
<td>De docent geeft voorbeelden van kwalitatief goed werk.</td>
<td>Combined to a more neutral formulation</td>
<td>‘Poor quality’ was too negative according to some of the teachers.</td>
</tr>
<tr>
<td>Succescriteria verschillen per leerling (bv: sommige leerlingen krijgen extra werk)</td>
<td>Rewritten into an item</td>
<td>The items for this indicator did not cover this, but it is an important part of it, so it became a new item.</td>
</tr>
<tr>
<td>De docent stelt samen met de leerlingen vast wanneer er aan een succescriterium voldaan wordt.</td>
<td>Replaced to item 9</td>
<td>Most respondents found this example more fitted for another item.</td>
</tr>
<tr>
<td>De docent stelt vragen aan de leerlingen.</td>
<td>Replaced to item 5</td>
<td>Most respondents found this example more fitted for another item.</td>
</tr>
<tr>
<td>Wanneer een leerling de vraag niet begrijpt, stelt de docent de vraag op een andere manier, zodat de leerling het beter begrijpt.</td>
<td>Replaced to item 5</td>
<td>Most respondents found this example more fitted for another item.</td>
</tr>
<tr>
<td>De leerlingen zijn in staat uit te leggen wat ze op dat moment aan het leren zijn</td>
<td>Replaced to item 9</td>
<td>Most respondents found this example more fitted for another item.</td>
</tr>
<tr>
<td>De docent kijkt tijdens de les waar de leerlingen mee bezig zijn.</td>
<td>Replaced to item 7</td>
<td>Because this item was removed, the examples that were not up for discussion were replaced to item 7, the item that is very closely related to his item.</td>
</tr>
<tr>
<td>De docent maakt notities tijdens de les.</td>
<td>Removed</td>
<td>Did not have added value.</td>
</tr>
<tr>
<td>De docent brengt de voorkennis van de leerlingen in kaart.</td>
<td>Replaced to item 3</td>
<td>Most respondents found this example more fitted for another item.</td>
</tr>
<tr>
<td>De docent gebruikt verschillende methoden gedurende de les om te kijken of leerlingen begrijpen wat er verteld is.</td>
<td>Replaced to item 3</td>
<td>Most respondents found this example more fitted for another item.</td>
</tr>
<tr>
<td>De docent weet wat de sterke en zwakke punten van elke leerling zijn en speelt hierop in</td>
<td>Replaced to item 7</td>
<td>Because this item was removed, the examples that were not up for discussion were replaced to item 7, the item that is very closely related to his item.</td>
</tr>
<tr>
<td>De docent stelt open vragen.</td>
<td>Combined with ‘de docent stelt vragen aan de leerlingen’</td>
<td>These examples were too closely related to keep them both, so they are combined.</td>
</tr>
<tr>
<td>Als een leerling een fout antwoord geeft, helpt de docent de leerling om tot het juiste antwoord te komen.</td>
<td>Replaced to item 6</td>
<td>Most respondents found this example more fitted for another item.</td>
</tr>
<tr>
<td>De docent maakt gebruik van antwoordmogelijkheden waarbij meerdere leerlingen tegelijk antwoord kunnen geven (bv de leerlingen het antwoord laten opschrijven en omhoog houden).</td>
<td>Example removed</td>
<td>The example did not have added value and is therefore removed.</td>
</tr>
<tr>
<td>De docent vernietigt feedback op basis van vergelijkingen met andere leerlingen.</td>
<td>Removed</td>
<td>Disagreement among respondents, where no consensus was reached, therefore this example is removed.</td>
</tr>
<tr>
<td>Item</td>
<td>Description</td>
<td>Change</td>
</tr>
<tr>
<td>------</td>
<td>-------------</td>
<td>--------</td>
</tr>
<tr>
<td>6</td>
<td>Docent geeft leerlingen feedback op hun leerproces.</td>
<td>Replaced to item 6</td>
</tr>
<tr>
<td>6</td>
<td>Docent geeft inhoudelijke feedback (meer dan alleen ‘goed’ of ‘fout’) die de leerlingen richting geeft om verder te komen in hun proces.</td>
<td>Replaced to item 6</td>
</tr>
<tr>
<td>7</td>
<td>Docent geeft de leerlingen informatie over hun leerproces.</td>
<td>Replaced to item 7</td>
</tr>
<tr>
<td>8</td>
<td>Docent vraagt leerling in welke mate ze uitgedaagd worden in de les.</td>
<td>Replaced to item 8</td>
</tr>
<tr>
<td></td>
<td>Docent gebruikt peer activiteiten om de leerlingen alternatieve strategieën aan te leren.</td>
<td>Removed</td>
</tr>
</tbody>
</table>

*Item number corresponds with the first draft of the instrument.*
### 6.5 Appendix V: Instrument version II (revision after phase 2)

