Performance Assessments in Vocational Education:

A Qualitative Follow-Up Study

Thomas Stege

University of Twente, Faculty of Behavioural Sciences

Abstract

This study concerns the preconditions necessary for fair appraisal of performance assessments. There is a focus on the number of assessors needed, the type of assessors who would be best for the job, and to what extent they need to be acquainted with the student. The training someone would need to become a better assessor is also researched. This qualitative study augments earlier results from a systematic review. Theory is used as a basis, and two direct interviews, one panel conversation and a number of e-mail interviews are used to identify stakeholder's opinions on this matter. Experts agree that more than one assessor is ideal, and that a small-scale training is sufficient. However, there is some disagreement concerning the ideal type of assessor, objective or subjective, and the needed acquaintance level. A multidisciplinary rating team seems ideal, but stakeholders' disagreement may be too strong to maintain this.

Performance Assessments in Vocational Education: A Qualitative Follow-Up Study[1]

Introduction

On the subject of performance assessment in vocational education, research has only just started. It has been pointed out that a different view on education in general has had its influences on assessment processes (Stege, 2013). Without completely restating what has already been mentioned, it is important to notice that the increasingly common performance assessments are an interesting case. Direct examination of practical skills involves a whole different assessing scheme than traditional knowledge tests. Generally, they are specified on the contents of the work someone should be able to perform, which offers a broad range of possibilities and some difficulties measuring it (Van Berkel & Bax, 2006, p. 209).

This research, however, does not focus specifically on the contents of the performance assessment, or on the exact criteria. Instead, it focuses on the assessors. It is still commonplace for students in a practical vocational training to be rated by a single supervisor. This is someone who saw the student work for a certain amount of time at a company, while training the student on the job. However, as stated in earlier research (Stege, 2013): *"According to [the Dutch inspectorate of education], it is ideal not to be acquainted with the student, because someone with former knowledge may be too biased. It is especially not recommended by them for the workplace trainer to be the assessor. […] They also recommend more than one examiner to rate the student, especially in complex situations."* (The Dutch Inspectorate of Education, 2009).

The educational inspectorate gives this pretty clear opinion on assessors with performance assessments in general in mind. It is plausible to assume this also holds for practical vocational training. Common in vocational education in the Netherlands, these vocational trainings include internships and learning-on-the-job situations. In both cases, the performance on a job should be assessed properly. As vocational education is the main supplier for the labor market (MBO Raad, 2009), and the needs of the labor market have become more important in the current educational system (Dochy, Segers & De Rijdt, 2002), vocational education is the main focus area of this study. For all types of vocational education, it is a great advantage if it is clear whether employees in training are doing a good job or not.

The guidelines presented by the inspectorate of education are important, because schools will have to keep them into account, but seemed counter-intuitive and lacked a theoretical framework at first (Stege, 2013). The framework from the systematic review by Stege (2013) was used as a basis for this study. Because a systematic review only focuses on theory from previous studies, a qualitative study was conducted in order to get more practical implications.

Research questions

In order to have a useful framework for further defining the area of research, four research questions were defined. These questions are focused respectively on objectivity of the assessor, acquaintance and affect, number of assessors, and training of assessors. The first three are related to the issues presented by the inspectorate of education, while the fourth became an important topic of research during the study's progression. Because the questions remain largely the same as in previous research (Stege, 2013), they will be restated here:

- Should performance assessments in an mbo setting be evaluated by objective assessors?
- Should objective assessors be acquainted with the student's work? If so, to what extent?
- Is it desirable and realistic to involve more than one assessor?

[1] The primary supervisor for this bachelor thesis is Marianne Hubregtse. The secondary supervisor is Bernard Veldkamp.

- Should assessors be trained to improve their assessment quality? If so, what kind of training?

Of course, these questions were already answered to a certain extent by Stege (2013) in a systematic review. Based on that study, it is hypothesized that a team of assessors should evaluate performance assessments, at least one of whom should be an objective, external assessor who is not acquainted with the student. Also, a small-scale training system should suffice.

