General discussion
Background
This thesis was initiated to explore how assessment can best be deployed in order to optimise postgraduate medical training. When the research project began, the Dutch Legislative Board of Medical Specialists (CGS) had just called for a vast increase in the number and variety of assessment moments. Loosely structured mastery-apprentice style learning had given way to competency-based, outcome-focused training. Subsequently, a huge influx of formative assessment moments had to ensure that postgraduate training became as efficient and transparent as possible by combining feedback to optimise learning with a measurable learning curve. Examples of formatively used assessment instruments include knowledge progress testing, the mini-clinical examination exercise (mini-CEX, KKB/KPB), direct observation of procedural skills (DOPS), objective structured assessment of technical skills (OSATS), multisource feedback (MSF) and the portfolio. However, even though they are theoretically sound and have an apparent face-validity, at the time of introduction there was little evidence available regarding the actual educational impact of these assessment instruments.

In this chapter we will explain how the previous chapters are interlinked. Furthermore, the general picture that emerges from the research findings and the implications for postgraduate medical training will be discussed. We conclude the chapter by pointing out the direction of future research on the subject of how assessment can be used to optimise postgraduate medical training.

Thesis
The research project started in 2004 with an evaluation of the utility of postgraduate formative knowledge progress testing. The Dutch Legislative Board of Medical Specialists (CGS) had just decreed that all postgraduate training programmes should introduce mandatory formative knowledge progress testing as a means of assessing the knowledge level and knowledge growth of trainees. This assessment instrument was carefully chosen for its feedback potential. Testing the knowledge of trainees in relation to the required knowledge level at graduation was expected to yield feedback on strengths and weaknesses and provide insight into individual learning curves without inciting exam-driven study behaviour. Formative knowledge progress testing was introduced in Dutch postgraduate O&G training several years prior to the official restructuring of postgraduate assessment by the CGS. Therefore, there was a significant amount of data available for analysis. We set out to evaluate the progress test by posing the following research question:

What is the utility of formative knowledge progress testing in postgraduate obstetrics and gynaecology?

The utility was determined by evaluating the reliability and construct validity (evidence of knowledge growth over training years) of postgraduate progress testing and the acceptability and educational impact of postgraduate progress testing. Chapter 2 and 3 report the findings of these studies. The reliability was determined by calculating Cronbach’s alpha, which ranged between 0.5 and 0.8. This is barely acceptable, even for formative assessment. To determine the construct validity the differences in test scores between training years were tested for significance. We found evidence for construct validity only for the lower
training years. To learn more about the acceptability and educational impact of postgraduate knowledge progress testing we devised a questionnaire directed at both trainees and educational supervisors. Acceptability, the degree to which the assessment instrument meets the needs of and is approved by stakeholders, was good, though trainees questioned the authenticity of test content. The perceived educational impact of the test, that is, the effect of the test on learning, was limited.

As we hypothesised that improving the authenticity of the test might improve its reliability and validity, we designed a sequel study that tested the addition of time-restricted internet access on the utility of postgraduate knowledge progress testing. Trainees frequently commented that the progress test contains too many factual questions and is not representative of daily practice because nowadays most doctors have easy access to high-speed internet. They claimed that, considering information can quickly be found in daily practice, the role of memorisation has become less important. In chapter 4 we report on our pre- and post intervention study that we designed to answer the next research question:

+ Can the utility of formative knowledge progress testing in postgraduate obstetrics and gynaecology be advanced by improving the authenticity of the test?

We found that time-restricted internet access during the progress test does increase test scores and the perception of authenticity of the test. However it does not have a measurable effect on ranking or construct validity or reliability. We also found that accessing the internet costs time, as trainees changed an average of 8 questions in 30 minutes, of which 6 were changed to a correct answer and 2 to an incorrect answer. As time is not indefinitely available in daily practice, this finding demonstrates that memorisation is still necessary, even in the internet age.

Finding a disappointing educational impact from a theoretically well-wrought assessment instrument made us curious about the perceptions of trainees and educational supervisors regarding the subject of assessments’ ability to improve postgraduate medical training. It is an axiom that acceptability is vital for the successful implementation of new assessment instruments. This is even more true for assessments that are to be used to optimise learning and training. For this reason we designed a qualitative study in which we used focus groups, or semi-structured group discussions, with trainees and educational supervisors to answer the subsequent research question:

+ What factors determine active engagement in formative assessment of postgraduate trainees and supervisors?

