Windesheim

UNIVERSITY OF APPLIED SCIENCES

Evaluation of the learning path independent assessments of Windesheim used in the educational domain of Health and Well-being in part-time education

Name: Kimberly de Jonge – s1443445
Email: k.dejonge@student.utwente.nl

First Supervisor: J.W. Luyten
Email: j.w.luyten@utwente.nl
Department: OMD

Second Supervisor: D. den Otter
Email: d.denotter@utwente.nl
Department: OMD

External organization: Hogeschool Windesheim
Supervisor: Jose Uitdewilligen
Email: JJM.Uitdewilligen@windesheim.nl

Keywords:
Lifelong learning, Competency assessments, Evaluation, Higher Education.

Summary

To assure everybody can participate in society, and to give everybody a chance to work in a fast-changing labour market, the Dutch government has implemented Lifelong Learning. Due to this educational shift, more flexible education is needed. That is why the Dutch government started the pilot flexibilization in higher education. Windesheim participates in this pilot.

Due to this educational change, the way of assessing needs to change too. The assessments need to become learning path independent. This means that it does not matter where when and how the learning outcomes are achieved to pass the assessment. These assessments need to be evaluated to come up with recommendations for making new assessments in the future and adjusting the assessments currently used.

To evaluate the new assessments used by Windesheim in the field of Health and Well-being the self-evaluation procedure of Baartman, Bastiaens, Kirschner, & van der Vleuten (2006) will be used, which uses a mixed methods approach. This self-evaluation procedure uses the following criteria: acceptability, authenticity, cognitive complexity, comparability, costs & efficiency, educational consequences, fairness, fitness for purpose, fitness for self-assessment, meaningfulness, reproducibility of decisions and transparency.

First a student and teacher questionnaire which is based on this framework will be used with a 5-point Likert scale. Second the outcomes of this questionnaire will be discussed in focus groups, one with teachers and one with students, which will result in strong and weak points per criterion. Based on these results, recommendations will be made.

Content

1. Introduction

1.1 Problem Statement

In the Netherlands Lifelong learning is implemented by the Dutch government to assure everybody can participate in society, and to give everybody a chance to work in a fast-changing labour market. To accomplish lifelong learning, more flexibility in higher education has become a topic of interest. More flexible education can make education more accessible and attractive, and better aligned to the characteristics and the needs of students (OCW, 2016), which will improve lifelong learning. In the pilot flexibilization, the schools for higher education get more flexibility in determining the learning goals from the professional profiles developed by the Dutch government (Rijksoverheid, n.d.). Windesheim uses Learning outcomes to accomplish the learning goals developed by the government(Olthof, Stulen, & Mossel, 2017). In the pilot the learning outcomes don't have to be time bound anymore but need to fit on the needed competencies of the students.

Because of the change in more flexible education the assessments need to be more flexible too. To accomplish this, new assessments are made. The assessments used are learning path independent, which means that they can be made independent from the education provided by Windesheim in the domain of Health and Well-being (Olthof et al., 2017; van Berkel, 2017). It does not matter where when and how the learning outcomes are achieved to pass the assessment. These assessments need to validate the needed knowledge and skills to accomplish the learning outcomes (Windesheim, 2017). Box 1 contains more information about the assessments of Windesheim. When all the learning outcomes are reached it will result in a diploma. To get accreditation for the educational program the assessments need to be of good quality (Nederlands Vlaamse Accreditatieorganisatie, 2018).

According to the literature reliability and validity are one of the most important quality criteria. Quality of assessments is important, and validity, intended use, is one of the most important criterium for assessment quality (Wools, 2012). According to a literature review about assessment quality, reliability is mentioned most frequent in articles about assessment quality (Maassen et al., 2014). According to Baartman, Gulikers, & Dijkstra (2013), the development, implementation and evaluation

Box 1. Assessments of Windesheim.

To pass a module of 30 EC, students can achieve the learning outcome by choosing to make a portfolio or the learning path independent assessments. When the students choose to do the assessments, they have to make 2 or 3 assessments per module. These assessments focus on the attitude, skills and knowledge of the students, which they learned during the education on Windesheim, the online education provided and/or the knowledge and skills of the people at the work place of the students. These assessments take place in the work context of the students and have to be made using a clear format. This format is described in the assessment form the students need to follow. Students can also validate their knowledge and skills using a portfolio, which has a less strict format but is graded using the same assessment criteria as used by the assessments. These criteria are based on the learning outcome of the module. For more information about the learning path independent assessments Windesheim uses in the domain of Health and well-being see: Handleiding Leerwegonafhankelijk toetsen (Windesheim, 2017) and Toolkit Flexibel hoger onderwijs voor volwassenen (Olthof et al., 2017).

of such assessments are not straightforward and require careful and critical consideration of the current assessment practices. To do this, a self-evaluation procedure has been developed by Baartman (2006) to evaluate the assessments used, to come up with improvements or to develop new assessments(Baartman, Bastiaens, Kirschner, & van der Vleuten, 2006; Baartman, Prins, Kirschner, & van der Vleuten, 2011; Dijkstra & Baartman, 2011).

To evaluate the current assessments of Windesheim in the domain of health and well-being, the self-evaluation procedure of Baartman (2013) can be used, because the Competency Assessment Programmes (CAP) fit the learning path independent assessments of Windesheim. First, questionnaires can be used to get an overview of the quality. Second, focus groups can be used to come up with recommendations for improvement. The criteria that measure the quality according to Baartman (2013) are acceptability, authenticity, cognitive complexity, comparability, costs & efficiency, educational consequences, fairness, fitness for purpose, fitness for self-assessment, meaningfulness, reproducibility of decisions and transparency (Baartman et al., 2006). By using the self-assessment procedure to assess the assessments used, the assessment quality can be improved by Windesheim without using an external party.

This research is set out to evaluate the quality of the learning path independent assessments of Windesheim. It is aimed to do so by using questionnaires to gather quantitative data about the quality and qualitative data by using focus groups with students and teachers who are involved in the learning path independent assessments. The focus group data is additional to the questionnaire data, the qualitative data will justify the quantitative data found and will make the weak and strong points more specific. It will result in recommendations to increase the quality of the learning path independent assessments of Windesheim. The research will take place in the period of April till October 2018.

1.2 Theoretical conceptual framework

1.2.1. Evaluation of assessments

The literature review of Gerritsen-van Leeuwenkamp, Joosten-ten Brinke, and Kester (2017) found that staff and students are the biggest perspectives in evaluating the assessment quality in tertiary education. According to the study of Gulikers, Biemans, and Mulder (2009), who studied the differences in experience between students, employees and developers of an assessment, developers and teachers are more critical about the quality of the assessments then the employees and students are.

Various researchers have focussed on evaluating the quality of competency assessments (Aea, 2011; Epstein & Hundert, 2002; J. T. M. Gulikers, Baartman, & Biemans, 2010; McMullan et al., 2003).

Education assessments are used to make high-stake decisions about learners. Developing adequate assessment methods is important because a strong relationship exists between the learning of students and the assessment of students (Baartman et al., 2006). According to Van Der Vleuten & Schuwirth (2005) reliability, validity and educational impact contribute to the quality of the assessments. Reliability is defined here as the reproducibility of the scores obtained from assessments. Validity is defined here as to whether an instrument actually does measure what it is supposed to do. The educational impact can be seen as a part of the validity and is defined as the impact of assessment on learning (Van Der Vleuten & Schuwirth, 2005).

According to Gerritsen-van Leeuwenkamp et al. (2017), the quality of assessments can be determined by the reliability, validity and transparency of the assessment. Where transparency refers to that the assessments need to be clear for the stakeholders, for example that students need to know what is expected from them, what the criteria are and whether the assessment counts for their diploma (Baartman, Prins, Kirschner, & van der Vleuten, 2007b, 2007a; Gerritsen-van Leeuwenkamp et al., 2017).

1.2.2 Evaluating competency assessments

Because researchers state that the term validity is confusing and the term reproducibility need to be defined differently for competency assessments and standard tests, Baartman (2006) developed a new framework (Baartman et al., 2006) which focusses on quality criteria for competence assessment programs. This framework has been developed first through a literature review to come up with criteria. Second, a focus group is carried out with international professionals in the field of assessment to come up with quality criteria. Finally, an adapted and improved framework is presented with the results of the focus group. In this framework, validity is a container concept and is measured by almost all the criteria used. The criteria: reproducibility of decisions and comparability, match with reliability.

1.2.3. Self-evaluation procedure

This framework can be used to conduct a self-evaluation of the competency assessment used in current education. The self-evaluation procedure consists of a quantitative part, were a questionnaire is used to measure the ratings on the 12 criteria, and a qualitative part, were group interviews are used to support the ratings on the questionnaire (Baartman et al., 2007a). The results of Baartman (2007) showed that the group interview used was very important because the different perspectives on the competency assessment programs were assembled into an overall picture of the assessment's quality. According to (Baartman et al., 2007b) the group interview seemed to be important to come up with evidence for the score of a criterion and it also served to correct misunderstandings of group members. The group interview led also to spontaneous ideas to improve the competency assessment program (Baartman et al., 2007b).

In the study of Dijkstra & Baartman (2011) of the academies in the Avans Hogeschool, the self-evaluation procedure turned out to have a positive effect on the evaluation and improvement of assessments. Besides that the self-evaluation procedure gave them guidelines to improve the study program as a whole, they indicate that the self-evaluation prepared them for their accreditation(Dijkstra & Baartman, 2011).

In the study of Baartman et al., (2013), the results show that in the quantitative part of the self-evaluation procedure the overall high scores on (almost) all indicators were found for fitness for purpose, fairness and accountability. Rather good scores were found for transparency, acceptability, costs and efficiency, authenticity and complexity. The low scores were found for reproducibility of decisions and development of self-regulated learning. In the qualitative part of the self-assessment procedure strong and weak points were given to all the criteria. A lot of points for improvement are mentioned to improve the internal quality of the assessments used (Baartman et al., 2013).

The goal of the research of Baartman et al., (2011) was to contribute to the validation of the self-evaluation method that was developed by Baartman, Prins, Kirschner, & van der Vleuten (2007a). This research concluded that the method of the self- evaluation procedure seemed to support the validity of the self-evaluation program to a great extent. According to

Baartman (2013) participants found some questions of the self-evaluation procedure difficult to answer, not all questions where being answered by student. In particular these questions where about teacher opinions. As recommendation, two separate questionnaires were suggested, one for students and one for teachers. Also, Baartman 2013 indicates that some questions of the procedure could be improved to increase understanding of the participants.

This framework is already used in the Netherlands for self-evaluation of education programs (Baartman et al., 2007b; Dijkstra & Baartman, 2011). The framework is described in Figure 1. Figure 2 explains the 12 criteria used in the framework of Baartman (2007).

In the centre of the framework the criterion fitness for purpose is stated, which prescribes that all the assessments used must be aligned with the learning goals(Baartman et al., 2006). The next inner layer consists of the criteria: comparability, reproducibility of decisions, acceptability and transparency. These criteria are the most basic ones which are already used in practice for evaluating assessments. The outer layer consists of the criteria: fairness, authenticity, cognitive complexity, meaningfulness and fitness for self-assessment. These criteria are generally newer, than the ones in the inner circle, in the assessment culture. It tends to be that the criteria in the inner layer are prerequisite for the criteria in the outer layer. The criteria costs & efficiency and educational consequences are outside the wheel, because these represents the broader educational space in which the assessments take place (Baartman et al., 2006).