**Algemene informatie**
Datum observatie:
School:
Geobserveerde groep, les en docent:
Aantal leerlingen:
Naam observant:

Observeer het gedrag van de docent gedurende de les. Geef de items een score van 1 t/m 4 (1 = overwegend zwak, 2 = meer zwak dan sterk, 3 = meer sterk dan zwak, 4 = overwegend sterk). De voorbeelden dienen als hulpmiddel om een beter beeld van het bijbehorende item te krijgen en hoeven dus niet te worden gescoord.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Leerdoelen en succescriteria duidelijk maken en delen | 1 | De docent maakt duidelijk wat de leerdoelen zijn van de les. | 1 2 3 4 | - De docent vertelt de leerdoelen aan het begin van de les.  
- De docent herhaalt de leerdoelen gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de leerdoelen in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent vertelt voor activiteiten tijdens de les wat de leerdoelen zijn.  
- De docent formuleert de leerdoelen in woorden die kennis, vaardigheden, concepten en/of houdingen benadrukken.  
- De leerdoelen zijn gefocust op wat er geleerd wordt. |
| | 2 | De leerlingen laten zien dat ze begrijpen wat de leerdoelen zijn door hiermee aan de slag te gaan. | 1 2 3 4 | - De leerlingen hebben de leerdoelen voor zichzelf opgeschreven.  
- De leerlingen werken gericht aan de leerdoelen.  
- Leerlingen strepen een leerdoel af wanneer ze het behaald (denken te) hebben.  
- Wanneer de leerlingen vastlopen bij een leerdoel, vragen ze hulp aan medeleerlingen of aan de docent. |
| | 3 | De docent maakt duidelijk wat de succescriteria van de les zijn. | 1 2 3 4 | - De docent vertelt de succescriteria voorafgaand aan de les.  
- De docent herhaalt de succescriteria gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de succescriteria in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent geeft voorbeelden van de te verwachten kwaliteit van het werk.  
- De docent behandelt mogelijke misvattingen omtrent de succescriteria. |
| | 4 | De leerlingen laten zien dat ze begrijpen wat de succescriteria zijn door hun werk hierop af te stemmen. | 1 2 3 4 | - De leerlingen hebben de succescriteria voor zichzelf opgeschreven.  
- De leerlingen werken toe naar het behalen van de succescriteria  
- De leerlingen strepen een succescriterium door wanneer ze (denken) hieraan (te) voldoen.  
- Wanneer leerlingen een succescriterium niet begrijpen, vragen ze hulp aan medeleerlingen of aan de docent. |
| | 5 | Succescriteria verschillen per leerling | 1 2 3 4 | - De docent geeft sommige leerlingen extra werk  
- De docent geeft sommige leerlingen gemakkelijker werk  
- De docent geeft sommige leerlingen moeilijker werk |
**Master thesis: What do we learn from assessment?**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| **Construeer effectieve klassegesprekken en taken die bewijs van leren ontlokken** | 6 | De docent gebruikt verschillende methoden om uit te vinden wat de voorkennis van de leerlingen is. | 1 2 3 4 | - De docent maakt gebruik van brainstormen.  
- De docent initieert een klasendiscussie.  
- De docent brengt de voorkennis van de leerlingen in kaart.  
- De docent gebruikt verschillende methoden gedurende de les om te kijken of leerlingen begrijpen wat er verteld is. |
|                                | 7 | De docent gebruikt verschillende strategieën om effectief vragen te stellen aan de leerlingen. | 1 2 3 4 | - De docent stelt open vragen aan de leerlingen.  
- De docent hanteert een adequate wachtijd met betrekking tot het antwoorden van leerlingen.  
- De docent laat alle leerlingen het antwoord opschrijven alvorens een leerling antwoordt.  
- De docent bepaalt op een willekeurige manier welke leerling moet antwoorden.  
- De docent maakt gebruik van antwoordmogelijkheden waarbij meerdere leerlingen tegelijk antwoord kunnen geven.  
- De docent behandelt zowel feitelijke kennis als diepere kennis tijdens klassikale discussies/vragen.  
- Als een leerling de vraag niet begrijpt, stelt de docent de vraag op een andere manier, zodat de leerling het beter begrijpt. |
| **Feedback geven die de leerling verder helpt.** | 8 | De docent geeft leerlingen feedback op wat er geleerd wordt. | 1 2 3 4 | - De docent geeft feedback op basis van criteria.  
- De feedback van de docent beschrijft specifieke punten waar verbetering nodig is.  
- De feedback van de docent bevat strategieën om de nodige verbeteringen te maken.  
- Als een leerling een fout antwoord geeft, helpt de docent de leerling om tot het juiste antwoord te komen.  
- De docent geeft leerlingen feedback op hun leerproces.  
- De docent geeft inhoudelijke feedback (meer dan alleen ‘goed’ of ‘fout’) die de leerlingen richting geeft om verder te komen in hun proces.  
- De geschreven feedback gaat in op de inhoud van de taak. |
|                                | 9 | De docent gebruikt de kennis over de voortgang van de leerlingen om feedback te geven. | 1 2 3 4 | - De docent kijkt tijdens de les waar de leerlingen mee bezig zijn.  
- De docent geeft de leerlingen informatie over hun leerproces.  
- De docent weet wat de sterke en zwakkere punten van elke leerling zijn en speelt hierop in.  
- De docent geeft feedback op basis van toetsen die de leerlingen eerder gemaakt hebben.  
- De docent gebruikt de voortgang van de leerlingen om feedback te geven.  
- Als de docent positieve feedback geeft, wordt ook aangegeven hoe de leerling dit niveau kan handhaven. |
### Indicator: Leerlingen activeren als eigenaar van hun eigen leerproces

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| 10   | De docent stimuleert leerlingen hun eigen vooruitgang in de gaten te houden. | 1 2 3 4 | - De leerlingen zijn in staat uit te leggen wat ze op dat moment aan het leren zijn.  
- De docent leert de leerlingen vaardigheden in zelf-monitoring te ontwikkelen.  
- De docent stimuleert de leerlingen hun eigen werk te beoordelen.  
- De docent stimuleert de leerlingen hun eigen voortgang bij te houden.  
- De docent reserveert tijd gedurende de les om leerlingen hun eigen werk te laten nakijken.  
- De docent moedigt de leerlingen aan verschillende vormen te gebruiken om hun eigen werk te beoordelen (bv. rubrics, checklists, nakijkbladen)  
- De docent stelt samen met de leerlingen vast wanneer er aan een succescriterium voldaan wordt.    |
| 11   | De docent vraagt feedback op het lesgeven aan de leerlingen              | 1 2 3 4 | - De docent vraagt of zijn/haar uitleg duidelijk is.  
- De docent vraagt leerlingen aan te geven wat er anders kan m.b.t. de instructie.  
- De docent vraagt leerlingen of ze uitgedaagd worden in de les. |