This follow-up research includes an extra dimension to provide additional possibilities for finding plausible answers. Therefore, two extra research questions were added:

- What combinations of assessors are actually used for performance assessments in vocational education?
- Why are these combinations of assessors used?

By answering these questions, more light can be shed on the practical situations concerning number and type of assessors, so the answers to the first four research questions can be improved. Opinions from the field form valuable input for more accurate answers. Once it is established why a current practice takes place, it will be easier to give recommendations based both on theory and practice.

Theoretical framework

On the subject of number of assessors, most studies agree that more than one assessor is preferred or even necessary for accurate performance assessments (Eckes, 2008; Gulikers, Biemans & Mulder, 2009; Lievens, 2002; Williams et al., 2012a; Williams et al., 2012b). Different raters rate in a much too different way to be interchangeable (Eckes, 2008). Although the mood of an assessor does not seem to play a significant role (Calderon, 1998), an assessor's opinion on a single performance may vary over time (Clauser, Clyman & Swanson, 1999). It even matters whether an assessor is rating something individually or in comparison to something else (Hofstee, 1970). All this is why, by the words of Williams et al. (2012b), *"Multiple raters should rate each resident to control for rater idiosyncrasies."* Using more than one assessor improves the rating accuracy, although the exact amount of assessors is rarely specified.

On the matter of what types of assessors to use, and whether or not the assessors should be acquainted with the students, the opinions differ. Some researchers claim that being acquainted with a student's previous performance helps form a clearer opinion on their job (Freeberg, 1969; Kozlowski, Kirsch & Chao, 1986). Even personal knowledge about a student, 'affect', would have a beneficial effect on rating accuracy (Freeberg, 1969; Van Scotter et al., 2007). This would be the case because acquainted raters differentiate on assessment subjects more easily. They know more about the student and better capable of categorizing that knowledge (Van Scotter et al., 2007).

On the other hand, knowing too much about a student can be detrimental to rating accuracy (Govaerts et al., 2010; Levy & Williams, 2004; Lievens, 2002; Williams et al., 2012a). This produces the so-called 'halo effect', an affect-driven distortion (Van Scotter et al., 2007). Objective assessment criteria are important (Dawes, Faust & Meehl, 1989), and an assessor should be as objective as possible according to some research (Lievens, 2002). It doesn't help, though, that the halo effect even takes place when the assessor only has trivial knowledge about a student, such as physical attractiveness (Clifford & Walster, 1973) or handwriting (Greifeneder et al., 2012). Arguably the best theoretical solution for these issues is presented by Gulikers, Biemans & Mulder (2009). In order to optimize rating accuracy, a multidisciplinary assessor team is needed with at least one objective, external assessor and at least one trainer or supervisor.

When looking at assessor training, it is important to note that an extensive training system is probably not necessary, because novice trainers tend to increase their rating accuracy quickly (Lim, 2011). The contents of an assessor training could include information related to the assessment method (Gulikers, Biemans & Mulder, 2009; Lievens, 2002), intra-rater reliability (Clauser, Clyman & Swanson, 1999), and equal importance of assessment criteria (Eckes, 2008). It should also be noted that assessor training does not solve all problems with assessor differences; raters will never be completely interchangeable (Eckes, 2008).

Method

As stated before, a systematic review (Stege, 2013) made it possible to summarize the findings from literature. This qualitative study was conducted because detailed answers from specific stakeholders could provide input literature could not. In this case, a qualitative study is more relevant than a quantitative study, because the more specific data made it possible to hear direct and nuanced opinions from stakeholders in the field. Specifically, a phenomenological approach was taken, because the stakeholder ideas are important to this research in this stage.