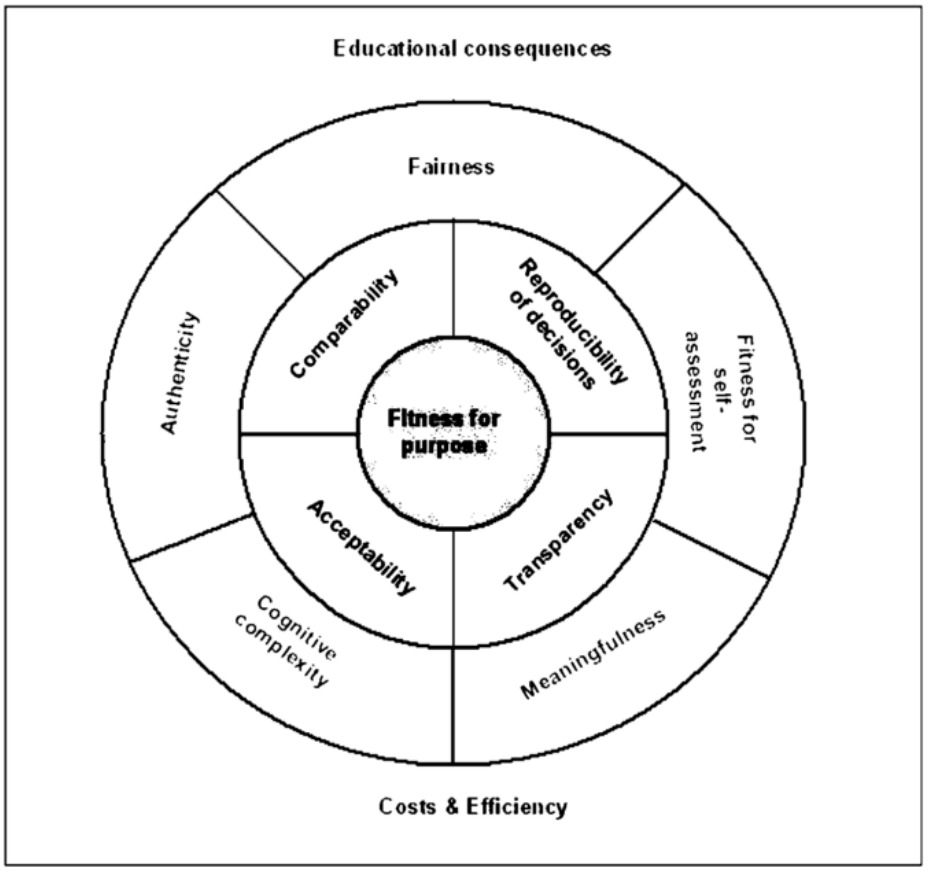

Figure 1. The wheel of competency assessment (L. K. J. Baartman et al., 2006).

Table 1. Quality criteria for CAPs (based on Baartman, Prins et al. 2007).

Criterion	Short description
Fitness for purpose	Alignment between curriculum goals and what and how is assessed. Criteria and standards should address all competences and the mix of methods should be fit to assess competence (Brown 2004; Miller and Linn 2000)
Cognitive complexity	CAPs should enable the judgment of thinking process, besides assessing the product or outcome (Maclellan 2004)
Self-assessment	CAPs should stimulate self-regulated learning, for example by using self-assessments, and letting students formulate their own learning goals (Tillema, Kessels, and Meijers 2000)
Authenticity	The degree of resemblance of a CAP to the future workplace (Gulikers, Bastiaens, and Kirschner 2004)
Transparency	CAP should be clear and understandable for all stakeholders (Frederiksen and Collins 1989; Linn, Baker and Dunbar 1991)
Comparability	Assessment tasks, criteria, working conditions and procedures should be consistent with respect to key features of interest (Baartman, Bastiaens et al. 2007)
Reproducibility of decisions	Decisions about students should be based on multiple assessors, multiple tasks and multiple situations (Moss 1994; van der Vleuten and Schuwirth 2005)
Fairness	Students should get a fair chance to demonstrate their competences, for example by letting them express themselves in different ways and making sure the assessors do not show biases (Dierick and Dochy 2001; Hambleton 1996; Linn, Baker and Dunbar 1991)
Acceptability	All stakeholders should approve of the assessment criteria and methods (Stokking et al. 2004)
Meaningfulness	CAPs should be learning opportunities in themselves and generate useful feedback for all stakeholders (Linn, Baker and Dunbar 1991)
Educational consequences	The degree to which the CAP yields positive effects on learning and teaching (Messick 1994; Schuwirth and van der Vleuten 2004)
Costs and efficiency	The feasibility of carrying out the CAP for assessors and students (Hambleton 1996; Linn, Baker and Dunbar 1991)

Figure 2 Quality criteria for CAP's (Baartman et al., 2013; Baartman, Prins, Kirschner, & van der Vleuten, 2007)

1.3 Research question and model

The goal of this thesis is to come up with a recommendation for Windesheim to improve the quality of their learning path independent assessments in the education in the field of health and well-being. This will be done using the self-evaluation procedure of Baartman, Prins, Kirschner, & van der Vleuten (2007a). The framework fits current research because it focusses on already designed assessments, it is made in the Dutch context, and it is already used in several studies in (higher) educational programs (Baartman et al., 2013, 2007b, 2007a, 2011; Dijkstra & Baartman, 2011), it does not only evaluate the assessments but also to evaluate the entire range of assessments set within the curriculum, and it can be used by all stakeholders involved. First, quantitative data will be collected using a student and teacher questionnaire. Second, qualitative data will be collected to underpin the outcomes of the questionnaire, and to come up with strong and weak points.

The research question of this thesis is: How do the teachers and students evaluate the learning path independent assessments, and how can improvements be made?

1 How do the students evaluate the learning path independent assessments of Windesheim?
2 How do the teachers evaluate the learning path independent assessments of Windesheim?

3 Is there a difference in the evaluation of students and teachers of the leaning path independent assessments of Windesheim?
4 How can the learning path independent assessments of Windesheim be improved using the evaluation of students and teachers?

1.4 Scientific & Practical Relevance

According to Maassen et al. (2014) a lot of research is done about the evaluation of assessments, according to them more specific assessments in higher education could be researched to accomplish good assessment quality.

According to Baartman (2013), the self-evaluation procedure in general led to many points for improvement. These points for improvement can be useful for internal quality improvement of the learning path independent assessments of Windesheim.
Also, the self-evaluation procedure can be helpful in conducting new assessments in the future. The flexibilization of higher education will be improved if the assessments are also flexible and of good quality for the students and the teachers.

Also, it can help with the accreditation procedure of Windesheim. To give an accreditation to the educational program, the assignments used have to be of good quality. When the NVAO gives accreditation for an educational program, it guarantees the quality of the higher education and the value of the diploma given. (Nederlands Vlaamse Accreditatieorganisatie, 2018).

The self-evaluation procedure will be adjusted according to recommendations given in research by Baartman (2013). Instead of one questionnaire that is used by Baartman et al. (2006), two different questionnaires will be used in this research, one for teachers and one for students, to gather data. This can be helpful in future research to validate the self-evaluation procedure. By using the self-evaluation procedure in the pilot flexibilization, it can be tested if the self-evaluation procedure fits that specific context. Also, more knowledge will be generated about the quality of assessments in flexible education.
This research will make a contribution to literature by comparing existing literature about the self-evaluation procedure and the outcomes of this research.

2. Research Design and Methods

2.1 Research Design

A mixed methods approach was done to come up with an answer on the research question. First, quantitative research was done using a questionnaire which is based on the Wheel of Competency assessment, Figure 1 and 2. Also, qualitative additional data was gathered using focus groups with teachers and students. These focus groups were also based on the framework and research of Baartman et al. (2006).

2.2 Respondents

2.2.1 Questionnaire

All 75 students who already did a learning path independent assessment were approached during their education by their teacher and were asked to fill in the questionnaire on paper or online. All these students were involved in the part-time education in the field of Health and Well-being. In the end, 43 students filled in the questionnaire. These students had an average age of 34.4 years and had already on average 13.1 years of work experience in their work field. More woman (n=34) than man (n=9) had filled in the questionnaire, what is representative of the student population. Of these students 12 do social work, 21 do nursing and 9 do the associate degree of social work. 28 of the students did module 1, 39 of the students did module 2 and 9 of the students who filled in the questionnaire did module 3.

The teachers who conducted and graded the learning path independent assessments were also asked to fill in a questionnaire. They were approached using e-mail and providing paper and pencil questionnaire in their offices. The mail addresses were provided by Windesheim. A total of 20 teachers filled in the questionnaire. From those 3 where male and 17 were female. Their average age was 49,65(SD = 9,34), their work experience in education was on average 12,72 years (SD = 9,62) and their work experience on Windesheim was on average 9,44 years (SD = 7,72). 10 of the teachers were involved with module 1, 14 with module 2 and 10 with module 3.

This choice to ask students and teachers was made, because the literature review of Gerritsen-van Leeuwenkamp, Joosten-ten Brinke, & Kester (2017), which describes that for evaluating assessments, the staff and the students have the biggest perspectives, and because Baartman (2013) would like to use different questionnaires for teachers and students.

2.2.2 Focus Groups

To gather additional data, focus groups were held separate with teachers and students. To make sure the respondents are comfortable during the focus group. The respondents were asked at the end of the online questionnaire if they were willing to participate in a focus group about the learning path independent assessments. These respondents were contacted with a date and time at which the focus group took place.

The focus group with teachers was with only 2 teachers, of which 1 only gives lessons and grades assessments and the other is a chairman of a module and leads a learning team besides giving lessons and making assessments.

The focus group with students was with 2 students in total. One studied social work and the other nursing, both did 2 modules last year.

Due to the end of the school year and work load of the teachers and the students, the focus groups had a low attendance.

2.3 Instrumentation

2.3.1 Questionnaire

The instrument used was an existing questionnaire using the 12 criteria from the wheel of competency assessment. The criteria were operationalized in the different article of Baartman (Baartman et al., 2013, 2007a, 2007b, 2011). The original questionnaire of Baartman is used with some adjustments to fit into the context of the learning path independent assessments. The criteria and the operationalization are described in Appendix 1, together with the distribution of questions between the students and the teachers. These distribution choices were made by the researcher. The student questionnaire is shown in Appendix 2 and the teacher questionnaire in Appendix 3. The measurement level of the variables was ordinal using a 5-point Likert-scale, from totally disagree to totally agree.

The questionnaire was first pilot tested by teachers and students, to make it appropriate for the learning context of Windesheim.

2.3.2 Focus Group

To structure the focus groups a protocol, informed consent (Appendix 3) and a topic list were used. The protocol contained information about the content and the format of the focus group. The informed consent had to be signed by the researcher and interviewee before the start of the interview, it contained information about the anonymity of the taped data and that the data will not be distributed by the researcher, also the opt-out option was included for the respondent which stated that the respondents could leave the focus group when they wanted.

During the focus group the criteria and questions asked in the questionnaire were used to structure the focus group. These are captured in the topic list. The outcomes of the questionnaire were provided, and the respondents could react to them with underpinnings that could explain the results. Also points of improvement could be discussed during the focus group.

2.4 Procedure

2.4.1 Questionnaire

A pilot tested questionnaire was used. To assure the quality of the questionnaire, an already existing questionnaire was used, which is used in several studies in the Dutch (higher) education context and was already proven valid in literature. To assure the anonymity of the data, the personal data of the respondents and the answers on the questionnaire were saved in a different document on a different server. The data was for research only and was not distributed to other parties involved.

The students were asked to fill in the questionnaire during their lessons, in their learning teams. The learning teams, are lessons in which students can ask questions to students within their learning team, or the teacher who leads the learning team. It can be compared with a Mentor Class. The questionnaire was provided on paper. When a visit during the learning teams was not possible, the link to the online questionnaire was mailed by the teacher of the learning team to the students in their learning team.

The teachers got a copy of the teacher questionnaire on their desk to fill it in, also a link was send to them to fill in the questionnaire. It took about 5-10 minutes to fill in the questionnaire.

2.4.2 Focus group

After the questionnaire, teachers and students were asked if they want to give additional information about their experience with the learning path independent assessments in a focus group. The topic list and protocol were mailed as information about the proceedings of the focus group. At the start of the focus group the informed consent form was to be signed by the respondent and the researcher. The focus group was taped and transcribed.

When the outcomes about the quality of the learning path independent assessments of Windesheim were clear, recommendations were made, and the outcomes were presented to the teachers and the management of the Life Long Learning team.
The focus groups took about 90 minutes.

2.4.3 Ethical Considerations

The ethical committee of the university of Twente approved the approach of the questionnaires and interviews used in this research by the code 18510.