### Indicator: Leerlingen activeren als instructiebronnen voor andere leerlingen.

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| 12   | De docent stimuleert leerlingen het werk van andere leerlingen te beoordelen en feedback te geven aan andere leerlingen. | 1 2 3 4 | - De docent reserveert tijd om gedurende de les het werk van medeleerlingen te beoordelen.  
- De docent stimuleert de leerlingen om vragen aan elkaar te stellen alvorens ze naar de docent stappen.  
- De docent stimuleert het werken in groepen wanneer een opdracht individueel te lastig is.  
- De docent gebruikt geanonimiseerde voorbeelden van leerlingen om veelgemaakte fouten aan de leerlingen te laten zien.  
- De docent stimuleert leerlingen om vragen aan elkaar te stellen gedurende de les. |
6.6 Appendix VI: Instrument version III (revision after phase 3, round 1)

Algemene informatie
Datum en tijd (lesuur) observatie:__________________________
School:_______________________________________________
Geobserveerde groep en docent:_______________________________________________________________________________________________
Geobserveerd vak en soort les (instructie, zelfstandig werken, toets etc.) ____________________________________________________________________
Aantal leerlingen:______________________________________
Naam observant:_______________________________________

Observeer het gedrag van de docent gedurende de les. Geef de items een score van 1 t/m 4 (1= overwegend zwak, 2= meer zwak dan sterk, 3= meer sterk dan zwak, 4= overwegend sterk). De voorbeelden dienen als hulpmiddel om een beter beeld van het bijbehorende item te krijgen en hoeven dus niet te worden gescoord.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Leerdoelen en succescriteria duidelijk maken en delen | 1 | De docent maakt duidelijk wat de leerdoelen zijn van de les op een manier die de leerlingen begrijpen. | 1 2 3 4 | - De docent vertelt de leerdoelen aan het begin van de les.  
- De docent herhaalt de leerdoelen gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de leerdoelen in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent vertelt voor activiteiten tijdens de les wat de leerdoelen zijn.  
- De docent formuleert de leerdoelen in woorden die kennis, vaardigheden, concepten en/of houdingen benadrukken.  
- De leerlingen laten zien dat ze de leerdoelen begrijpen door hiermee aan de slag te gaan. |
| Leerdoelen verschillen per leerling. | 2 | 1 2 3 4 | - De docent geeft sommige leerlingen extra werk  
- De docent geeft sommige leerlingen gemakkelijker werk  
- De docent geeft sommige leerlingen moeilijker werk |
| Succescriteria verschillen per leerling. | 3 | De docent maakt duidelijk wat de succescriteria van de les zijn op een manier die de leerlingen begrijpen | 1 2 3 4 | - De docent vertelt de succescriteria voorafgaand aan de les.  
- De docent herhaalt de succescriteria gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de succescriteria in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent vertelt de succescriteria in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent geeft voorbeelden van de te verwachten kwaliteit van het werk.  
- De docent behandelt mogelijke misvattingen omtrent de succescriteria.  
- De leerlingen laten zien dat ze de succescriteria begrijpen door hun werk hierop af te stemmen. |
| Succescriteria verschillen per leerling. | 4 | 1 2 3 4 | - De docent geeft sommige leerlingen extra werk  
- De docent geeft sommige leerlingen gemakkelijker werk  
- De docent geeft sommige leerlingen moeilijker werk |
<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| **Construeer effectieve klassengesprekken en taken die bewijs van leren ontlokken** | 5  | De docent gebruikt verschillende methoden om uit te vinden wat de voorkennis van de leerlingen is. | 1 2 3 4 | - De docent maakt gebruik van brainstormen.  
- De docent initieert een klasendiscussie.  
- De docent stelt klassiskaal vragen om de voorkennis van de leerlingen in kaart te brengen.  
- De docent stelt vragen aan individuele leerlingen om de voorkennis in kaart te brengen.  
- De docent laat leerlingen opdrachten maken om zo de voorkennis in kaart te brengen.  
- De docent gebruikt verschillende methoden gedurende de les om te kijken of leerlingen begrijpen wat er verteld is.  
- De docent stelt open vragen aan de leerlingen.  
- De docent hanteert een adequate wachtijd met betrekking tot het antwoorden van leerlingen.  
- De docent laat alle leerlingen het antwoord schrijven alvorens een leerling antwoordt.  
- De docent bepaalt op een willekeurige manier welke leerling moet antwoorden.  
- De docent maakt gebruik van antwoordmogelijkheden waarbij meerdere leerlingen tegelijk antwoord kunnen geven.  
- Als een leerling de vraag niet begrijpt, stelt de docent de vraag op een andere manier, zodat de leerling het beter begrijpt. |
| **Feedback geven die de leerling verder helpt.** | 7  | De docent geeft leerlingen feedback op wat er geleerd wordt. | 1 2 3 4 | - De docent geeft feedback op basis van criteria.  
- De feedback van de docent beschrijft specifieke punten waar verbetering nodig is.  
- De feedback van de docent bevat strategieën om de nodige verbeteringen te maken.  
- Als een leerling een fout antwoord geeft, helpt de docent de leerling om tot het juiste antwoord te komen.  
- De docent geeft leerlingen feedback op hun leerproces.  
- De docent geeft inhoudelijke feedback (meer dan alleen ‘goed’ of ‘fout’) die de leerlingen richting geeft om verder te komen in hun proces.  
- De feedback gaat in op de inhoud van de taak.  
- De docent geeft feedback op basis van toetsen die de leerlingen eerder gemaakt hebben.  
- De docent gebruikt de voortgang van de leerlingen om feedback te geven.  
- Als de docent positieve feedback geeft, wordt ook aangegeven hoe de leerling dit niveau kan handhaven.  