The results of this study are conveyed in chapter 5. In the discussions the individual perspectives of trainees and supervisors on feedback emerged as a central theme. Whether the trainee and/or supervisor initiates and pursues a feedback cycle including assessment is largely determined by ownership as expressed by the trainee and the achievement goal orientation (mastery vs. performance) of both trainees and supervisors. Whereas people with a mastery-goal orientation tend to focus on acquiring and developing competence, welcoming all feedback as an opportunity to improve performance, the focus of people with
a performance-orientation tends to be on demonstrating one’s competence and out-performing others. The latter usually valuing the clear standards and consequences of summative assessment. Furthermore, the perception of the credibility of feedback and/or feedback-givers appeared to play a pivotal role in the educational impact of feedback. A supportive learning environment consisting of dedicated supervisors, dedicated teaching time and clear assessment procedures with defendable standards and consequences for substandard performance was frequently mentioned as an important facilitator.

As we hypothesised that formative assessment could play an important role in determining which professional activities can be entrusted to a trainee, we also included a question on this subject during the focus groups. The discussions that ensued provided such rich data on the subject of becoming competent and progressive independence that we decided to report these findings in a separate paper, which is included here as chapter 6. The question that was put forward was:

What is the role of assessment in assessing competence and determining progressive independence?

From the discussion it became clear that determining the level of competence of a trainee and the degree of independence that is allowed is a largely implicit process. Moreover, the degree of independence is usually determined for each procedure again by both trainee and supervisor. Central factors in the determination process are the self-assessment and self-efficacy of the trainee, the number of performed procedures, patient safety issues and the ‘audacity’ of the supervisor. It was remarkable that both trainees and supervisors were apprehensive regarding formal assessment of competence preceding more independent practice. They felt the result of one assessment is not representative and consequentially not generalisable to future performance. Furthermore, trainees feared being declared competent too soon and then being left alone in situations where they would have preferred supervision.

The research project was concluded by a systematic review of contemporary literature on the educational impact of assessment. The review was inspired by the fact that the findings in both our primary research studies—the evaluation of the utility of knowledge progress testing and the exploration of perceptions on assessment for learning—fell in line with our experiences from daily clinical practice: the effectiveness of commonly used formative assessment instruments (e.g., the mini-CEX, DOPS, OSATS, MSF, knowledge progress testing and portfolio) appears to be very variable and largely dependent on individual inclinations. This made us curious about the current state of affairs regarding the educational impact of postgraduate assessment, as more than a decade has passed since the first introduction of new assessment instruments. With the review we tried to answer the following question:

What is the evidence considering the educational impact of both formative and summative assessment?
In chapter 7 we present the findings of the literature review. The literature indicates that it is possible to perform high quality research that is capable of detecting a genuine impact of assessment on learning. There is empirical support for the use of several assessment instruments; however, no evidence beyond Kirkpatrick level 1 (the perception of trainees and/or supervisors) can be found for the mini-CEX, DOPS, OSATS and the portfolio, at least not in the way they have been implemented in daily practice. By performing a realist synthesis we uncovered the following characteristics shared by the assessment instruments that have a genuine effect on learning:

- Trainee performance is compared to a gold standard (guidelines, licensing examinations, experts)
- Sufficient time is assigned for structured feedback
- Feedback is part of a feedback cycle

**Meaningful formative assessment**

This research started with the following question: How should assessment be deployed in order to optimise postgraduate medical training? Our research has provided us with the following insights into how a measurable educational impact can be achieved using formative postgraduate assessment:

1 **‘Feedback needs comparison to an agreed-upon standard’**

Feedback in clinical education is defined as ‘specific information about the comparison between a trainee’s observed performance and a standard, given with the intent to improve the trainee’s performance’ [1]. One of the major challenges of formative assessment in clinical education is the lack of a ‘gold’ standard. For the majority of patient problems no definite guidance is available. This implies that there is still a lot of room for authority/experience-based medicine. Moreover, since trainees have alternating supervisors, sometimes on a day-by-day basis, it is understandable why feedback is often conflicting in the perception of trainees. Examples of agreed-upon standards are national or local guidelines, protocols, and involving experts in the assessment of the trainee. For those areas for which no standards are available as yet, a defensible alternative would be ongoing discussions between supervisors and trainees, addressing all aspects of clinical and professional decisions [3].