2.5 Data analysis

2.5.1 Questionnaire

The quantitative data from the questionnaires were analyses by SPSS statistics and Excel. The Cronbach's alpha was calculated for every criterion. When the Cronbach's alpha was above 0.6 the criterion is used as a whole, when the Cronbach's Alpha is below 0.6 the means of the questionnaire items were used in further calculations.

When a score is above 3.5 the score was market as sufficient. Under 3.5 the score was marked as insufficient. This margin was set by the researcher. The answers of the teachers and students were normally distributed. To check if there are differences between the opinion of both groups about the assessments, t-test were conducted for every question.

2.5.2 Focus groups

The additional focus groups are transcribed using Microsoft Word and coded using Atlas.ti. The transcribes have been read and coded by the different criteria used in the interview. Second, the strong and weak points of the criteria were coded, which fits the research of Baartman et al. (2006).

These outcomes were summarized to come up with point for improvement for internal quality improvement. Recommendations will be based on these points for improvement.

3. Results

This chapter contains the results of the questionnaires and the focus groups. First the result of the student questionnaire and focus groups will be described in paragraph 3.1. Paragraph 3.2 contains the results of the teacher questionnaire and the teacher focus group. At last, these results will be compared with each other in paragraph 3.3.

3.1 Students

This paragraph contains the results that relate to the research question: How do students evaluate the learning path independent assessments of Windesheim? First the results of the quantitative data from the questionnaire will be described. Second the additional qualitative data will be described to provide a full evaluation of the learning path independent assessments used by Windesheim from the students.

3.1.1 Questionnaire

The results from the questionnaire are described in Table 1.

Table 1
Results Student questionnaire.

Criterion	Mean	St.Dev	Cronbach's alpha	Question	Mean	St.Dev
Acceptability	3.61	0.643	0.781	1.1	3.70	0.803
				1.2	*3.47*	0.909
				1.4	3.59	0.785
				1.5	3.67	0.892
Authenticity	3.64	0.843	*0.550*	2.1	3.56	1.076
				2.4	3.74	0.939
Cognitive complexity	3.93	0.597	0.742	3.1	3.72	0.797
				3.2	4.37	0.618
				3.3	3.83	0.803
				3.4	3.74	0.875
Comparability	*3.17*	0.557	*0.512*	4.1	*3.31*	0.924
				4.2	*2.68*	0.850
				4.3	*3.44*	0.776
				4.4	*3.32*	0.879
Educational consequences	*3.48*	0.752	0.706	6.1	3.51	0.798
				6.2	*3.47*	0.909
Fairness	3.66	0.783	0.787	7.3	3.53	0.987
				7.4	3.86	0.710
				7.5	3.60	1.027
Fitness for Purpose	3.69	0.522	0.652	8.1	3.76	0.692
				8.2	3.98	0.423
				8.3	*3.41*	0.894
				8.4	3.51	0.960
				8.5	3.81	0.852
Self-assessment	*3.47*	0.652	0.703	9.1	*2.51*	0.952
				9.2	*3.44*	0.983
				9.3	3.84	0.871
				9.4	3.56	0.934
				9c	4.05	0.999
Meaningfulness	3.71	0.622	0.779	10.1	3.77	0.895
				10.2	3.67	0.837
				10.3	3.93	0.856
				10.4	3.83	0.794

				10.5	*3.33*	0.898
Reproducibility of decisions	*3.04*	0.726	0.815	11.2	*3.07*	0.894
				11.3	*3.00*	0.795
				11.5	*3.19*	0.862
Transparency	*3.43*	0.609	0.729	12.1	*3.35*	0.923
				12.2	3.51	0.910
				12.3	*3.23*	0.996
				12.5	*3.29*	0.891
				12.6	*3.43*	0.929

Note: Italic: scores are below the 3.5 benchmark or Cronbach's alpha scores below the 0.6 benchmark.

The Cronbach's alphas are calculated for every criterion. As seen in Table 1he Cronbach alphas are insufficient (<.6) for the criteria: authenticity and comparability. This means that the mean score for that criterion cannot be taken into account, so the scores per question are used for those criteria. The low score for authenticity can be explained by the fact that it was only measured by two questions.

According to the outcomes of the student questionnaire, some criteria score insufficient. These are comparability, educational consequences, self-assessment, reproducibility of decisions and transparency. Whereby educational consequences, self-assessment and transparency score just below the insufficient limit of 3.5.

3.1.2 Focus group

During the focus group the results of the questionnaire were discussed and explained by the respondents. These results are stated per criterion below. A more detailed overview of the results of the focus group with the strong and weak point per criterion is described in Appendix 5.

Acceptability

The score of accessibility of the student questionnaire was sufficient and has a high Cronbach's alpha. According to the students the assessments of module 2 had a good connection with the work context the students experienced. Also, the assessments of module 2 had a format in which the students could work easily.

As weak points, the students mention that the assessment and the rubric should be better aligned to make the assessment more useful for the students and their work context. Module 3 is less aligned with the work context of students and did not included a clear format.

Authenticity

Authenticity scores sufficient according to the student questionnaire. As a weak point was mentioned by the students that assessments are described in a generic context. Which makes it hard for them to do an assessment when a work context does not fit the assessment criteria properly.

Strong points according to the students are that the assessments and the criteria fit the work context of the students and are feasible in the context of a social worker or nurse. Also, as a student you can gain knowledge about all sides of a problem or context in a work context used for an assessment.

Cognitive complexity

According to the student questionnaire the cognitive complexity of the assessments scores high, *M* = 3,93. This can be explained according to the students by the fact that they are triggered to make thinking steps when they make the assessments. The assessment gives those thinking steps already in the explanation of the assessment by describing what the student has to do in which order. These thinking steps are already implemented unconsciously by the students at their workplace. Also, the assessment criteria provide information about the support and underpinning a student needs to give when it is needed. This can be when a student needs to make a choice and have to support this choice.

Weak points concerning the cognitive complexity are that when an assessment does not fit the work context, the thinking steps are hard to implement during daily practise. Also, the students do not implement the thinking steps consciously in their workplace.

Comparability

Comparability scores insufficient according to the student questionnaire. When looked at the separate questions, all the questions score individually insufficient. The following explanation is given by the students during the focus group.

All students have a different work context which makes it difficult to customize the assessments to make them fit the work context, which makes students not comparable. Also, not all the students have the same opportunities in their work situation which can make things like filming difficult for students. Also, assessments of different students are not comparable due to the different work contexts. The assessment procedure is not clear to the students, in particular the part what makes an assignment sufficient or insufficient. Students also mention that assessors are prejudiced when they assess multiple assessments of them. Finally, assessors give different kinds of feedback to students on the assessment form, some are very short and others more specific.

Some strong points mentioned are that the assessments are the same for all students. For module 2 the assessments of all students are comparable because everyone used the same format in their assessment. According to the students, all assessors are open towards giving more or more specific feedback when a student asks.

Educational consequences

Educational consequences scores insufficient according to students, but scores close to sufficient (*M* = 3.48). The insufficient score can be explained by the fact that students sometimes do an assignment just for getting the points to get their diploma instead of evoking a learning process for the students. This can be caused by the fact that assignments do not always fit the work context of students properly, which makes students less motivated to do the assignment. This way the students only look at the assessment criteria to pass their assessment instead of looking at the assignment itself, which makes that they are not positive affected by the assessments.

This was not the case for the assessments of module 2, according to students these where really nice to do and it resulted in a desired educational process for them.

Fairness

According to the student questionnaire, fairness scored sufficient (*M* = 3,66). Students do not think that the teachers are prejudiced. Also, the assessments fit the professional code of the profession they are learning.

Students do think that there is a difference in how teachers mark an assessment and the way they provide feedback on an assessment. A student also experienced that a teacher marks multiple assessment from him/her, and that those assignments where kept in mind while marking another assessment. Also, students think that the weight of the assessments do not fit the working load of the assessments made.

Fitness for purpose

Students rate the fitness for purpose criterion sufficient, but when looked at the individual questions the question about the use of different forms of assessment scored low. This can be explained by the fact that only 2 forms of assessments are currently used to accomplish a learning outcome, these are an assessment or a portfolio.

Students mention that there are no moments during the educational program to get formative feedback. Also, the lessons sometimes do not fit the assessment that has to be made. When formative feedback can be given during the learning teams, it is difficult to manage, because everybody in the learning team does different modules, so different assessments. Also, assessments from the same module can differ between students because of different work contexts of the students in which the assessments are made, which makes giving feedback also hard.

Students do indicate that the assessments do fit with the goal of the educational program by measuring attitude and behaviour at once. Besides this, students indicate that the learning team is open towards giving feedback, but it is hard to manage.

Self-assessment

The self-assessment criterion scores insufficient, but close to sufficient *(M* = 3.47) according to the student questionnaire. This insufficient mark can be caused by the fact that the educational program has not built in moments for peer feedback. Also, students indicate that the assessments are strict in the way they have to look like, which makes it sometimes hard to come to the learning outcome the way a student wants.

The strong points mentioned by the students are that students can choose themselves what they use to make the assessment, they can use the online modules, the lessons given at Windesheim or their work context. Students can even choose how to use their learning team, in which specific feedback can be asked and given or support can be given concerning the assessments. The feedback of professionals in their workplace focusses more on the practical part of the assessments but is meaningful for the students. Students indicate that the assessments support their own professional development.

Meaningfulness

According to the student questionnaire, the meaningfulness criterion scores sufficient. This can be explained by that the assessments are made in their workplace. Also, the professionals of the workplace of the students think these assessments and the assessment criteria are meaningful for the students.

Students mention that there is a big difference in education between the fulltime students who do an internship at their workplace, and the dual education students, which is hard for the workplace. Also, according to students, the APA-guidelines are not clear.

Reproducibility of decisions

The criterion reproducibility of decisions has the lowest score of the criteria used in the student questionnaire (*M* = 3,04). This can be partly explained by the fact that students are unsure if there is always a second assessor asked in case of an insufficient mark (which is mandatory). Also, students indicate that they are not sure if they get the same mark when they hand in their assignment another time. At last students think that there is a difference in who of the assessors assesses an assignment, because students think some assessors are stricter than others.

On the other hand, students know that there are calibration sessions among assessors to increase the reproducibility of decisions. Also, students indicate that assessors can't be prejudiced because it is unsure who assesses an assignment from them.

Transparency

The last criterion in the student questionnaire was transparency, which also scored insufficient (*M* = 3.43). Students indicate that they know the assessment criteria of the EVL according to the questionnaire.

Weak points mentioned by the students are that not all the modules are developed yet, which makes the educational program not transparent. Second, teacher and students are still looking for ways to do the lessons and the assessments because it is a whole new educational program which started in 2017. Third, students indicate that the background information is not clear, which makes it hard for them to understand the educational processes involved in the educational program. And at last, students indicate that it is for them not clear what some assessments need to look like.

3.2 Teachers

This paragraph contains the results of the evaluation of the learning path independent assessments used by Windesheim in the educational field of health and well-being by the teachers to answer the following research question: How do the teachers evaluate the learning path independent assessments of Windesheim?

First, the results of the teacher questionnaire are described in Table 2. Second, the results of the questionnaire and focus group are combined per criterion to get an overview of the evaluation of the learning path independent assessments used.

3.1.1 Questionnaire

The results from the questionnaire are described in Table 2.

Table 2
Results Teacher questionnaire.