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| 8         | 8  | De docent gebruikt de kennis over de voortgang van de leerlingen om feedback te geven. | 1 2 3 4 | - De docent kijkt tijdens de les waar de leerlingen mee bezig zijn.  
- De docent geeft de leerlingen informatie over hun leerproces.  
- De docent weet wat de sterke en zwakke punten van elke leerling zijn en speelt hierop in.  
- De docent geeft feedback op basis van toetsen die de leerlingen eerder gemaakt hebben.  
- De docent gebruikt de voortgang van de leerlingen om feedback te geven.  
- Als de docent positieve feedback geeft, wordt ook aangegeven hoe de leerling dit niveau kan handhaven.  
| 9         | 9  | De docent gebruikt feedback van de leerlingen op het lesgeven direct om instructie aan te passen. | 1 2 3 4 | - De leerlingen geven aan dat de uitleg niet duidelijk is, waarna de docent de uitleg op een andere manier geeft.  
- De leerlingen geven aan dat ze een bepaalde manier van instructie niet prettig vinden, waarna de docent dit aanpast.  
- Wanneer leerlingen uitdaging lijken te missen, zoekt de docent naar een manier om deze uitdaging te bieden.  
- Wanneer de stof te lastig lijkt voor de leerlingen, zoekt de docent een manier om het toch begrijpbaar te maken.  

...
### Indicator: Leerlingen activeren als eigenaar van hun eigen leerproces

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| 10 | De docent stimuleert leerlingen hun eigen vooruitgang in de gaten te houden. | 1 2 3 4 | - De leerlingen zijn in staat uit te leggen wat ze op dat moment aan het leren zijn.  
- De docent leert de leerlingen vaardigheden in zelf-monitoren te ontwikkelen.  
- De docent stimuleert de leerlingen hun eigen werk te beoordelen.  
- De docent stimuleert de leerlingen hun eigen voortgang bij te houden.  
- De docent reserveert tijd gedurende de les om leerlingen hun eigen werk te laten nakijken.  
- De docent moedigt de leerlingen aan verschillende vormen te gebruiken om hun eigen werk te beoordelen (bv. rubrics, checklists, nakijkbladen)  
- De docent stelt samen met de leerlingen vast wanneer er aan een succescriterium voldaan wordt. |
| 11 | De docent vraagt feedback op het lesgeven aan de leerlingen.         | 1 2 3 4 | - De docent vraagt of zijn/haar uitleg duidelijk is.  
- De docent vraagt leerlingen aan te geven wat er anders kan m.b.t. de instructie.  
- De docent vraagt leerlingen of ze uitgedaagd worden in de les. |

### Indicator: Leerlingen activeren als instructiebronnen voor andere leerlingen.

<table>
<thead>
<tr>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| 12 | De docent stimuleert leerlingen het werk van andere leerlingen te beoordelen en feedback te geven aan andere leerlingen. | 1 2 3 4 | - De docent reserveert tijd om gedurende de les het werk van medeleerlingen te beoordelen.  
- De docent stimuleert de leerlingen om vragen aan elkaar te stellen alvorens ze naar de docent stappen.  
- De docent stimulateert het werken in groepen wanneer een opdracht individueel te lastig is.  
- De docent gebruikt geanonimiseerde voorbeelden van leerlingen om veelgemaakte fouten aan de leerlingen te laten zien.  
- De docent stimuleert leerlingen om vragen aan elkaar te stellen gedurende de les. |

Overige opmerkingen:

______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________
______________________________________________________________________________
6.7 Appendix VII: Instrument version IV (revision after phase 3, round 2)
Algemene opmerkingen bij het gebruik van het observatie instrument voor AfL

Voor de observatie:
- Zorg ervoor dat je, voor je dit instrument gebruikt, kennis hebt opgedaan over AfL en voornamelijk over de vijf strategieën die gebruikt worden om AfL te implementeren. Dit is noodzakelijke kennis om het instrument te kunnen begrijpen.
- Wanneer je met meerdere personen tegelijk observeert, bespreek dan van tevoren hoe je de scoring-opties en de items/voorbeelden interpreteert. Dit bespreken zorgt voor minder verwarring achteraf. Zorg ook dat je samen een testles kunt observeren (dit kan bijvoorbeeld ook door een video van een les op YouTube te bekijken) en bespreek deze achteraf. Eventuele interpretatieverschillen die eerder nog niet duidelijk waren, kunnen zo alsnog gevonden en opgelost worden.
- Wanneer je dit instrument gebruikt om het handelen van een docent te evalueren, is het raadzaam te observeren gedurende een aantal verschillende lessen, bijvoorbeeld een les waarin een toets besproken wordt, een instructie en een les waarin zelfstandig gewerkt wordt. Tijdens het testen van het instrument was te zien dat niet alle zaken in één les aan bod komen en dat het handelen van een docent dus ook niet volledig in één les te observeren is. Let er wel op dat de les die geobserveerd wordt telkens bij dezelfde klas is en tijdens hetzelfde lesuur op dezelfde dag; dit zijn externe factoren die invloed kunnen hebben op het handelen van een docent die we graag zo klein mogelijk houden.