Data collection and instruments

Several methods for data collection were used, because different stakeholders could be contacted in different ways, and had their own preference in ways to participate in this study. First, an e-mail was sent out to a group of people (about 40 people) who were active in the last year as assessors in an internship. All of these assessors are experienced with assessing students on an mbo level, in the context of a performance assessment. The participants can all be classified as subjective assessors (trainers or supervisors) who are acquainted with the students. The potential respondents were given three options to participate. They could apply for a panel conversation at a given location, they could have an interview at their own workplace, or they could decide to answer a list of questions by e-mail. In the end, the first option was removed, because not enough people were interested in a panel conversation outside of their own workplace.

Aside from the above, a panel conversation with objective, external assessors was arranged. For this, a spokesperson from ECABO was contacted[2]. This offers an interesting contrast between the two types of assessors included in this research.

For the respondents in the first group, subjective assessors, who wanted to answer by e-mail, a questionnaire with twelve questions was developed, most of them with subquestions. The questions were all based on the research questions and on the findings from literature. This questionnaire was sent to the assessors who wanted to respond by e-mail, and it was also used as a basis for the interviews with people from both assessor groups. For the interviews, some questions were formulated differently in advance; in order to maintain an orderly conversation, the questions could not all be read literally, but they were all used, because they could give relevant information. Questions regarding the general tasks of the company and the respondent were included in order to clear up the background. Also, if a reply to a question gave some interesting insights in the respondent's view on a different relevant topic, that topic was brought up accordingly and the order of the questions was changed. Some questions were altered for the panel conversation with the objective assessors, to prevent them from being inapplicable. Some translated examples of questions from the questionnaire are: "Can you describe how performance assessments are being judged right now? Who is the assessor?" and "What would you prefer: assessing by an assessor from an external company, or assessing by the internship supervisor?"[3]

After the data collection, the available responses were processed by comparing them to each other, and to the findings from the theoretical framework. For each question, it was determined whether it was part of an answer to a research question, and whether the answer was clear enough to be included. The results are presented in a way based on the research questions. Because the responses regarding these subjects tend to overlap, the sections 'types of assessors' and 'affect and acquaintance' were merged together.

[2] ECABO is a company specialized in several projects concerning vocational education, and their assessors rate students who come over for performance assessments.
[3] The full questionnaire is not included in this study, but can be requested to the author.

Respondents

Out of the people from the first group who were contacted (about 40 subjective assessors who were contacted by e-mail)[4], about ten people said they wanted to answer by e-mail, and in the end, this yielded five responses. One of those five responses could not be used in this study, because the respondent answered that too many questions were not applicable to their situation. Three respondents answered that they preferred a real life interview, but with one of them, attempts to arrange a meeting were unsuccessful. Therefore, this yielded two more responses for a total of six subjective assessors. The interviewees were both male, Caucasian and middle-aged. Specific sample descriptors of the other four participants are not known.

The panel conversation with objective assessors from ECABO was originally planned with three participants. One of them was unable to make it the day of the interview, and thus, two objective assessors were included in the study. They answered the questions together, and they were both female, Caucasian and middle-aged. A detailed description of all respondents, as far as possible, can be found in table 1.

Table 1

List of respondents and sample descriptors

Respondent #	Responded by	Assessor type	Gender	Company	Age
1	E-mail	Subjective	Unknown	Unknown	Unknown
2	E-mail	Subjective	Unknown	Unknown	Unknown
3	E-mail	Subjective	Unknown	Unknown	Unknown
4	E-mail	Subjective	Unknown	Unknown	Unknown
5	Interview (one on one)	Subjective	Male	Aldipress B.V.[5]	About 55
6	Interview (one on one)	Subjective	Male	Duifhuizen Lederwaren[6]	About 50
7	Panel conversation	Objective	Female	ECABO	About 55
8	Panel conversation	Objective	Female	ECABO	About 40

During the interviews, all relevant information needed for an answer to any of the research questions and relevant background information was written down. These notes were summarized by the researcher, and then used as a basis for determining the results of this study. A direct coding scheme was not used, but the available information was compared to each of the research questions.

[4] The assessors were contacted by Marianne Hubregtse, from her database of assessors who participated in previous research by her.

[5] Aldipress B.V. is a distribution center for magazines, situated in the Dutch town of Duiven, Gelderland.