Criterion	Mean (if item removed)	St.Dev	Cronbach's alpha (item removed)	Question	Mean	St.dev
Acceptability	3.78	0.672	0.871	1.1	3.60	0.883
				1.2	3.65	0.933
				1.3	3.70	0.979
				1.4	3.82	0.809
				1.5	3.85	0.875
Authenticity	4.18	0.654	0.676	2.1	4.50	0.513
				2.4	3.85	0.933
Cognitive complexity	3.78	0.656	0.737	3.1	3.65	0.875
				3.2	4.32	0.478
				3.3	*3.47*	1.020
				3.4	3.75	1.020
Comparability	3.78 (3,75) *	0.317	0.665 (4.3) *	4.1	3.95	0.705
				4.2	*3.06*	0.899
				4.3	4.15	0.489
				4.4	4.16	0.375
Costs				5a	*3.39*	1.092
Educational consequences	3.66	0.765	0.892	6.1	*3.50*	0.946
				6.2	3.84	0.765
				6.3	3.72	1.018
				6.4	3.94	0.772
				6.5	4.00	0.767
Fairness	3.51	0.489	*0.574*	7.1	3.61	0.698
				7.3	*3.42*	0.769
				7.4	3.85	0.813
				7.5	*3.26*	0.653
				7.b	3.67	0.686
Fitness for Purpose	3.76 (3,98) *	0.435	0.847(8.4) *	8.1	4.05	0.394
				8.2	4.11	0.875
				8.3	3.79	0.787
				8.4	*2.84*	0.765
				8.5	4.00	0.562
Self-assessment	*3.46*	0.542	0.749	9.1	*3.06*	0.938
				9.2	3.53	0.697
				9.3	3.89	0.567
				9.4	*3.05*	0.911
				9c	3.72	0.575
Meaningfulness	3.56	0.451	0.725(10.1) *	10.1	3.71	0.849

	(3,58) *			10.2	3.76	0.664
				10.3	*3.20*	0.768
				10.4	3.55	0.826
				10.5	3.85	0.813
Reproducibility of decisions	*3.39*	0.515	*0.482*	11.2	*3.11*	0.758
				11.3	*3.32*	0.749
				11.4	3.61	0.979
				11.5	3.84	0.834
				11.6	*3.28*	1.074
Transparency	3.61	0.571	0.806	12.1	3.60	0.940
				12.2	*3.50*	0.827
				12.3	*3.50*	0.827
				12.4	4.10	0.718
				12.5	*3.50*	0.688
				12.6	3.95	0.510

Note: Italic: scores are below the 3.5 benchmark or Cronbach's alpha scores below the 0.6 benchmark.
** if item removed to increase the Chronbach's Alpha*

3.2.2 Focus group

As by the student focus group, the scores on the questionnaire were used as input for the focus group. The results are reported per criterion below. A more detailed description of the strong and weak points per criterion is given in Appendix 6.

Acceptability

According to the teacher survey, the criterion acceptability scores sufficient. Teachers explain that they experience trust in the assessments used. Also, teachers have the idea they can work with it.

Teachers experience from the students that they indicate that the assessment criteria are not very clear, not guiding and that they are formulated in an abstract way. The teachers explain that the assessments and assessment criteria are all relative new, and minor adjustments need to be made in the assessments and the assessment criteria used.

Authenticity

Authenticity scores relatively high on the teacher questionnaire (*M* = 4.18). This is because all the assessments students need to do are done in the work context of the students.

Teachers indicate that it is hard sometimes to adjust the assessments to the work situation of specific students. This takes time and effort of the students to accomplish.

Cognitive complexity

The cognitive complexity criterion scores sufficient according to the teacher questionnaire (*M* = 3.78). This sufficient score is explained by the fact that the assessments call up the necessary thinking steps needed.

On the other hand, teachers mention that the assessments are hard to do for a beginning professional. Second the assessment criteria and the assessment do not always match with each other which can make it difficult for students to make the assessment.

Comparability

According to the teacher questionnaire, the comparability criterion scores sufficient (*M* = 3.78 and (*M* = 3,75 if 4.3 is removed). Teachers indicate that when students deviate from the

assessments, they can still get a sufficient grade when they underpin the deviation well. Also, the assessments, the assessment criteria and the assessment procedure are comparable for all students.

Weak points mentioned by the teachers are first that the work situation of students can differ, which makes some of the assessments not fitting with their work context. Second, students sometimes think that they are not assessed fairly, this can be caused by the fact that not all the assessors provide feedback the same way on the assessment form. Finally, according to teachers, (external) assessors are not always approachable or available for students, which can cause a burden for students to ask for feedback on their assignment.

Time and costs

This criterion only describes if the teachers can do the assessments in the available time. This scored insufficient (*M* = 3.39). Teachers have 6 hours per student per module, which is sometimes not enough, especially when students have to do a retake. Second, teachers all have different job packages, which can intervene with the assessments they have to mark, which makes the load too much. Teachers indicate that grading assessments should be divided better between teachers with different job packages to make the work load better divided in the future when more work load is coming.

This year teachers were able to grade most of the assessments in 3 weeks after the due date, because not everybody handed it in at the same time. When more students hand it in the same time, problems can occur.

Educational consequences

The educational consequences criterion scores sufficient according to teachers (*M* = 3.66). the assessments are in the work context of the students, which motivates them when it is appreciated and puts them into work. This way the students experience a learning process when they make an assessment.

Teachers mention that when an assessment is carried out for the second time, the experiences of the first time the assessment was carried out can be shared. These former experiences can be shared between the teacher and the students, but also between students in their learning team.

Fairness

Fairness scores according to teachers just sufficient (*M* = 3.51). According to teachers the assessors are not prejudiced. When a student thinks his assessment is not marked fairly, a procedure is described in the EER (Education and Examination Regulations) which the student can look up. When assessors make mistakes in marking an assignment they can and are willing to adjust it.

According to teachers, there where incidents with assessments, but these incidents do not happen structural. In the focus group teachers indicated that they thought the score for this criterion is very low, which they had not expected.

Fitness for purpose

Fitness for purpose scores as criterion a sufficient (*M* = 3.76 and *M* = 3.98 if 8.4 is removed). When question 8.4 is removed the Cronbach's alpha is also sufficient. The score of question 8.4 is insufficient, this question is about formative and summative assessments during the educational program. This is caused by the fact that students have to ask feedback on their

own initiative in the learning teams or the lessons. They do not get it automatically and structurally throughout their educational program.

Teachers indicate that the summative assessment criteria can also be used in a formative way. Also, they indicate that feedback can be provided during the learning teams on the initiative of the students themselves.

The other questions score sufficient and score relatively high (around 4). Which seems to indicate that the assessments fit their purpose according to teachers.

Self-assessment

The self-assessment criterion scores insufficient in the teacher questionnaire (*M* = 3.46). This is also explained in the fitness for purpose criterion, which described that formative feedback is not given often. Second, students don't come up themselves with giving peer feedback to each other. Third, students are not obliged to form their own learning goals during the lessons or during the learning teams.

On the other hand, teachers indicate that there is room for (peer)feedback. Also, when looked specifically to the assessment, a part with intermediate feedback, and how this can be accomplished is included. This part mentions for example that students can ask intermediate feedback in their learning team or work context. Teacher try to stimulate discussion in their lessons, which can be seen as a form of feedback.

Meaningfulness

Meaningfulness scored sufficient according to the teacher questionnaire (*M* = 3.56 and *M*=3.58 if 10.1 is removed). According to the teachers, students think that the assessments are useful for their (future) workplace.

On the other hand, teachers mentioned that some students think the assessments are too generic, given the specific context they experience on their workplace. For those students the assessments are not meaningful for their learning process which is focussed on their work process.

When students get formative feedback, they experience it as meaningful according to teachers.

Reproducibility of decisions

Reproducibility of decisions scores insufficient according to the teacher questionnaire (*M* = 3.39). The teacher asked in the focus group think this is too low. Strong points according this criterion are first that different backgrounds are used in assessing students. This is achieved by using the calibration sessions among assessors. The different backgrounds are, on the other hand, not used consciously during the assessments. Second, assessors need to get their BQE (Basic Qualification Examination, Basis Kwalificatie Examinering in Dutch), to make sure the teachers have enough knowledge and skills to assess students and need to do a portfolio assessment training. This makes the grades reproducible.

On the other hand, teacher mention that assessors assess assessments alone, and when a student scores an insufficient mark a second assessor is used. The assessments only measure one work context of the student, which explains the low score on this question.

Teacher of the focus group would have rated this criterium higher.

Transparency

The last criterion in the questionnaire, transparency, scores sufficient according to teachers. This can be explained by the fact that students have access to the assessment and assessment criteria and that they come up with questions about those. These questions always are answered according to the teachers.

Weak points about transparency are first that the workplace of the students assesses the assessment of their students in different way compared to the assessors from Windesheim, which can cause friction. Second, teachers who lead a learning team do not always have the information about all the different modules, which makes it hard to answer questions in their learning team or lessons about a specific assessment of a module they do not teach in.

3.3 Students and teacher scores compared

This paragraph shows the answer on the 3th research question: Is there a difference in the evaluation of students and teachers of the leaning path independent assessments of Windesheim? To compare the questionnaire data of the students and the teachers, a number of t-tests were conducted with an alpha of 5%. In Table 3 below the results of the t-test are reported.

Table 3
Results t-test of the comparison of the teacher and student scores.

			Teacher		Student		
Criterium	**Question**		**Mean**	**St.dev**	**Mean**	**St.dev**	**P-value**
Acceptability	1.1	Students approve of criteria	3.60	0.883	3.70	0.803	0.665
	1.2	Students approve of procedure	3.65	0.933	*3.47*	0.909	0.459
	1.3	Teachers approve of assessments and procedure	3.70	0.979			
	1.4	Employers approve of assessments and procedure	3.82	0.809	3.59	0.785	0.315
	1.5	Confidence in quality of assessments and procedure	3.85	0.875	3.67	0.892	0.467
Authenticity	2.1	Assessment tasks resemble job	4.50	0.513	3.56	1.076	0.000**
	2.4	Assessment criteria resemble job	3.85	0.933	3.74	0.939	0.662
Cognitive	3.1	Tasks trigger thinking steps	3.65	0.875	3.72	0.797	0.751
complexity	3.2	Explain choices	4.32	0.478	4.37	0.618	0.726
	3.3	Criteria address thinking steps	*3.47*	1.020	3.83	0.803	0.149
	3.4	Tasks require thinking level	3.75	1.020	3.74	0.875	0.982
Comparability	4.1	Assessment tasks comparable	3.95	0.705	*3.31*	0.924	0.010**
	4.2	Working conditions comparable	*3.06*	0.899	*2.68*	0.850	0.137
	4.3	Assessment criteria comparable	4.15	0.489	*3.44*	0.776	0.000**
	4.4	Assessment procedure comparable	4.16	0.375	*3.32*	0.879	0.000**
Costs	5a	Time and money estimated	*3.39*	1.092			
Educational	6.1	Desired learning process stimulated	*3.50*	0.946	3.51	0.798	0.960
consequences	6.2	Positive influence on students	3.84	0.765	*3.47*	0.909	0.120
	6.3	Positive influence on teachers	3.72	1.018			
	6.4	Improved if negative effects	3.94	0.772			
	6.5	Curriculum adapted if assessments and/or procedure warrants	4.00	0.767			
Fairness	7.1	Procedure to rectify mistakes	3.61	0.698			
	7.3	Assessors not prejudiced	*3.42*	0.769	3.53	0.987	0.688
	7.4	Various types of assessment tasks	3.85	0.813	3.86	0.710	0.959
	7.5	Student think assessments and procedure are fair	*3.26*	0.653	3.60	1.027	0.188
	7.b	Teacher think assessments and procedure are fair	3.67	0.686			
Fitness for Purpose	8.1	Coverage of competence profile	4.05	0.394	3.76	0.692	0.089
	8.2	Integrated assessment of K/S/A	4.11	0.875	3.98	0.423	0.442
	8.3	Mix of different assessment forms	3.79	0.787	*3.41*	0.894	0.123
	8.4	Both summative and formative forms	*2.84*	0.765	3.51	0.960	0.009**
	8.5	Forms match with educational goals	4.00	0.562	3.81	0.852	0.378
Self-assessment	9.1	Self- and peer-assessment	*3.06*	0.938	*2.51*	0.952	0.047
	9.2	Giving and receiving feedback	3.53	0.697	*3.44*	0.983	0.736
	9.3	Reflection on personal development	3.89	0.567	3.84	0.871	0.793
	9.4	Formulation of personal learning goals	*3.05*	0.911	3.56	0.934	0.052
	9c	Feedback work place useful	3.72	0.575	4.05	0.999	0.203
Meaningfulness	10.1	Feedback formative useful	3.71	0.849	3.77	0.895	0.809
	10.2	Feedback summative useful	3.76	0.664	3.67	0.837	0.693
	10.3	Assessment is opportunity to learn	*3.20*	0.768	3.93	0.856	0.002**
	10.4	Students think criteria meaningful	3.55	0.826	3.83	0.794	0.200

	10.5	Teacher/employers think criteria meaningful	3.85	0.813	*3.33*	0.898	0.035*
Reproducibility of decisions	11.2	Several assessors	*3.11*	0.758	*3.07*	0.894	0.870
	11.3	Assessors with different backgrounds	*3.32*	0.749	*3.00*	0.795	0.154
	11.4	Equal discussion between assessors	3.61	0.979			
	11.5	Trained and competent assessors	3.84	0.834	*3.19*	0.862	0.008**
	11.6	Several work situations	*3.28*	1.074			
Transparency	12.1	Student know of formative of summative	3.60	0.940	*3.35*	0.923	0.321
	12.2	Students know criteria	*3.50*	0.827	3.51	0.910	0.961
	12.3	Students know procedures	*3.50*	0.827	*3.23*	0.996	0.301
	12.4	Teachers know and understand	4.10	0.718			
	12.5	Employers know and understand	*3.50*	0.688	*3.29*	0.891	0.347
	12.6	External party can audit	3.95	0.510	*3.43*	0.929	0.025*

*Note: Italic: scores are below the 3.5 benchmark. *p<0.05 **p<0.01*

To compare the overall results of the students and the teachers about the evaluation of the learning path independent assessments of Windesheim, the results of 3.1 and 3.2 will be compared per criterion below.