Tijdens de observatie:
- Vul altijd alle informatie in die gevraagd wordt. Dit is belangrijk, omdat dit invloed kan hebben op hetgeen er geobserveerd wordt (bijvoorbeeld: een les op maandagochtend verloopt vaak wat anders dan het laatste uur op vrijdag).
- Alle items kunnen gescoord worden tijdens zowel klassikale als individuele instructie, ook als dit niet expliciet vermeld is.
- Bij het eerste item ontstaat nog wel eens verwarring over de term ‘leerdoel’. Het is belangrijk dat je deze ook echt als leerdoel en niet als lesdoel opvat. Het gaat er dus om dat een docent vertelt wat het doel is van wat ze leren en niet zozeer dat de docent vertelt wat ze die les gaan doen. Vertellen wat je gaat doen kan natuurlijk een eerste stap hierin zijn, maar het is niet waar je uiteindelijk naar toe wil. Mocht je dit zien, dan is het handig om dat onderaan het instrument aan te geven en het eventueel met de docent te bespreken.
Na de observatie:

- Dit instrument is bedoeld om te evalueren en niet om te beoordelen. Het is dan ook niet zo dat de docent een bepaalde score dient te behalen; de scores zijn alleen een hulpmiddel om aan te geven waar welke mate van verbetering gewenst is.

- Bespreek de resultaten met de docent. Dit kan na één les, maar omdat dit niet altijd een compleet beeld geeft, is het raadzaam dit pas te doen nadat alle lessen geobserveerd zijn. Maak dan wel duidelijke notities, zodat je de docent ook nog duidelijke aanwijzingen kunt geven over zaken die tijdens de eerste les zijn opgevallen. Probeer hierbij ook de voorbeelden te betrekken; deze geven vaak een goede schets van wat er zich in de praktijk afspeelt en kunnen duidelijke handvatten zijn voor docenten.
**Algemene informatie**

Datum en tijd (lesuur) observatie: __________________________
School: __________________________
Geobserveerde groep en docent: ____________________________________________
Geobserveerd vak en soort les (instructie, zelfstandig werken, toets etc.) ____________________________
Aantal leerlingen: ____________________________
Naam observant: ____________________________

**Observeer het gedrag van de docent gedurende de les. Geef de items een score van 1 t/m 4 (1= overwegend zwak, 2= meer zwak dan sterk, 3= meer sterk dan zwak, 4= overwegend sterk). De voorbeelden dienen als hulpmiddel om een beter beeld van het bijbehorende item te krijgen en hoeven dus niet te worden gescoord.**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Leerdoelen en succescriteria duidelijk maken en delen | 1 | De docent maakt duidelijk wat de leerdoelen zijn op een manier die de leerlingen begrijpen. | 1 2 3 4 | - De docent vertelt wat de leerdoelen voor deze les zijn.  
- De docent vertelt wat de leerdoelen op lange termijn zijn.  
- De docent vertelt de leerdoelen aan het begin van de les.  
- De docent herhaalt de leerdoelen gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de leerdoelen in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent vertelt voor activiteiten tijdens de les wat de leerdoelen zijn.  
- De docent formuleert de leerdoelen in woorden die kennis, vaardigheden, concepten en/of houdingen benadrukken.  
- De leerdoelen zijn gefocust op wat er geleerd wordt.  
- De leerlingen laten zien dat ze de leerdoelen begrijpen door hiermee aan de slag te gaan. |
| | 2 | De docent past het lesgeven aan op basis van de verschillen in tempo van leerlingen. | 1 2 3 4 | - De docent geeft sommige leerlingen extra werk.  
- De docent geeft sommige leerlingen minder werk.  
- Sommige leerlingen mogen eerder gaan, nadat ze hun werk hebben afgerond.  
- Sommige leerlingen blijven wat langer zitten om hun werk af te kunnen maken. |
| | 3 | De docent maakt duidelijk wat de succescriteria zijn op een manier die de leerlingen begrijpen | 1 2 3 4 | - De docent vertelt wat de succescriteria voor deze les zijn.  
- De docent vertelt wat de succescriteria op lange termijn zijn.  
- De docent vertelt de succescriteria voorafgaand aan de les.  
- De docent herhaalt de succescriteria gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de succescriteria in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent geeft voorbeelden van de te verwachten kwaliteit van het werk.  
- De docent behandelt mogelijke misvattingen omtrent de succescriteria.  
- De leerlingen laten zien dat ze de succescriteria begrijpen door hun werk hierop af te stemmen. |
| | 4 | De docent past het lesgeven aan op basis van de verschillen in niveau van leerlingen. | 1 2 3 4 | - De docent geeft sommige leerlingen gemakkelijker werk  
- De docent geeft sommige leerlingen moeilijker werk |
<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Construkteer effectieve     | 5 | De docent gebruikt verschillende methoden om uit te vinden wat de   | 1 2 3 4 | - De docent maakt gebruik van brainstormen.  
- De docent initieert een klasendiscussie.  
- De docent stelt klassikaal vragen om de voorkennis van de leerlingen in kaart te brengen.  
- De docent stelt vragen aan individuele leerlingen om de voorkennis in kaart te brengen.  
- De docent laat leerlingen opdrachten maken om zo de voorkennis in kaart te brengen.  
- De docent gebruikt verschillende methoden gedurende de les om te kijken of leerlingen begrijpen wat er verteld is.                                                                                                                                                                           |
| klassengezamen en taken die  |    | voorkennis van de leerlingen is.                                    |       |                                                                                                                                                                                                                                                                                                                                                                                                    |
| bewijs van leren ontkomen    |    |                                                                      |       |                                                                                                                                                                                                                                                                                                                                                                                                    |
|                               | 6 | De docent gebruikt verschillende strategieën om effectief vragen    | 1 2 3 4 | - De docent stelt open vragen aan de leerlingen.  
- De docent hanteert een adequate wachtijd met betrekking tot het antwoorden van leerlingen.  
- De docent laat alle leerlingen het antwoord opschrijven alvorens een leerling antwoordt.  
- De docent bepaalt op een willekeurige manier welke leerling moet antwoorden.  
- De docent maakt gebruik van antwoordmogelijkheden waarbij meerdere leerlingen tegelijk antwoord kunnen geven.  
- De docent behandelt zowel feitelijke kennis als diepere kennis tijdens klassikale discussies/vragen.  
- Als een leerling de vraag niet begrijpt, stelt de docent de vraag op een andere manier, zodat de leerling het beter begrijpt.                                                                                                                                                                                                                     |
|                               |    | te stellen aan de leerlingen (klassikaal of individueel).           |       |                                                                                                                                                                                                                                                                                                                                                                                                    |