[6] Duifhuizen Lederwaren is a wallet, purse and belt shop, situated in the Dutch town of Oud-Beijerland, Zuid-Holland.

Results

A global overview of the results that followed from this research can be found in table 2. The respondent numbers correspond to the numbers in table 1, and are shown in that order.

Table 2

Overview of research data

Respondent #	Current situation	Preferred number of assessors	Preferred assessor type	Affect and acquaintance	Assessor training; points of interest
1	Multiple types of subjective assessors	More than one	Multiple types	Negative effect for affect; positive for relevant acquaintance	Clear working method; getting students started
2	Unclear	More than one	Supervisor	No significant effect	Clear feedback
3	Multiple types of subjective assessors	Sometimes one, for financial reasons	Supervisor	Negative effect	No specific points of interest
4	Only one subjective assessor	More than one	Supervisor	No significant effect	Assessing on different levels
5	Multiple types of subjective assessors	More than one (should be mandatory)	Multiple types	No significant effect	Clear expectations and feedback
6	Only one subjective assessor	Only one, except in rare cases	Supervisor	Negative effect for affect; positive for relevant acquaintance	No assessor training at all
7	Multiple objective assessors	Two	External assessor	Negative effect	Objective reasoning
8	Multiple objective assessors	Two	External assessor	Negative effect	SMART assessment criteria

Number of assessors

From literature, it has become clear that an assessment situation ideally has more than one assessor (e.g. Eckes, 2008; Lievens, 2002). Five out of eight assessors (respondents 1, 3, 5, 7 and 8) who provided insight in practice point out that this is already the case. The respondents themselves are one of the assessors. For the subjective assessors, this means that they are fulfilling the role of both trainer and assessor. The second assessor is usually a teacher or supervisor from school (respondents 1, 3 and 5 agree). After the performance assessment, there would be discussed how to evaluate the student's

performance. For objective assessors, it just means that multiple objective assessors are used (respondents 7 and 8 agree).

However, two respondents pointed out that they are actually the only ones who assess their interns (respondents 4 and 6) in a specific performance assessment. These respondents did not say that they see direct problems with this situation, however, one of them (respondent 6) did say that it remains to be seen whether an appraisal like this is really valuable[7]. The other respondent (respondent 4) would prefer to see more than one assessor. Both respondents who are usually the only assessor, and even two respondents who are not (respondents 3 and 5), say that they would prefer to be in closer contact with the schools. Even if schools send someone to be an assessor at a performance assessment, they often do not have anything to do with the internship process.

Out of eight respondents, seven (all but respondent 6) see at least some advantages in using more than one assessor. One of them is very clear about this: in all performance assessment situations, it should be compulsory to have more than one assessor[8]. The others have a more nuanced view about this, but are certainly in favor of multiple assessors. Three respondents (respondents 1, 7 and 8) mention the possibility of a completely objective, external assessor, as described in the theoretical framework (e.g. Lievens, 2002). The one respondent that does not see an advantage in getting involved with a second assessor claims that this is mostly because of his specific situation. The work that his students are performing – for example, speaking to clients in the shop – is in his opinion too general in nature to be able to assess based on a single performance.

Reasons some respondents mention for preferring multiple assessors agree with the theoretical views. For example, two respondents (7 and 8) acknowledge that every assessor rates according to their own norms and values, although it is pointed out that they were trained to be as objective as possible. This is also mentioned in literature by Eckes (2008), who points out that assessors are not interchangeable. The respondents (7 and 8) also point out that there would be another, more practical problem with a single assessor. In their assessment situations, usually a larger number of students (typically a maximum of eighteen) gather in a simulation of a workplace situation. When one assessor would have to assess all eighteen of them, they would get too little time to see how each student works. If a smaller number of students are present in the simulation, then a single assessor could work. However, this is only a plausible idea if there is at least one other assessor who takes a look at the final results. This is necessary to maintain a consistent rating.