Acceptability

Students and teachers scored both sufficient on the acceptability criterion of the questionnaire. Teachers had on this criterion one question more than the students had, this question was about to what extent teachers approve of the assessment's goal, criteria and procedure.

Teachers experience trust in the assessments and have the idea that they can work with it. Students had that feeling about the assessments of module 2, these had a clear format and had a strong connection with the work context of the students.

Teachers and students both mention that the assessments and the assessment criteria in the rubric could be better aligned with each other.

Students think the assessments of module 3 should be better aligned with the lessons provided in the module.

Authenticity

The amount of which the assessment tasks resemble the job according to students, scores significant lower than the teacher score. Students and teachers comment on this item that it is sometimes hard to adjust the assessment to a specific work context of the students. Where teachers think this is possible, students think it is hard and it takes a lot of time and effort. Students would like more choice in assessment possibilities to prevent these problems for them.

Students and teachers both think that the assessments and the criteria resemble the job as a social worker or nurse. Also, all the assessments are feasible in the context of a social worker or a nurse, but not in all the different specific work situations students work in during their education.

Cognitive complexity

The scores to the cognitive complexity questions for students and teacher are not significant different from each other. In the focus group, both parties agree that the assessments trigger the thinking steps needed to perform the assessments properly in practice. An example of such a thinking step is: You declare the contemporary professional context from

historical perspective using the online module “Historical context. Students mention that they implement these thinking steps implicit in their workplace.

Teachers mention that the assessments, in particular module 3, are hard for a beginning practitioner. Second, they mention that the assessments with the thinking steps, do not always match the assessment criteria.

Students find it hard to implement the thinking steps when an assessment does not properly fit their work context. Also, the thinking steps are not implemented consciously in their work context, which would be preferable implemented conscious by them.

Comparability

According to the data of the questionnaires, students score significantly lower on 3 of the 4 questions asked in this criterion. These questions asked if the assessment tasks, assessment criteria and the assessment procedure are comparable for students.

Students mention that these scores are caused by a difference in work context among the students. Some assessments fit in their work context and other assessments do not, which makes the assessments not comparable for all the students. Also, the assessment procedure is not comparable between students, in the way in which teachers provide feedback on the assessment form students receive from their assessor. Finally, the students do not think the assessment procedure is the same for every student. To them it is not clear what makes a sufficient and what makes an insufficient mark.

Teachers on the other hand describe that students can deviate from the assessment when they justify it in a correct way. According to teachers, the assessments, the assessment criteria and the assessments are all similar for every student, but they agree on the fact that it is sometimes hard for students to make their assessment fit their work context.

A strong point is according to teachers and students, that all the assessors are open towards giving more or more specific feedback. It makes it sometimes harder when an assessor is extern or is often not available to schedule an appointment, but the possibilities are present to gain more feedback.

Costs

This question is only asked to the teachers.

Educational consequences

Students answered only two questions about the criterion educational consequences. These were about if the desired learning process of the students was stimulated and if the assessments had a positive influence on the students.

Students were positive about module 2 were the assessments resulted in a nice educational process, which fits the opinion of the students. On the other hand, students only experience an educational process when an assessment fits their learning context, otherwise they only do the assessment to get their points to get to a diploma.

Teachers think the assessments put students to work, especially when the work context of the students appreciates the outcomes and procedure of the assessment they have to do. Also, teachers mention that the second time an assessment and module is done, they use the previous experiences to support the students in their assessment and adapt the lessons and assessments when needed.

Fairness

The students filled in two questions less than the teachers about fairness. These questions were about the procedure to correct mistakes and if the teachers think the assessments and the procedure are fair.

Both students and teachers think the assessors are not prejudiced. And students think the assessments follow the professional code of their profession. According to students there is a difference in which assessors provide feedback about the assessment. Also, students think the weight of the assessment does not fit the working load of the assessment.

According to teachers, accidents happen with assessments, but these happen occasionally and not structurally. When an accident happens, a procedure is set in the EER, which details what a student needs to do when he or she is not treated fair in opinion. When an assessor has made a mistake in assessing an assessment, or filling in a mark, it will be adjusted.

Fitness for purpose

There is a significant difference between the score of one question in the fitness for purpose criterion between the student and teacher score. This question is about the use of both summative and formative forms of feedback. This difference can be explained by the fact that students experience formative feedback in the learning teams, while teachers think that is minimal use of feedback while this only happens on the initiative of the student.

According to students the assessments fit with the goal by measuring attitude and behaviour simultaneously. Teachers mentioned that the summative assessment form can be used in a formative way to by the students when they make their assessment.

Both students and teachers mentioned that formative feedback is given in the learning teams, but it can only be given when a student shows initiative to get feedback on his or her assignment.

Self-assessment

There is a significant difference between the sore of one question in the self-assessment criterion between the students and teacher score. This question is about self- and peer-assessment. Students score significant lower, but both score insufficient.

Both students and teachers mention that students do not come up with giving peer feedback to each other themselves. Also, both students and teachers mention that students do not have to come up with their own learning goals during their education.

Students mention that the assessments support the own professional development of the students at their work-place. The feedback their work-place gives can be handy for them but is focussed more on the practice than the theoretical part. Also, students can adjust their education by choosing to follow the lessons, the online module or to use people in their work context.

Teachers described that they stimulate discussion in their lessons which can also be seen as a form of feedback. Second, teachers describe that in the assessment a part is included which describes how students can get intermediate feedback during their assessments. This part mentions the learning teams, the assessment form and the work context of the students to get feedback. Also, self-assessment is mentioned in the assessments by describing that students have to check their assessment when it is done if it fits the assessment criteria.

Meaningfulness

The meaningfulness criteria consisted of 5 questions. The question: if the assessment is an opportunity to learn, scores significant lower according to teachers in comparison with students. This can be explained by the fact that the assessments are generic, and maybe not specific enough for all the possible work contexts of the students. The question about if the criteria are meaningful for the work context of the students, scores significant lower for students than for teachers. This can be explained by the fact that the criteria for full time students differ a lot form the criteria of part time students, which is hard for the work context to understand, but on the other hand, the professionals who work in the work context of the students do think the criteria are meaningful.

Teachers describe that the students think that the assessments are useful for the work they have to do in their (future) workplace.

Reproducibility of decisions

To measure reproducibility of decisions, 5 questions were asked to the teachers and 4 to students, the students did not answer the question about an equal discussion between assessors.

There is a significant difference between the answers of students and the answers of teachers about trained and competent assessors. Students scored significantly lower on this question. This difference can be explained by the fact that students feel that there is a difference in who assesses an assessment, some assessors are stricter than others.

Teachers mentioned that there is a calibration moment among assessors, in which assessors with different backgrounds participate. Students are also aware of this moment.

Teachers also mention that assessors need to get their BQFE diploma (Basic Qualification Examination, Basis Kwalificatie Examinering in Dutch) and have to do a portfolio assessment training when they assess assessments, which makes them trained and competent. On the other hand, an assessment is only assessed by one assessor, except for insufficient scores. Students know this but are uncertain if this happens. Finally, only one work context is assessed by the assessments used.

Transparency

Transparency of the assessments used is asked using 5 questions, in which one scores significant lower for students than for teachers. Teachers score higher on the question if an external party (professional of the work context) can audit than students do

Students mention that not every module is developed yet, and teachers are still looking for ways to do the assessments and the lessons. Which makes it hard for students to know what an assessment needs to look like. Second, students are not always informed about the background of an assessment, which makes it hard for them to understand the educational process needed to accomplish an assessment.

According to teachers, students know how to use the assessment criteria and the assessments, and when they have questions, these are answered by the teachers. Teachers mention that it is hard for them to know everything about the whole education. When they have a learning team in a different module they teach in, it is difficult to know the right information for the students.

4. Discussion and conclusion

This chapter discusses the results found when evaluating the learning path independent assessments of Windesheim in the educational field of health and well-being. First the results will be described shortly, second these results will be compared with the literature. After this the limitations of this research will be described. At last, the conclusion will be given.

Results

In this paragraph the results will be summarized by giving the strong and weak points of the assessments used in the part time education in the field of health and wellbeing of Windesheim. The strong and weak points are summarized in Table 4.

Table 4

Summary of the strong and weak points of the learning path independent assessments.

Strong points	Weak points
The assessments resemble the job as a nurse or social worker	The assessment, the assessment criteria and the thinking steps needed are not aligned.
Students can deviate from the assessment when they underpin their decisions.	The assessments are generic and sometimes hard to fit into specific work situations, which does not contribute to the educational process of the students.
Teachers are open toward explaining their feedback or giving more specific feedback on an assessment.	Assessors assess assessments differently from each other. Especially in giving feedback.
Former experiences are shared between teachers and students, or between students about assessments.	Students are unsecure about the assessment procedure, what makes an insufficient, sufficient, good grade?
Formative feedback can be given during the learning teams, which are comparable with mentor classes.	Assessors think the assessment is hard to do in the available hours, especially when students have to do a retake, or when more students hand in their assessments.
The assessment form can be used for self-assessment. This is also mentioned in the assessment.	The assessment weight does not resemble the working load of the assessments according to students.
Students can choose their own study route to accomplish an assessment.	Formative feedback is only given on initiative of the students themselves.
There is calibration among teachers who assess an assessment, these teachers all have different backgrounds.	There is little use of peer-feedback and self-assessment.
When a student thinks he or she is treated unfair, a procedure in the EER details how to solve this.	Some parts of the education are not developed yet.
Mistakes in the marking of assessments, the assessment itself and the assessment criteria are corrected when needed.	Why students have to do a specific assessment as a nurse or social worker is not always explained.

According to the students and teachers of Windesheim several weak and strong points about the learning path independent assessments are mentioned: the content of the assessments, the assessment criteria, and the alignment between those, and the assessment procedure.

The assessments are described in generic terms, which can make it hard for students to fit the assessment in their work context, but teachers mention that students can deviate from the assessment when their choices are explained well. Students can choose their own study route, in which they can choose between following the lectures, following the online study route or learn it in their work context. Students do not always understand why they have to do a specific assessment in their work context. This is not explained by their teacher, but on the other hand, the assessments used resemble the overall job of a nurse or social worker in all the work contexts possible. Finally, students mention that the assessment weight does not resemble the work load an assessment takes. This can be explained by the

fact that in the learning path independent assessments the assessment weight does not resemble the working load., According to the OCW the learning outcomes are not time bound anymore (OCW, 2016).

The assessment, the assessment criteria and the thinking steps needed in the assessment are not aligned with each other. This can be explained by the fact that everything is rather new. When teachers mention a mistake or a deviation they change it in the assessment or in the assessment criteria. Among assessors there is always a calibration meeting in which these mistakes can be discussed, and corrections can be made.