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Feedback geven die de leerling | 7 | De docent geeft leerlingen feedback op wat er geleerd wordt.         | 1 2 3 4 | - De docent geeft feedback op basis van criteria.  
- De feedback van de docent beschrijft specifieke punten waar verbetering nodig is.  
- De feedback van de docent bevat strategieën om de nodige verbeteringen te maken.  
- Als een leerling een fout antwoord geeft, helpt de docent de leerling om tot het juiste antwoord te komen.  
- De docent geeft leerlingen feedback op hun leerproces.  
- De docent geeft inhoudelijke feedback (meer dan alleen ‘goed’ of ‘fout’) die de leerlingen richting geeft om verder te komen in hun proces.  
- De feedback gaat in op de inhoud van de taak.  
- Als een docent positieve feedback geeft, wordt ook aangegeven hoe de leerling dit niveau kan handhaven.                                                                                                                                                                                                                      |
| verder helpt.                 |    |                                                                      |       |                                                                                                                                                                                                                                                                                                                                                                                                    |
|                               | 8 | De docent gebruikt de kennis over de voortgang van de leerlingen om  | 1 2 3 4 | - De docent kijkt tijdens de les waar de leerlingen mee bezig zijn.  
- De docent geeft de leerlingen informatie over hun leerproces.  
- De docent weet wat de sterke en zwakke punten van elke leerling zijn en speelt hierop in.  
- De docent geeft feedback op basis van toetsen die de leerlingen eerder gemaakt hebben.  
- De docent gebruikt de voortgang van de leerlingen om feedback te geven.  
- Als de docent positieve feedback geeft, wordt ook aangegeven hoe de leerling dit niveau kan handhaven.                                                                                                                                                                                                                      |
|                               |    | om feedback te geven.                                                |       |                                                                                                                                                                                                                                                                                                                                                                                                    |
## Leerlingen activeren als eigenaar van hun eigen leerproces

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leerlingen activeren als eigenaar van hun eigen leerproces</td>
<td>9</td>
<td>De docent stimuleert leerlingen hun eigen vooruitgang in de gaten te houden.</td>
<td>1 2 3 4</td>
<td>- De leerlingen zijn in staat uit te leggen wat ze op dat moment aan het leren zijn. - De docent leert de leerlingen vaardigheden in zelf-monitoren te ontwikkelen. - De docent stimuleert de leerlingen hun eigen werk te beoordelen. - De docent stimuleert de leerlingen hun eigen voortgang bij te houden. - De docent reserveert tijd gedurende de les om leerlingen hun eigen werk te laten nakijken. - De docent moedigt de leerlingen hun eigen voortgang bij te houden. - De docent reserveert tijd gedurende de les om leerlingen hun eigen werk te laten nakijken. - De docent stimuleert de leerlingen hun eigen zaak aan te nemen bij het leren. - De docent stimuleert de leerlingen hun eigen werk te beoordelen (bv. rubrics, checklists, nakijkbladen). - De docent stelt samen met de leerlingen vast wanneer er aan een succes criterium voldaan wordt.</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leerlingen activeren als instructiebronnen voor andere leerlingen.</td>
<td>11</td>
<td>De docent stimuleert leerlingen het werk van andere leerlingen te beoordelen en feedback te geven aan andere leerlingen.</td>
<td>1 2 3 4</td>
<td>- De docent stimuleert leerlingen het werk van medeleerlingen te beoordelen. - De docent stimuleert de leerlingen om vragen aan elkaar te stellen alvorens ze naar de docent stappen. - De docent stimuleert het werken in groepen wanneer een opdracht individueel te lastig is. - De docent gebruikt geanonimiseerde voorbeelden van leerlingen om veelgemaakte fouten aan de leerlingen te laten zien. - De docent stimuleert leerlingen om vragen aan elkaar te stellen gedurende de les.</td>
</tr>
</tbody>
</table>

Overige opmerkingen:

____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
6.8 Appendix VIII: Instrument version V, final version (revision after phase 3, round 3)

Algemene opmerkingen bij het gebruik van het observatie instrument voor AfL

Voor de observatie:

- Zorg ervoor dat je, voor je dit instrument gebruikt, kennis hebt opgedaan over AfL en voornamelijk over de vijf strategieën die gebruikt worden om AfL te implementeren. Dit is noodzakelijke kennis om het instrument te kunnen begrijpen.
- Wanneer je met meerdere personen tegelijk observeert, bespreek dan van tevoren hoe je de scoring-opties en de items/voorbeelden interpreteert. Dit bespreken zorgt voor minder verwarring achteraf. Zorg ook dat je samen een testles kunt observeren (dit kan bijvoorbeeld ook door een video van een les op YouTube te bekijken) en bespreek deze achteraf. Eventuele interpretatieverschillen die eerder nog niet duidelijk waren, kunnen zo alsnog gevonden en opgelost worden.
- Wanneer je dit instrument gebruikt om het handelen van een docent te evalueren, is het raadzaam te observeren gedurende een aantal verschillende lessen, bijvoorbeeld een les waarin een toets besproken wordt, een instructie en een les waarin zelfstandig gewerkt wordt. Tijdens het testen van het instrument was te zien dat niet alle zaken in één les aan bod komen en dat het handelen van een docent dus ook niet volledig in één les te observeren is. Let er wel op dat de les die geobserveerd wordt telkens bij dezelfde klas is en tijdens hetzelfde lesuur op dezelfde dag; dit zijn externe factoren die invloed kunnen hebben op het handelen van een docent die we graag zo klein mogelijk houden.

Tijdens de observatie:

- Vul altijd alle informatie in die gevraagd wordt. Dit is belangrijk, omdat dit invloed kan hebben op hetgeen er geobserveerd wordt (bijvoorbeeld: een les op maandagochtend verloopt vaak wat anders dan het laatste uur op vrijdag).
- Alle items kunnen gescoord worden tijdens zowel klassikale als individuele instructie, ook als dit niet expliciet vermeld is.
- Bij het eerste item ontstaat nog wel eens verwarring over de term ‘leerdoel’. Het is belangrijk dat je deze ook echte als leerdoel en niet als lesdoel opvat. Het gaat er dus om dat een docent vertelt wat het doel is van wat ze leren en niet zozeer dat de docent vertelt wat ze die les gaan doen. Vertellen wat je gaat doen kan natuurlijk een eerste stap hierin zijn, maar het is niet waar je uiteindelijk naar toe wil. Mocht je dit zien, dan is het handig om dat onderaan het instrument aan te geven en het eventueel met de docent te bespreken.
Na de observatie:

- Dit instrument is bedoeld om te evalueren en niet om te beoordelen. Het is dan ook niet zo dat de docent een bepaalde score dient te behalen; de scores zijn alleen een hulpmiddel om aan te geven waar welke mate van verbetering gewenst is.

- Bespreek de resultaten met de docent. Dit kan na één les, maar omdat dit niet altijd een compleet beeld geeft, is het raadzaam dit pas te doen nadat alle lessen geobserveerd zijn. Maak dan wel duidelijke notities, zodat je de docent ook nog duidelijke aanwijzingen kunt geven over zaken die tijdens de eerste les zijn opgevallen. Probeer hierbij ook de voorbeelden te betrekken; deze geven vaak een goede schets van wat er zich in de praktijk afspeelt en kunnen duidelijke handvatten zijn voor docenten.
### Algemene informatie

Datum en tijd (lesuur) observatie: ____________________________

School: ________________________________________________

Geobserveerde groep en docent: __________________________________________________________________________________________

Geobserveerd vak en soort les (instructie, zelfstandig werken, toets etc.) _______________________________________________________

Aantal leerlingen: ______________________________________

Naam observant: _________________________________________

Observeer het gedrag van de docent gedurende de les. Geef de items een score van 1 t/m 4 (1= overwegend zwak, 2= meer zwak dan sterk, 3= meer sterk dan zwak, 4= overwegend sterk). De voorbeelden dienen als hulpmiddel om een beter beeld van het bijbehorende item te krijgen en hoeven dus niet te worden gescoord.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Leerdoelen en succescriteria duidelijk maken en delen | 1  | De docent maakt duidelijk wat de leerdoelen zijn op een manier die de leerlingen begrijpen. | 1 2 3 4 | - De docent vertelt wat de leerdoelen voor deze les zijn.  
- De docent vertelt wat de leerdoelen op lange termijn zijn.  
- De docent vertelt de leerdoelen aan het begin van de les.  
- De docent herhaalt de leerdoelen gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de leerdoelen in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent vertelt voor activiteiten tijdens de les wat de leerdoelen zijn.  
- De docent formuleert de leerdoelen in woorden die kennis, vaardigheden, concepten en/of houdingen benadrukken.  
- De leerdoelen zijn gefocust op wat er geleerd wordt.  
- De leerlingen laten zien dat ze de leerdoelen begrijpen door hiermee aan de slag te gaan. |
|                            | 2  | De docent maakt duidelijk wat de succescriteria zijn op een manier die de leerlingen begrijpen | 1 2 3 4 | - De docent vertelt wat de succescriteria voor deze les zijn.  
- De docent vertelt wat de succescriteria op lange termijn zijn.  
- De docent vertelt de succescriteria voorafgaand aan de les.  
- De docent herhaalt de succescriteria gedurende de les of heeft deze voor alle leerlingen zichtbaar in de klas hangen.  
- De docent vertelt de succescriteria in begrijpelijke taal die aansluit bij de leerlingen.  
- De docent geeft voorbeelden van de te verwachten kwaliteit van het werk.  
- De docent behandelt mogelijke misvattingen omtrent de succescriteria.  
- De leerlingen laten zien dat ze de succescriteria begrijpen door hun werk hierop af te stemmen. |
**Master thesis: What do we learn from assessment?**