There appear to be no significant practical problems when assessing with multiple assessors. No respondents recalled specific cases in which assessors strongly disagreed with each other when coming to a conclusion about the assessment. Instead, one respondent (respondent 5) claims that it is usually fairly easy to come to an agreement when the results a student gets seem to contradict themselves at first. Other respondents gave similar answers.

Financial and practical issues do not seem to play an important role for four out of five respondents who see an advantage in using more than one assessor. One respondent (respondent 3) said that schools usually have very limited time to visit a company where an internship takes place at all[9]. It would be unrealistic to assume that they would be present with every examination, because teachers and supervisors at schools tend to have other things to do. However, this does assume that the second assessor would be someone from the school. The respondent did not give an opinion about an objective second assessor from an external source.

---

[7] Quote (in Dutch): “Ik zet wel mijn vraagtekens bij de waarde die deze beoordeling heeft”

[8] Quote (in Dutch): “Het moet in alle examensituaties verplicht zijn om meerdere beoordelaars te hebben.”

[9] Quote (in Dutch, verbatim from questionnaire): “De contacten in het algemeen vind ik erg weinig. (zeker bij snuffelstages) Veelal komen ze vooraf , en bellen bv. maar 1 keer. Zeker voor de leerling is dit te weinig, maar ook om meer inzicht te krijgen hoe de leerling functioneeert.”

In the end, there are enough reasons to assume that a performance assessment should have multiple assessors. In addition to the theoretical basis (e.g. Williams et al., 2012a; Williams et al., 2012b), most assessors agree that there are significant advantages to using more than one assessor. Even if they do not, they admit that the value of the assessment may be flawed.

Types of assessors, affect and acquaintance

In spite of research suggesting that at least one objective assessor is a good idea (e.g. Clifford & Walster, 1973; Greifeneder et al., 2012), the subjective assessors in this study (respondents 1 through 6) have no experience with an external assessor. Everyone involved in the assessment process has got at least some previous knowledge about the students. The assessors from the schools tend to have at least some personal knowledge and are also acquainted with previous results of students. These results may or may not be relevant to the constructs measured in the performance assessment. Also, the assessors from the company have followed their students during the internship or work process.

Although the subjective assessors lack experience with objective assessors, at least one of the respondents (respondent 1) can see the benefit of including them in the assessment process[10]. When asked whether an external assessor or an assessor from the own company is preferred, this respondent says that both types of assessors are possible, but the external one will be more neutral. Another respondent (respondent 5) points out that someone from the internship company would know the student better, but an external assessor may provide some additional insight.

Some subjective assessors acknowledge some form of halo effect. However, two out of six respondents (respondents 1 and 6) relate this to situations where an assessor has to rate a student they already knew in person before there was an assessor-student relationship. This could be someone from their own family or neighborhood. They generally do not feel it is a disadvantage to have previous knowledge about a student, and when it's relevant to the assessment, they feel that it's actually an advantage. Two others (respondents 2 and 5) imply that there is no such thing as the halo effect at all, and that assessors should always be 'objective', even if they know the student. One respondent claims that someone who is acquainted with the student would be a better assessor, yet also says that someone acquainted is generally more negative about a student's capabilities. They say that negative experiences with student behavior influences their opinion, but assessors are sometimes surprised by how capable these students turn out to be. Something similar was described by Van Scotter et al. (2007).

When asked whether someone from the school should be involved as an extra assessor, considering that schools are always responsible for student's assessments in the end, all six subjective assessors (respondents 1 through 6) answer positively. As stated before, assessors from internship companies tend to be unsatisfied with the amount of involvement from schools. It could be summarized that subjective assessors generally want more involvement from another type of subjective assessor.

It should be noted that no subjective assessor who responded to the questions thought that affect was too destructive for rating accuracy. Some of them said that adding an objective assessor could be worthwhile, but none of them implied that an objective assessor is generally superior to a subjective assessor. However, the objective assessors (respondents 7 and 8) themselves did lean this way. According to their experience, objectivity is way too important and the halo effect way too strong for subjective assessors to play an important role in the decision making process. This opinion is somewhat comparable to the one presented by Lievens (2002), who discourages subjectivity in assessors altogether.