When assessments are made, difference between these assessments are mentioned by both students and teachers. This relates especially to the way feedback is provided by the assessors on the assessment form. This is not consistent. Also, students are unsure which makes a specific mark, so what makes an insufficient, or a sufficient or good mark. Teachers have to mark the assessments holistic, which is hard to understand for students. When a student needs more information about the feedback on their assessment, they can always ask their assessor. These are open towards this. Because the assessments and everything around it are rather new, teachers and assessors are still experiencing new things which can be shared to overcome uncertainties of students. These problems with new assessments are not over yet, because parts of the overall education are still being developed.

To continue on the feedback, not only summative feedback can be given to students, but also formative feedback. This feedback is now given to students during their learning team on the student's own initiative. Students are stimulated to do self-assessment, which is also mentioned in the assessment they have to do, but this is not always done by them. They can do this by using the summative assessment criteria for formative purposes. Also, there is little use of peer-feedback among members of a learning team. Assessing takes a lot of time from teachers, and they mentioned that it is now doable in the given time but when more students hand it in, it will be very hard to accomplish, especially when students need to do a retake.

When a student is not satisfied about their assessed assessment, a procedure in the EER details what a student can do in case he or she thinks it is unfair. This happens incidental.

Literature

According to Gulikers et al., (2009) teachers are more critical than students are in an evaluation of assessments. According to this research students are more critical, they score on 8 questions significant lower than teacher did. Teachers scored on 2 questions significant lower than the students. This can be explained by the fact that the whole flexible educational concept is rather new, and students have to get use to this specific type of education. Teachers have developed everything and have more background information, which can make them less critical. Also, students who had an insufficient mark can be more critical or negative than students who had a sufficient mark.

When Baartman et al., (2013) conducted their self-evaluation procedure, they found an overall high score on: fitness for purpose, fairness and accountability. This research found overall high scores for acceptability, authenticity and cognitive complexity. Baartman et al., (2013) found a rather good score for: transparency, acceptability, cost & efficiency, authenticity and complexity. This research found a rather good score for educational consequences, fairness, fitness for purpose and meaningfulness. And finally, Baartman et al., (2013) found low scores on reproducibility of decisions and development of self-regulated learning. This research found that the criteria: comparability, cost & efficiency, development

of self-regulated learning, reproducibility of decisions and transparency scored low. When these results are compared with the research of Baartman et al. (2013) only development of self-regulated learning and reproducibility of decisions score similar. The difference in scores can be explained by the fact that these assessments are rather new for students and teachers at Windesheim, and the assessments used in the research of Baartman (2013) are used for a longer time.

Windesheim

The assessments are rather new for the students and the teachers, they are implemented in the study year of 2017/2018. This means that everybody needs to learn how to work with it, which can cause problems for the students and the teachers. When these assessments are used longer, teacher can use the previous experiences to help students with their assessments and help other teachers. Also, students can help each other in the learning teams when they are further in their education or when they are further in their assessments.

A limitation is the small sizes of the focus groups of students and teachers due to the end of the schoolyear. Both of the focus groups consisted of 2 people. For the students this were students from different classes of which one studies nursing and the other social work. Because of this mix, both studies were discussed during in the focus group which ensures that the results relate to both educational fields. The teacher focus group also contained two different teachers. One teacher only grading assessments and the other teaches, assesses, leads a learning team and is chairman of a module. This can make the results representative for the teacher population. Finally, the focus groups were used for additional data to understand and interpret the questionnaire data which is leading. More participants for the focus group would be desirable, but for the goal of the focus groups this population was sufficient.

The research is done using an already existing method, the self-evaluation procedure of Baartman (2006), which is proven valid in other self-assessment procedures. Also, a combination of quantitative data and qualitative data is used which had outcomes that are aligned. The results are not reliable on a longer term, because the assessments are rather new, and students and teacher are still learning to work with them and still learn from experiences. When this is clearer, the results of the evaluation could have been more positive. Because of the use of students with different backgrounds and teacher with different work situations the results do represent the entire group of students and teacher of the part-time education in the field of health and well-being of Windesheim. This can also mean that the results are not valid, but because of a valid method the results are valid at this moment, but are not valid further in time, when the students and teacher are more familiar with the assessments or when modifications are made in the assessments, the assessment criteria and/or the assessment procedure.

When another party had done an evaluation, different results could have been found. This can be explained by the fact that opinions are used in this research and another party could look more into the assessments, assessments criteria and the assessment procedure and how this is designed.

By doing this research at the learning path independent assessments Windesheim uses in the part-time education in the domain of health and well-being, the self-evaluation procedure is validated even more. By using this procedure, the internal quality of the assessments of Windesheim can be increased. This research showed that this procedure fits the learning path independent assessments used in the pilot flexibilization, so this procedure

could also be used by other schools in this pilot to evaluate their assessments. Finally, this research contributes to the literature of self-evaluation of assessments in higher education.

In a follow-up study in a later stage, the self-evaluation procedure can also be used by Windesheim to increase the internal quality of the assessments of Windesheim even more over time.

Conclusion

According to the self-evaluation procedure, modifications need to be made in the content of the assessments, assessment criteria and the whole assessment procedure. The assessment procedure includes the whole procedure in relation to the assessments. First the assessments and the assessment criteria need to be better aligned. Second the students all have different work contexts which makes assessments sometimes not fitting. And at last, feedback should be provided more and better aligned, assessors all provide feedback differently on assessments and formative feedback needs to be implemented more.

The assessments and the flexible education in which they are implemented are relatively new, (implemented in 2017/2018) and had a relatively good start but need some modifications to fulfil the expectations of the students and the teachers of the part-time education in the field of health and well-being of Windesheim.

5. Recommendations

According to the results the following steps could be made by Windesheim to improve the learning path independent assessments:

1. The content of the assessments and assessment criteria need to be better aligned. The developers of the different modules need to check if their assessments are aligned with the assessment criteria and the assessment procedure, which involves grading, used. The developer or the chairman of the module can do it themselves or using other assessors during calibration sessions. Also, students can be to indicate where things are unclear for them during their assessments.
2. Students need sometimes help with fitting the assessment to their work context, because of the broad range of work contexts. To solve this, developers of a module have to think about how to help students at the start of the assessment with fitting it into their work context. This can prevent stress and problems in the end of the assessment for students. Also, teachers than have to assess less, because it can prevent some insufficient marks and retakes in the end. This role can be taken by the mentor of the learning team of the student or the teacher in the module.
3. Feedback by the teachers need to be better aligned with another teacher's feedback. On the assessment criteria form teachers can provide feedback. This could be more structured. This can be accomplished by using a form which structures the feedback for the students or by providing a training about giving feedback to students.
4. There needs to be more formative feedback for students. This can be done by implementing self-assessment or peer feedback. Self-assessment can be done using the summative assessment criteria by students during their assessment to help them to get a sufficient grade. Also, peer-feedback can be used whereby students can assess assignments form other students. This again can be done by using the summative assessment criteria. This peer feedback can be arranged between students of a learning team. To motivate students to use self-assessment, students can be asked to fill in the assessment criteria for their own assessment and bring that to their class or learning team in which questions could be asked with regards to their own self-assessment.

Reference list

Aea. (2011). European Framework of Standards for Educational Assessment 1.0, 29. Retrieved from http://www.aea-europe.net/images/downloads/SW_Framework_of_European_Standards.pdf

Baartman, L. K. J., Bastiaens, T. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in Educational Evaluation*, *32*(2), 153–170. https://doi.org/10.1016/j.stueduc.2006.04.006

Baartman, L. K. J., Gulikers, J., & Dijkstra, A. (2013). Factors influencing assessment quality in higher vocational education. *Assessment & Evaluation in Higher Education*, *38*(8), 978–997. https://doi.org/10.1080/02602938.2013.771133

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007a). Determining the Quality of Competence Assessment Programs: a Self-Evaluation Procedure. *Studies in Educational Evaluation*, *33*(3–4), 258–281. https://doi.org/10.1016/j.stueduc.2007.07.004

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2007b). Kwaliteitsmeting van Competentie Assessment Programma ' s via zelfevaluatie, 17–26.

Baartman, L. K. J., Prins, F. J., Kirschner, P. A., & van der Vleuten, C. P. M. (2011). Self-evaluation of assessment programs: A cross-case analysis. *Evaluation and Program Planning*, *34*(3), 206–216. https://doi.org/10.1016/j.evalprogplan.2011.03.001

Dijkstra, A., & Baartman, L. K. J. (2011). Zelfevaluatie van de kwaliteit van assessment. *OnderwijsInnovatie (Open Universiteit Nederland)*, (maart), 17–25.

Epstein, R. M., & Hundert, E. M. (2002). Defining and assessing professional competence. *JAMA*, *287*(2), 226–35. Retrieved from www.bristol-inquiry.org.uk/final_report/report/sec2chap25_4.htm

Gerritsen-van Leeuwenkamp, K. J., Joosten-ten Brinke, D., & Kester, L. (2017). Assessment quality in tertiary education: An integrative literature review. *Studies in Educational Evaluation*, *55*(September), 94–116. https://doi.org/10.1016/j.stueduc.2017.08.001

Gulikers, J., Biemans, H., & Mulder, M. (2009). Developer, teacher, student and employer evaluations of competence-based assessment quality. *Studies in Educational Evaluation*, *35*(2–3), 110–119. https://doi.org/10.1016/j.stueduc.2009.05.002

Gulikers, J. T. M., Baartman, L. K. J., & Biemans, H. J. A. (2010). Facilitating evaluations of innovative, competence-based assessments: Creating understanding and involving multiple stakeholders. *Evaluation and Program Planning*, *33*(2), 120–127. https://doi.org/10.1016/j.evalprogplan.2009.07.002

Maassen, N., Otter, D. den, Wools, S., Hemker, B., Straetmans, G., & Eggen, T. (2014). Kwaliteit van toetsen binnen handbereik, (September), 1–6.

McMullan, M., Endacott, R., Gray, M. A., Jasper, M., Miller, C. M. L., Scholes, J., & Webb, C. (2003). Portfolios and assessment of competence: A review of the literature. *Journal of Advanced Nursing*, *41*(3), 283–294. https://doi.org/10.1046/j.1365-2648.2003.02528.x

Nederlands Vlaamse Accreditatieorganisatie. (2018). NVAO Accreditatieorganisatie,. Retrieved April 5, 2018, from https://www.nvao.net/

OCW. (2016). Handreiking pilots flexibilisering hoger onderwijs, (april), 1–39.

Olthof, M., Stulen, E., & Mossel, R. van. (2017). Toolkit Flexibel hoger onderwijs voor volwassenen. Retrieved June 28, 2018, from https://elo.windesheim.nl/CMS/Windesheim algemeen/Onderwijsadvies/Ontwerpteam flexibel deeltijd/Toolkit_Flexibel_Deeltijd_deployment/index.htm#topic3

Rijksoverheid. (n.d.). Pilots flexibilisering. Retrieved March 19, 2018, from https://www.rijksoverheid.nl/onderwerpen/hoger-onderwijs/experimenten-om-deeltijdonderwijs-flexibeler-te-maken/pilots-flexibilisering

van Berkel, A. (2017). De assessmentdriehoek voor leerstofonafhankelijk toetsen en begeleiden. *Onderwijsinnovatie*, *2*(juni 2017), 17–25.

Van Der Vleuten, C. P. M., & Schuwirth, L. W. T. (2005). Assessing professional competence: From methods to programmes. *Medical Education*, *39*(3), 309–317. https://doi.org/10.1111/j.1365-2929.2005.02094.x

Windesheim. (2017). Handleiding leerwegonafhankelijke toetsing, (December).

Wools, S. (2012). Towards a Comprehensive Evaluation System for the Quality of Tests and Assessments. *Psychometrics in Practice at RCEC*, 115–124.