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Construeer effectieve klassengeprekken en taken die bewijs van leren ontloken | 3  | De docent gebruikt verschillende methoden om uit te vinden wat de voorkennis van de leerlingen is. | 1 2 3 4 | - De docent maakt gebruik van brainstormen.  
- De docent initieert een klasendiscussie.  
- De docent stelt klassikaal vragen om de voorkennis van de leerlingen in kaart te brengen.  
- De docent stelt vragen aan individuele leerlingen om de voorkennis in kaart te brengen.  
- De docent laat leerlingen opdrachten maken om zo de voorkennis in kaart te brengen.  
- De docent gebruikt verschillende methoden gedurende de les om te kijken of leerlingen begrijpen wat er verteld is. |
|           | 4  | De docent gebruikt verschillende strategieën om effectieve vragen te stellen aan de leerlingen. | 1 2 3 4 | - De docent stelt open vragen aan de leerlingen.  
- De docent hanteert een adequate wachtijd met betrekking tot het antwoorden van leerlingen.  
- De docent laat alle leerlingen het antwoord schrijven alvorens een leerling antwoordt.  
- De docent bepaalt op een willekeurige manier welke leerling moet antwoorden.  
- De docent maakt gebruik van antwoordmogelijkheden waarbij meerdere leerlingen tegelijk antwoord kunnen geven.  
- De docent behandelt zowel feitelijke kennis als diepere kennis tijdens klassikale discussies/vragen.  
- Als een leerling de vraag niet begrijpt, stelt de docent de vraag op een andere manier, zodat de leerling het beter begrijpt. |

<table>
<thead>
<tr>
<th>Indicator</th>
<th>#</th>
<th>Item</th>
<th>Score</th>
<th>Voorbeelden</th>
</tr>
</thead>
</table>
| Feedback geven die de leerling verder helpt. | 5  | De docent geeft leerlingen feedback op wat er geleerd wordt. | 1 2 3 4 | - De docent geeft feedback op basis van criteria.  
- De feedback van de docent beschrijft specifieke punten waar verbetering nodig is.  
- De feedback van de docent bevat strategieën om de nodige verbeteringen te maken.  
- Als een leerling een fout antwoord geeft, helpt de docent de leerling om tot het juiste antwoord te komen.  
- De docent geeft leerlingen feedback op hun leerproces.  
- De docent geeft inhoudelijke feedback (meer dan alleen ‘goed’ of ‘fout’) die de leerlingen richting geeft om verder te komen in hun proces.  
- De feedback gaat in op de inhoud van de taak. |
|           | 6  | De docent gebruikt de kennis over de voortgang van de leerlingen om feedback te geven. | 1 2 3 4 | - De docent kijkt tijdens de les waar de leerlingen mee bezig zijn.  
- De docent geeft de leerlingen informatie over hun leerproces.  
- De docent weet wat de sterke en zwakke punten van elke leerling zijn en speelt hierop in.  
- De docent geeft feedback op basis van opdrachten/toetsen die de leerlingen eerder gemaakt hebben.  
- De docent gebruikt de voortgang van de leerlingen om feedback te geven.  
- Als de docent positieve feedback geeft, wordt ook aangegeven hoe de leerling dit niveau kan handhaven. |
**Indicator** | # | Item | Score | Voorbeelden
--- | --- | --- | --- | ---
Leerlingen activeren als eigenaar van hun eigen leerproces | 7 | De docent stimuleert leerlingen hun eigen vooruitgang in de gaten te houden. | 1 2 3 4 | - De leerlingen zijn in staat uit te leggen wat ze op dat moment aan het leren zijn.  
- De docent leert de leerlingen vaardigheden in zelf-monitoren te ontwikkelen.  
- De docent stimuleert de leerlingen hun eigen werk te beoordelen.  
- De docent stimuleert de leerlingen hun eigen voortgang bij te houden.  
- De docent reserveert tijd gedurende de les om leerlingen hun eigen werk te laten nakijken.  
- De docent moedigt de leerlingen aan verschillende vormen te gebruiken om hun eigen werk te beoordelen (bv. rubrics, checklists, nakijkbladen)  
- De docent stelt samen met de leerlingen vast wanneer er aan een succescriterium voldaan wordt.

8 | De docent vraagt feedback op het leergeven aan de leerlingen. | 1 2 3 4 | - De docent vraagt of zijn/haar uitleg duidelijk is.  
- De docent vraagt leerlingen aan te geven wat er anders kan m.b.t. de instructie.  
- De docent vraagt leerlingen of ze uitgedaagd worden in de les.

**Indicator** | # | Item | Score | Voorbeelden
--- | --- | --- | --- | ---
Leerlingen activeren als instructiebronnen voor andere leerlingen. | 9 | De docent stimuleert leerlingen het werk van andere leerlingen te beoordelen en feedback te geven aan andere leerlingen. | 1 2 3 4 | - De docent reserveert tijd om gedurende de les het werk van medeleerlingen te beoordelen.  
- De docent stimuleert de leerlingen om vragen aan elkaar te stellen alvorens ze naar de docent stappen.  
- De docent stimuleert het werken in groepen wanneer een opdracht individueel te lastig is.  
- De docent gebruikt geanonimiseerde voorbeelden van leerlingen om veelgemaakte fouten aan de leerlingen te laten zien.  
- De docent stimuleert leerlingen om vragen aan elkaar te stellen gedurende de les.

Overige opmerkingen:
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________
____________________________________________________________________________________________________________________