According to the objective assessors, a performance assessment rating by someone from the internship company generally does not work at all. In rare circumstances, it could be possible to rate a student in the internship company, but only if the company includes a structured, special simulation

[10] Quote (in Dutch, verbatim from questionnaire): "Kan beide een externe is neutraler [...] Ja misschien een externe erbij om te kijken als er goed beoordeeld gaat worden."

environment. If someone who has trained the student throughout the working process becomes an assessor, there is a large risk of rating according to the standards of the company. According to them, these are not objective enough and do not provide a good measuring scale of a student's performance. This does not seem to be supported by literature, though.

It has also been stated by respondents 7 and 8 that contact between someone from school and an objective assessor should be very limited. When a group of students come in for a performance assessment, some teachers tend to declare to the assessors in advance which students are best. This is not a good idea, because it undermines the objectivity of the assessors, and therefore has a negative effect on the rating accuracy.

When asked if the fact that the students are assessed by only a single performance causes any problems, the objective assessors deny. According to them, the only problems that could arise include sickness, personal issues or severe exam anxiety. In case of sickness or personal issues, these problems can be solved by simply choosing another date for the assessment. The objective assessors (respondents 7 and 8) say that severe cases of exam anxiety are very rare, and call for solutions unrelated to this specific assessment environment. Milder cases usually solve themselves. Even though the company simulates workplace situations, stress factors are diminished by making the students feel at ease.

All in all, it seems that in spite of the agreement on the topic of multiple raters, there is a huge gap between the opinions of subjective and objective assessors when discussing this matter. The fact that all assessors tend to value their own work higher than someone else's work probably plays an important role. Nevertheless, this is an interesting dilemma presented by both types of assessors, that subjective assessors know the company and the student better but objective assessors are not as bothered by the halo effect. This dilemma is also supported by theory (e.g. Van Scotter et al., 2007).

Training of assessors

When the subjective assessors were asked whether or not they had some kind of training to be an assessor, five out of six respondents answered positively (respondents 1 through 5). Four of those five have had the training because their executives or someone else in the same company thought that would be a good idea. The fifth (respondent 4) simply wanted to improve their own capabilities as an assessor. The respondent that never had training (respondent 6) does have about ten years of experience as an assessor, and also as a supervisor of other employees. According to him, it was simply never offered by his executives, and he points out that there are no complaints about his assessor skills.

Considering the contents of assessor training, several aspects are mentioned. Both types of assessors were asked what they remembered from the training, and what aspects of the training they still use in their everyday job as an assessor. Some things named by the subjective assessors included: getting the students started immediately, using a planned approach, rating students on different levels of education, giving elaborate feedback, and clearing up expectations in advance.

The objective assessors (respondents 7 and 8), who both had training, mentioned the importance of objectivity again, and they find it important to have their assessment criteria formulated SMART (Specific, Measurable, Attainable, Relevant, Timely). They also point out that there is a system that is somewhat comparable to the BIG-register system in Dutch health care. In this system, doctors and other people with medical professions need to follow courses every once in a while to keep their knowledge and practical skills updated.[11] For objective assessors, something comparable also exists,

[11] The BIG-register (Beroepen in de Individuele Gezondheidszorg; Professions in Individual Health Care) is a Dutch database, in which several types of Dutch medical professionals, such as pharmacists and psychiatrists, are registered. Only people registered in this database are allowed to perform this profession. This guarantees that all of these people are actual professionals. In order to maintain

and it means that they need to renew their certification once every two years. This seems to be a thorough and very professional way to make sure that assessors are qualified for the job.

oneself in the database, someone must follow periodic courses to keep their knowledge and skills updated.