Appendix 1

Table 3. Operationalisation of the criteria

Criteria		Operationalization	Students questionnaire	Teacher questionnaire	Not asked
1. Acceptability	1	Students approve of criteria	x	x	
	2	Students approve of procedure	x	x	
	3	Teachers approve of CAP		x	
	4	Employers approve of CAP	x	x	
	5	Confidence in quality of CAP	x	x	
2. Authenticity	1	Assessment tasks resemble job	x	x	
	2	Working conditions resemble job	X	x	
	3	Social context resembles job	x	x	
	4	Assessment criteria resemble job	x	x	
3. Cognitive complexity	1	Tasks trigger thinking steps	x	x	
	2	Explain choices	x	x	
	3	Criteria address thinking steps	x	x	
	4	Tasks require thinking level	x	x	
4. Comparability	1	Assessment tasks comparable	x	x	
	2	Working conditions comparable	x	x	
	3	Assessment criteria comparable	x	x	
	4	Assessment procedure comparable	x	x	
5. Costs & Efficiency	1	Time and money estimated			X
	2	Deliberately choosing mix			X
	3	Yearly evaluation of efficiency			X
	4	Positive effects outweigh investment			x
6. Educational consequences	1	Desired learning process stimulated	x	x	
	2	Positive influence on students	x	x	
	3	Positive influence on teachers	x	x	
	4	Improved if negative effects		x	
	5	Curriculum adapted if CAP warrants		x	
7. Fairness	1	Procedure to rectify mistakes		X	
	2	Weights based on importance		X	
	3	Assessors not prejudiced		X	
	4	Various types of assessment tasks	x	x	
	5	Student think CAP is fair	x	x	
8. Fitness for purpose	1	Coverage of competence profile	x	x	
	2	Integrated assessment of K/S/A	x	x	
	3	Mix of different assessment forms	x	x	
	4	Both summative and formative forms	x	x	
	5	Forms match with educational goals	x	x	
9. Fitness for self-assessment	1	Self- and peer-assessment	X	x	
	2	Giving and receiving feedback	x	x	
	3	Reflection on personal development	x	x	

	4	Formulation of personal learning goals	x	x	
10. Meaningfulness	1	Feedback formative useful	x	x	
	2	Feedback summative useful	x	x	
	3	Assessment is opportunity to learn	x	x	
	4	Students think criteria meaningful	x	x	
	5	Teacher/employers think criteria meaningful		x	
11. Reproducibility of decisions	1	Several times	x	x	
	2	Several assessors		x	
	3	Assessors with different backgrounds		x	
	4	Equal discussion between assessors		x	
	5	Trained and competent assessors		x	
	6	Several work situations		x	
12. Transparency	1	Student know of formative of summative	x	x	
	2	Students know criteria	x	x	
	3	Students know procedures	x	x	
	4	Teachers know and understand	x	x	
	5	Employers know and understand	x	x	
	6	External party can audit	x	x	

Appendix 2

Beste Student,

Mijn naam is Kimberly de Jonge en ik studeer Educational Science and Technology aan de Universiteit Twente. Ik heb voor mijn opleiding de opdracht gekregen om de toetsing van het flexibele onderwijs te evalueren en ik ben nu bezig met het evalueren van de toetsen die gebruikt worden tijdens het onderwijs. Dit wil ik graag doen door studenten en docenten te vragen om deze vragenlijst in te vullen.

Zoals hierboven beschreven is het doel van mijn onderzoek het evalueren van de bestaande en gebruikte toetsen, om indien mogelijk of noodzakelijk deze toetsen te verbeteren, ook kan deze evaluatie helpen bij het ontwikkelen van nieuwe toetsen.

Het invullen van de vragenlijst duurt ongeveer 10 minuten. En alle resultaten worden anoniem verwerkt.

Alvast heel erg bedankt!

1= helemaal niet mee eens, 2= niet mee eens, 3=neutraal, 4= mee eens, 5=helemaal mee eens.

Criterium	**Vragen**	**1**	**2**	**3**	**4**	**5**	**-**
1. Acceptatie	Ik kan me vinden in de beoordelingscriteria van de EVL.						
	Ik kan me vinden in de wijze waarop de toets-opdracht uitgevoerd moet worden.						
	Mijn leerwerkbegeleider kan zich vinden in de beoordelingscriteria en de procedures van de toets-opdracht.						
	Ik heb vertrouwen in de kwaliteit van de toets-opdrachten en de beoordelingscriteria.						
2. Authenticiteit	De toets-opdrachten bevatten activiteiten die ik op de werkplek moet uitvoeren.						
	De beoordelingscriteria lijken op de criteria waaraan werknemers in het toekomstige moeten voldoen.						
3. Cognitieve complexiteit	De toets-opdrachten roepen de denkstappen op die beginnende beroepsbeoefenaren hanteren.						
	Bij het maken van een toets-opdracht moet ik uitleggen waarom ik bepaalde keuzes heb gemaakt.						
	De beoordelingscriteria zijn ook gericht op de gehanteerde denkstappen.						
	De toets-opdrachten vereisen het denkniveau dat beginnend beroepsbeoefenaren nodig hebben.						
4. Vergelijkbaarheid	De toets-opdrachten zijn voor alle studenten vergelijkbaar en eventuele verschillen worden verantwoord.						
	De werkomstandigheden zijn voor alle studenten vergelijkbaar en met eventuele verschillen wordt rekening gehouden in het oordeel.						
	De beoordelingscriteria zijn voor alle studenten vergelijkbaar en eventuele verschillen worden verantwoord.						
	De beoordelingsprocedure is voor alle studenten vergelijkbaar en eventuele verschillen worden verantwoord.						
5. Onderwijsgevolgen	De toets-opdracht roept bij mij de gewenste leerprocessen op in de voorbereiding naar een beoordeling.						
	Ik word op een positieve manier beïnvloed door de toets-opdracht.						
6. Eerlijkheid	De beoordeelaars zijn niet bevooroordeeld.						
	De toets-opdrachten zijn gevarieerd.						
	Ik ervaar de beoordeling als eerlijk.						
7. Geschiktheid voor onderwijsleerdoelen	De toets-opdrachten dekken de leeruitkomsten en eindtermen.						
	In de toets-opdrachten worden kennis, vaardigheden en attitude geïntegreerd beoordeeld.						
	Het toets-programma bestaat uit een mix van verschillende beoordelingsvormen.						
	De EVL bevat zowel summatieve als formatieve (feedback) beoordelingsvormen.						

Criterium	Vragen	1	2	3	4	5	-
	De gekozen beoordelingsvormen passen bij de leeruitkomsten van het onderwijs.						
8. Ontwikkeling van zelfsturend leren	De studenten beoordelen zichzelf of elkaar.						
	De toets-opdrachten stimuleren het op een goede manier (leren) geven en ontvangen van feedback.						
	De toets-opdrachten stimuleren het (leren) reflecteren op de eigen ontwikkeling.						
	De toets-opdrachten stimuleren het formuleren van eigen leerdoelen, gebaseerd om de eigen ontwikkeling.						
	Ik vind de feedback van mijn leerwerkbegeleider zinvol voor mijn eigen leerproces.						
9. Betekenisvolheid	Ik vind de feedback van de formatieve beoordelingsmomenten zinvol voor mijn leerproces.						
	Ik vind de feedback van de summatieve beoordelingsvormen (toets-opdracht) zinvol voor mijn leerproces.						
	Ik ervaar de beoordeling als een leermoment.						
	Ik vind de beoordelingscriteria betekenisvol met betrekking tot mijn toekomstige beroep.						
	Mijn werkgever vindt de beoordelingscriteria betekenisvol met betrekking tot de eisen die zij stellen aan toekomstige beroepsbeoefenaars.						
10. Herhaalbaarheid van beslissingen	Voor een summatief eindoordeel (cijfer) wordt het oordeel van meerdere beoordeelaars gecombineerd.						
	Bij een summatieve beoordeling worden beoordeelaars met verschillende achtergronden ingezet.						
	Een summatief oordeel wordt gebaseerd op beoordelingen in verschillende werksituaties.						
11. Transparantie	Ik weet of een beoordeling formatief of summatief bedoeld is.						
	Ik ken en begrijp de beoordelingscriteria van de EVL.						
	Ik weet en begrijp hoe ik de toets-opdracht uit moet voeren.						
	Mijn praktijkbegeleiders kennen en begrijpen het doe, de criteria en de procedure van de toets-opdracht.						
	Mijn werkgever (externe partij) kan op basis van een vastgestelde procedure en de beschrijving van de uitgevoerde toets-opdracht een controle uitvoeren.						
12. Informatie	Opleiding (PMK of SW of VPK)						
	Leeftijd (in jaren)						
	Geslacht						
	Toetsopdrachten gemaakt (EVL: 1.1,1.2,1.3,2.1,2.2,2.3,3.1,3.2,3,3)						
	Leerervaring (bijv. mbo, vmbo, hbo)						
	Werkervaring in de zorg (in jaren)						

Focusgroep

Graag zou ik een focusgroep willen organiseren om de resultaten te bespreken. Zou je daaraan mee willen werken?

- Ja, het liefst op een maandag
- Ja, het liefst op een dinsdag
- Ja, het liefst op een woensdag
- Ja, het liefst op een donderdag
- Ja, het liefst op een vrijdag
- Nee

Laat hier je e-mailadres achter wanneer je mee wil doen aan de focusgroep:

……

Heb je verder nog op- en/of aanmerkingen over de vragenlijst of de toestopdrachten?

Appendix 3

Beste Docent(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(Baartman et al., 2007a)(L. K. J. Baartman et al., 2007a)(L. K. J. Baartman et al., 2007a)(L. K. J. Baartman et al., 2007a)(L. K. J. Baartman et al., 2007a)(L. K. J. Baartman et al., 2007a),

In het kader van het evalueren van de nieuwe deeltijd zou ik graag willen vragen of u de onderstaande vragenlijst willen invullen **voor 9 juni.**

Mijn naam is Kimberly de Jonge en ik studeer Educational Science and Technology aan de Universiteit Twente. Ik heb voor mijn opleiding en de projectgroep de opdracht gekregen om de toetsing van het flexibele onderwijs te evalueren. Dit wil ik graag doen door studenten en docenten te vragen om deze vragenlijst in te vullen.

Zoals hierboven beschreven is het doel van mijn onderzoek het evalueren van de bestaande en gebruikte toetsen, om indien mogelijk of noodzakelijk deze toetsen te verbeteren, ook kan deze evaluatie helpen bij het ontwikkelen van nieuwe toetsen.

Het invullen van de vragenlijst duurt ongeveer 10 minuten. En alle resultaten worden anoniem verwerkt.

Alvast heel erg bedankt!

1= helemaal niet mee eens, 2= niet mee eens, 3=neutraal, 4= mee eens, 5=helemaal mee eens.

Criterium	**Vraag**	**1**	**2**	**3**	**4**	**5**	**-**
1. Acceptatie	De studenten kunnen zich vinden in de beoordelingscriteria van de EVL.						
	De studenten kunnen zich vinden in de wijze waarop de toets-opdrachten uitgevoerd moet worden.						
	Ik kan me vinden in het doel, de beoordelingscriteria en de procedure van de toets-opdracht.						
	De leerwerkbegeleiders kunnen zich vinden in de beoordelingscriteria en de procedures van de toets-opdracht.						
	Ik heb vertrouwen in de kwaliteit van de toets-opdrachten en beoordelingscriteria.						
2. Authenticiteit	De toets-opdrachten bevat activiteiten die studenten op de werkplek moeten uitvoeren.						
	De beoordelingscriteria lijken op de criteria waaraan werknemers in het toekomstige beroep moeten voldoen.						
3. Cognitieve complexiteit	De toets-opdrachten roepen de denkstappen op die beginnend beroepsbeoefenaren hanteren.						
	Bij het maken van een toets-opdrachtenmoeten een student uitleggen waarom bepaalde keuzes zijn gemaakt.						
	De beoordelingscriteria zijn ook gericht op de gehanteerde denkstappen.						
	De toets-opdrachten vereisen het denkniveau dat beginnend beroepsbeoefenaren nodig hebben.						
4. Vergelijkbaarheid	De toets-opdrachten zijn voor alle studenten vergelijkbaar en eventuele verschillen worden verantwoord.						
	De werkomstandigheden zijn voor alle studenten vergelijkbaar en met eventuele verschillen wordt rekening gehouden in het oordeel.						
	De beoordelingscriteria zijn voor alle studenten vergelijkbaar en eventuele verschillen worden verantwoord.						
	De beoordelingsprocedure is voor alle studenten vergelijkbaar en eventuele verschillen worden verantwoord.						
5. Tijd en kosten	Ik kan de toetsing binnen de beschikbare uren uitvoeren.						
6. Onderwijsgevolgen	De toets-opdrachtenroepen bij de studenten de gewenste leerprocessen op in de voorbereiding naar een beoordeling.						
	De studenten worden op een positieve manier beïnvloed door de toets-opdracht.						