Discussion

Based on previous research (Stege, 2013), it was hypothesized that multiple assessors would be ideal in a performance assessment situation. It was also hypothesized that an ideal assessor team would be composed of at least one objective assessor – someone with no former acquaintance of the student – and at least one subjective assessor, such as a trainer or supervisor from the internship company. Finally, it is hypothesized that an assessor training can be useful, but a long and intense training is probably not necessary.

From the qualitative study, it was clarified that no significant objections to a multitude of assessors arise. Researchers on one hand, and objective and subjective assessors from the field on the other hand, generally agree that it is a good idea to rate students in an assessor team and not individually. There are several reasons for this. The main reason seems to be that not every assessor rates the same way, nor do they rate flawlessly, no matter how much training they have. Since the risk of overestimation is about as high as the risk of underestimation, using multiple assessors helps to average these effects and solve the problem of assessor differences. There are some other reasons, such as impracticalities of rating alone, or wanting different points of view in the assessment process. Conclusively, it is recommended to use more than one assessor in performance assessments. In practice, there are cases where only one assessor is used. This could lead to a low rating accuracy, as students who should pass may be getting a lower grade than they should, and students who should not pass may be getting a higher grade.

As in literature, the other issues turn out to be much more complicated. The dilemma as presented by Van Scotter et al. (2007), between a subjective assessor who suffers from the halo effect and an objective assessor whose rating is based on an insufficient amount of data, lingers. Research suggests that there is a huge gap between the opinions of objective and subjective assessors. Both types of assessors are clearly under the impression that their job is more important than the other assessor type's job. Although some subjective assessors do not object to an external assessor helping out, they are generally conservative on this matter. Objective assessors, on the other hand, do not feel that subjective assessors are suited for the job at all.

From literature (e.g. Gulikers, Biemans & Mulder, 2009), it seemed to be a good idea for objective and subjective assessors to cooperate. They would have to discuss their findings from the performance assessment after it has been completed. There is no evidence that anything close to this happens at all, and thus, it is not clear if something like this actually works. Therefore, an experiment involving a setup with a multidisciplinary rating team may be a very interesting idea for follow-up research. The skepticism about the quality of subjective assessors by objective assessors probably does not help this case. However, with no clear evidence against a cooperation of this type, for now this can still be recommended.

The idea that a small-scale training is sufficient can be confirmed to an extent. Most assessors recall some important information from their own training, even if it was only a small-scale training and if it was some time ago. More research on the exact contents of an assessor training is necessary before a definitive statement can be made. Something interesting was mentioned by the objective assessors, though. A system in which assessors need to update their knowledge on a regular basis to be qualified for the job, akin to the BIG-register system in Dutch health care, could be an interesting option. This could also be used for subjective assessors, as they need to know the same basics. A small-scale training in this system will probably still be sufficient, and an extensive training is probably not realistic due to time and financial reasons anyway.

A limitation of this research is the types of respondents involved. It may be the case that the respondents could not form an honest opinion on this matter because they were rooting for their own team. More elaborate qualitative research with different stakeholders, for example from the educational inspectorate, from schools or from the management level in companies, may address these issues. The opinions from students themselves may also be included in follow-up research, as they are stakeholders in the assessment process as well. More conclusive answers to the research questions

could be achieved this way. Another idea for follow-up research, related to this, is to study the bias in favor of an assessor's own profession.

Another limitation is the number of respondents. During the course of this research, not many respondents took the time to answer by e-mail or make an appointment. With a larger sample size, more relevant examples could have been found and more conclusive findings could have been presented. The sample size increase could also be achieved by finding more objective assessors, for example in other companies than ECABO. It could very well be the case that a type of objective assessment is present in another professional setting.

References

Calderon, R.F. (1998). *Mood and performance appraisal quality.* (Doctoral dissertation). The Ohio State University.

Clauser, B.E., Clyman, S.G., & Swanson, D.B. (1999). Components of Rater Error in a Complex Performance Assessment. *Journal of Educational Measurement, 36(1),* 29-45.

Clifford, M.M., & Walster, E. (1973). The effect of physical attractiveness on teacher expectations. *Sociology of Education, 46(2),* 248-258.