Criterium	Vraag	1	2	3	4	5	-
	Ik word op een positieve manier beïnvloed door de toets-opdracht.						
	De toets-opdrachten worden verbeterd als onverwachte negatieve gevolgen worden gevonden.						
	De leeractiviteiten wordt aangepast als de resultaten van de toets-opdrachten dit vereisen.						
7. Eerlijkheid	Er zijn procedures opgesteld voor het corrigeren van eventueel gemaakte fouten tijdens de beoordeling.						
	De beoordeelaars zijn niet bevooroordeeld.						
	De toets-opdrachten zijn gevarieerd.						
	Studenten ervaren de beoordeling als eerlijk.						
	Ik ervaar de beoordeling als eerlijk.						
8. Geschiktheid voor onderwijsleerdoelen	De toets-opdrachtendekken de leeruitkomsten en eindtermen.						
	In de toets-opdrachten worden kennis, vaardigheden en attitudes geïntegreerd beoordeeld.						
	Het toets-programma bestaat uit een mix van beoordelingsvormen.						
	De EVL bevat zowel summatieve als formatieve beoordelingsvormen.						
	De gekozen beoordelingsvormen passen bij de leeruitkomsten van het onderwijs.						
9. Ontwikkeling van zelfsturend leren	De studenten beoordelen zichzelf of elkaar.						
	De toets-opdrachten stimuleren het op een goede manier (leren) geven en ontvangen van feedback.						
	De toets-opdrachten stimuleren het (leren) reflecteren op eigen ontwikkeling.						
	De toets-opdrachten stimuleren het formuleren van eigen leerdoelen, gebaseerd op de eigen ontwikkeling.						
	De studenten vinden de feedback van de leerwerkbegeleiders zinvol voor hun eigen leerproces.						
10. Betekenisvolheid	De studenten vinden de feedback van de formatieve beoordelingsmomenten zinvol voor hun leerproces.						
	De studenten vinden de feedback van de summatieve beoordelingsvormen zinvol voor hun leerproces.						
	De studenten ervaren hun beoordeling als leermoment.						
	De studenten vinden de beoordelingscriteria betekenisvol met betrekking tot hun toekomstige beroep.						
	Ik vind de beoordelingscriteria betekenisvol met betrekking tot de eisen die ik en bedrijven stellen aan toekomstige beroepsbeoefenaars.						
11. Herhaalbaarheid van beslissingen	Voor een summatief eindoordeel wordt het oordeel van meerdere beoordeelaars gecombineerd.						
	Bij summatieve beoordeling worden beoordeelaars met verschillende achtergronden ingezet.						
	Tussen de verschillende beoordeelaars vindt een gelijkwaardig overleg plaats waarin iedereen zijn oordeel onderbouwt.						
	De beoordeelaars zijn getraind en competent voor de verschillende beoordelingsvormen.						
	Een summatief oordeel wordt gebaseerd op de beoordelingen in verschillende werksituaties.						
12. Transparantie	De studenten weten of een beoordeling formatief of summatief is bedoeld.						
	De studenten kennen en begrijpen de beoordelingscriteria van de EVL.						
	De studenten weten en begrijpen hoe de toets-opdrachten worden uitgevoerd.						
	Ik ken en begrijp het doel, de criteria en de procedure van de toets-opdracht.						
	De praktijkbegeleiders kennen en begrijpen het doel, de criteria en de procedure van de toets-opdracht.						

Criterium	Vraag	1	2	3	4	5	-
	Een externe partij kan op basis van de vastgestelde procedures en de beschrijving van de uitgevoerde toets-opdrachten een controle uitvoeren.						
13. Informatie	Bij welke opleiding bent u betrokken? (PMK of SW of VPK)						
	Wat is uw leeftijd? (In jaren)						
	Wat is uw geslacht?						
	Bij welke toets-opdrachten was u betrokken? (1.1,1.2,1.3,2.1,2.2,2.3,3.1,3.2,3,3)						
	Hoeveel werkervaring heeft u in het onderwijs? (In jaren)						
	Hoeveel werkervaring heeft u in het onderwijs binnen Windesheim en het domein Gezondheid & Welzijn? (In jaren)						

Focusgroep
Graag zou ik een focusgroep willen organiseren om de resultaten te bespreken. Zou u daaraan mee willen werken?

- Ja, het liefst op een maandag
- Ja, het liefst op een dinsdag
- Ja, het liefst op een woensdag
- Ja, het liefst op een donderdag
- Ja, het liefst op een vrijdag
- Nee

Laat hier je e-mailadres achter wanneer u mee wil doen aan de focusgroep:

……

Heeft u verder nog op of aanmerkingen over de vragenlijst of de toestopdrachten?

Appendix 4

Toestemmingsverklaring Focusgroep Kwaliteit Toets-opdrachten

Titel onderzoek: Kwaliteit toets-opdrachten
Verantwoordelijke onderzoekers: Kimberly de Jonge,

In te vullen door de deelnemer

Aan mij is op een duidelijke manier verteld over het onderzoek: over het doel, de methode en wat het van mij vraagt. Ik weet dat de gegevens en resultaten van het onderzoek alleen anoniem en vertrouwelijk worden gepresenteerd en gedeeld. Mijn naam komt dus niet terug in rapporten, presentaties of andere publicatievormen. Wat ik heb verteld, wordt alleen gedeeld op een vertrouwelijke manier. Ik ben tevreden over hoe mijn vragen zijn beantwoord.

Ik ga akkoord met het opnemen van de focusgroep m.b.v. audio-apparatuur. Ik begrijp dat geluidsmateriaal of bewerking daarvan uitsluitend voor analyse en/of (wetenschappelijke) presentaties en rapportages zal worden gebruikt.

Ik doe geheel vrijwillig mee met dit onderzoek.

Naam deelnemer: ……..

Datum: …….. 2018

Handtekening deelnemer: ……..

Ondergetekende verklaart dat de hierboven genoemde persoon zowel mondeling als schriftelijk over het bovenvermelde onderzoek geïnformeerd is. Hij/zij verklaart tevens dat een voortijdige beëindiging van de deelname door bovengenoemde persoon, geen enkele gevolgen zal hebben.

Naam Kimberly de Jonge

Functie Master EST, Utwente, Windesheim

Handtekening …….

Appendix 5.

Results focus group students

Criterium	Strong points	Weak points
Acceptability	Module 2: Had a good connection with the work context and had a format in which the students could work.	The assessment and the rubric should be better aligned. The assessment criteria are complicated for the work context. The rubric works better for the students than the assessment does. Module 3: the assignment did not fit into the work context of some students and missed a clear explanation.
Authenticity	The assessments and the assessment criteria fit with the work context as professional. You can get knowledge about all sides. All the assignments are feasible in the context of social worker or nurse.	No choice in the assessments when a work context does not fit the assessment.
Cognitive complexity	Students are triggered to make thinking steps by doing the assessment. The thinking steps are unconsciously implemented in the work context of the students. The rubric provides information about the substantiation a student needs to give in their assessments.	When an assessment does not fit the work context the thinking steps are hard to implement during daily practise. The students do not implement the thinking steps consciously.
Comparability	The assessments are for all students the same. All the assessors are open towards giving more feedback when a student asks. For module 2 (2.1) everybody had the same format they had to use, which makes the assessments comparable.	Module 3: It is not customized to different work contexts, which makes students not comparable. It is unclear what makes a sufficient or insufficient? Assessors assess on different ways and give different kinds of feedback. Every assessment is different which makes it hard to compare them.
Educational consequences	Module 2: Was really nice to do and resulted in an educational process.	Students think: I have to do this to get my diploma and look at the assessment criteria instead of the assessment used. Some assessments do not fit the work practice of the students, which makes them less motivated.
Fairness	Students don't think the teachers are prejudiced. The assessments fit the professional code.	There is a different in how teachers mark assignments and provide feedback. The weight of the assessment does not fit with the working load of the assessment.
Fitness for purpose	The assessments fit with the goal by measuring attitude and behaviour at once. The learning teams are open towards feedback.	There are no moments during the education to get formative feedback. The lessons do not fit the assessments used. Everybody does another assessment, so getting feedback during the learning teams is difficult to manage.
Self-assessment	The assessments support the own development of the students.	There are no moments for peer feedback.

	The feedback of the work place is focussed more on situations in practice but is meaningful. You can choose if you use the lessons, online modules or people in practice.	The assessments are strict in the way which they have to look like, which makes it hard to come to the learning goal your one way.
Meaningfulness	The working context of the students says the rubric is meaningful.	There is a big difference with the fulltime education for the work context. APA guidelines are not clear.
Reproducibility of decisions	There is calibration among the assessors. It is not sure who assesses an assignment beforehand, so they are not prejudiced.	There is always a first assessor, but it is unsure that a second assessor was asked in case of an insufficient. Students are not sure if they get the same grade when they hand it in at another time. There is a difference in who assesses an assignment, some are stricter than others.
Transparency		Not all the modules are developed. Everybody is still looking for ways to do the lessons and the assessments. The background information is not clear, which makes it hard to understand the educational process. It is not clear how the assessments need to look like.

Appendix 6

Criterium	Strong points	Weak points
Acceptability	There is trust in the assessment criteria and the assessments, and people have the idea they can work with it.	The students indicate that the assessment criteria are not very clear, not guiding and are formulated in an abstract way. It is all new, and minor adjustments need to be made in the assessments and the assessment criteria.
Authenticity	The assessments need to be done at the workplace of the students.	Sometimes it is hard to adjust the assessments to the work situation of the students, this takes time and effort from the students.
Cognitive complexity	The assessments call up the necessary thinking steps needed.	The assessments are hard for a beginning professional. The assessment criteria and the assessment do not always match with each other.
Comparability	A student can deviate from the assessment when it is justified. The assessments, the assessment criteria and the procedure are all similar for all the students.	The work situation of all the students are different which makes some assessments not fitting. Students sometimes have the idea that they are not assesses fairly. Not all assessors assess on the same way, in providing feedback for the students when filling in the assessment form. The (extern)assessors are not always even approachable or available.
Costs	It is reachable to assess the assessments within 3 weeks, but when more assessments are handed in it will become a challenge.	6 hours per student per module is sometime not enough, especially with retakes. Assessing does not fit in all the job packages, this should be aligned better between the assessors.
Educational consequences	The assessments put the students to work, especially when the work context appreciates it. The second time of the assessments experiences can be shared between the teacher and the students but also between students in a learning team.	
Fairness	The assessors are not prejudiced. When a student thinks his assessment is not fair, the student can follow a procedure which is set in the OER. When a assessor makes a mistake it will be undo.	There are incidents with assessments, but these happen incidental and not structural.
Fitness for purpose	The summative assessment criteria can be used formative to give feedback. Formative feedback can be given in the learning team of the student on their own initiative.	Students have to show initiative to get feedback during their learning team.
Self-assessment	During the learning teams, there is time for feedback. In the assessments a part with intermediate feedback mentioned in	Students don't come up themselves with giving peer feedback.

	which it is explained how students can get intermediate feedback. Teachers stimulate discussion during their lesson.	Students are not obligated to form learning objectives themselves during the lessons or the learning teams.
Meaningfulness	Students think that the assessments are useful for their (future) workplace.	Some students think the assessments are to generic instead of the specific context they have on their workplace.
Reproducibility of decisions	Different backgrounds are used in calibrating. Assessors need to get their BQE and do a portfolio assessment training.	Assessors assess alone, when it is insufficient a second assessor is asked. Different backgrounds are not used consciously. Assessing is done using one work context of the students.
Transparency	Students come with questions about the rubric and the assessments, and these are answered.	Sometimes the work place of the students assesses different than the assessors of WIndesheim. When teachers accompany a learning team of a module they do not teach in, they do not know all the specific information about the assessments made.