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science, New series, 243(4899),* 1668-1674.

The Dutch Inspectorate of Education. (2009). Examinering in de beroepspraktijk: veel gestelde vragen en antwoorden. Retrieved from http://www.onderwijsinspectie.nl/binaries/content/assets/Documents+algemeen/2009/Examinering+in+de+praktijk+-+FAQ.pdf, October 23, 2012.

Eckes, T. (2008). Rater types in writing performance assessments: A classification approach to rater variability. *Language testing, 25,* 155-186. Doi: http://dx.doi.org/10.1177/0265532207086780

Freeberg, N.E. (1969). Relevance of rater-ratee acquaintance in the validity and reliability of ratings. *Journal of Applied Psychology, 53(6),* 518-524.

Govaerts, M.J.B., Schuwirth, L.W.T., Van der Vleuten, C.P.M., & Muijtjens, A.M.M. (2010). Workplace-based assessment: effects of rater expertise. *Advances in Health Science Education, 16,* 151-165. Doi: http://dx.doi.org/10.1007/s10459-010-9250-7

Greifeneder, R., Zelt, S., Seele, T., Bottenberg, K., & Alt, A. (2012). Towards a better understanding of the legibility bias in performance assessments: The case of gender-based inferences. *British Journal of Educational Psychology, 82,* 361-374. Doi: http://dx.doi.org/10.1111/j.2044-8279.2011.02029.x

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

Hofstee, W.K.B. (1970). Comparative vs Absolute Judgments of Trait Desirability. *Educational and Psychological Measurement, 30,* 639-647.

Kozlowski, S.W.J., Kirsch, M.P., & Chao, G.T. (1986). Job Knowledge, Ratee Familiarity, Conceptual Similarity and Halo Error: An Exploration. *Journal of Applied Psychology, 71(1),* 45-49.

Levy, P.E., & Williams, J.R. (2004). The Social Context of Performance Appraisal: A Review and Framework for the Future. *Journal of Management, 30,* 881-906. Doi: http://dx.doi.org/10.1016/j.jm.2004.06.005

Lievens, F. (2002). Trying to Understand the Different Pieces of the Construct Validity Puzzle of Assessment Centers: An Examination of Assessor and Assessee Effects. Journal of Applied Psychology, 87(4), 675-686. Doi: http://dx.doi.org/10.1037//0021-9010.87.4.675

Lim, G.S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing, 28,* 543-561. Doi: http://dx.doi.org/10.1177/0265532211406422

MBO Raad. (2009). Dutch vocational education and training and adult education. Retrieved from http://www.mboraad.nl/?page/530112/About+us.aspx, January 11, 2013.

Stege, T.A.M. (2013). *Appraisal of Performance Assessments in Vocational Education.* Unpublished bachelor thesis, University of Twente, Enschede, The Netherlands.

Van Berkel, H., & Bax, A. (2006). *Toetsen in het hoger onderwijs* (2nd ed.). Houten, Netherlands: Bohn Stafleu van Loghum.

Van Scotter, J.R., Moustafa, K., Burnett, J.R., & Michael, P.G. (2007). Influence of prior acquaintance with the ratee on rater accuracy and halo. *Journal of Management Development, 26(8),* 790-803. Doi: http://dx.doi.org/10.1108/02621710710777282

Williams, R.G., Sanfey, H., Chen, X., & Dunnington, G.L. (2012a). A Controlled Study to Determine Measurement Conditions Necessary for a Reliable and Valid Operative Performance Assessment. *Annals of Surgery, 256(1),* 177-187. Doi: http://dx.doi.org/10.1097/SLA.0b013e31825b6de4

Williams, R.G., Verhulst, S., Colliver, J.A., Sanfey, H., Chen, X., & Dunnington, G.L. (2012b). A template for reliable assessment of resident operative performance: Assessment intervals, numbers of cases and raters. *Surgery*, *152*, 517-527. Doi: http://dx.doi.org/10.1016/j.surg.2012.07.004