2 **‘Feedback should be part of a feedback cycle’**

Even when the trainee judges feedback to be credible, it does not automatically result in learning. Learning requires that the feedback be part of a continuous feedback cycle [22,23]. A feedback cycle starts with the assessment of the performance of a trainee by comparison to an agreed-upon standard, following which feedback is provided. Subsequently, trainee and supervisor work out a plan of action for improvement and the trainee gets the opportunity to practice. Next, the trainee is assessed again on the same subject. A feedback cycle serves a dual purpose: it becomes much more important for the trainee to act on the feedback (overcoming the individual susceptibility to feedback), but it also provides information on the impact on training that was actually achieved. Assessment and feedback given outside a feedback cycle should be seen as merely snapshots of learning [24]. Important as they may be to guide learning in the workplace, if given in isolation with no clear follow-up or consequences, it will be hard to establish a measurable educational impact.
To counterbalance the effects of the missing gold standard and the tendency to remain stuck in feedback-snapshots we strongly recommend the re-valuation of master-apprentice-style training. Long term commitment between trainees and supervisors enables ongoing discussions on the ‘gold’ standard and reasons for choosing a specific diagnostic or therapeutic approach. This makes it much harder for trainees to just ‘shrug off’ feedback that they feel is not credible or is not in accordance with their self-assessment. Furthermore, it becomes possible to complete the feedback cycle one or more times, resulting in a measurable learning curve.

Certifying exams
When the CGS called for far-reaching changes in the assessment structure of postgraduate medical training she refrained from introducing certifying exams. This is in line with the Dutch academic tradition within which the freedom to learn, freedom to teach and a strong emphasis on individual responsibility are key elements. Consequentially, this has resulted in a strong aversion to high stakes testing, especially in higher education. The assumption is that this type of testing will invoke exam directed study behaviour and that the results of certifying exams are trivial, representing only a snapshot of trainee performance, and cannot be extrapolated over time. As trainees are seen as highly motivated adults, who need to be prepared for life-long learning, including numerous formative assessment moments in postgraduate training is believed to better support their needs. Exposing trainees regularly to formative assessment provides feedback for trainees, to guide further study activities, but also helps supervisors determine the trainee’s level of performance. At the same time, the freedom to learn and teach and the individual responsibility remain intact because of the non-committal nature of the feedback.

However, our studies also highlight the fact that there is a downside to giving trainees too much freedom and individual responsibility. We find that without a gold standard and without clear consequences for substandard performance trainees are at liberty to use feedback as they see fit, resulting in a non-measurable educational impact. One can argue that judging feedback, and having the freedom to act on it, is exactly what is necessary to prepare trainees for a career of lifelong learning and reflection. However, the non-committal nature of this way of receiving feedback may also lead to superficiality and misplaced confidence in one’s own insight. It is well known that physicians’ ability to accurately self-assess is limited. For this reason we strongly advice that high stakes assessment should become part of Dutch postgraduate medical training.

One way of raising the stakes of postgraduate assessment is by introducing more summative assessment, for example, in the form of certifying exams. The assumption that these exams result in unwanted study behaviour is only supported by a small number of studies that involved medical students. Moreover, our systematic review clearly shows that certifying exams are associated with better patient care. However, if certifying exams are to be introduced in Dutch postgraduate training, it is of paramount importance that they concentrate on testing relevant knowledge and skills, and that they are of sufficient validity and reliability. This is important, partially because the exams will require a considerable...
time-investment of the trainee, but foremost because both trainees and society deserve a reliable and defendable assessment structure.

There are other ways to increase the stakes of postgraduate assessment, for example, by facilitating exclusive trainee-supervisor relationships of sufficient length. Such relationships will allow trainees and supervisors to engage in ongoing feedback cycles, which can effectively make up for the lack of gold standards in clinical practice. Another way of raising the stakes would be to precede progressive independence in professional activities with summative assessment, involving the demonstration of knowledge and skills. Preferably, this assessment is combined with benchmarks indicating where and when a trainee must achieve a set of entrusted professional activities during his training.

**Recommendations for optimising Dutch postgraduate assessment:**

- Establish a healthy mix of formative and summative assessments.
- Continue workplace-based assessment but ensure that it is firmly embedded in a system that facilitates its use in a way that can have a genuine impact on postgraduate training.
- Entrusted professional activities should be preceded by assessment of knowledge, skills and performance.
- Combine knowledge progress testing with summative modular testing or certifying exams.
- Use portfolios for the collection of evidence of competence, but not for the assessment of competence.

**Strengths and limitations**

This research project has several strengths. First is its focus on the relation between assessment and learning. Although much has been written on the potential benefits of formative assessment and the drawbacks of summative assessment in medical education, little empirical research has been conducted to support these claims. Furthermore, the evaluation of assessment instruments tends to focus on validity, reliability and feasibility issues. However, if the aim of assessment is to stimulate learning, educational impact and acceptability must also be addressed. This is especially important because novel assessment instruments require a considerable investment of both trainees and supervisors. Another strength is that we sought to triangulate our findings by including both trainees and educational supervisors in the study on progress testing and in our focus groups. Thus the subject of learning assessment was evaluated from different perspectives, adding strength to our findings. We also believe that a significant strength is the fact that the research was performed by several clinicians, who are actually involved in the training of trainees, in cooperation with a medical educationist. The current thesis underlines the fact that educational interventions that may appear advantageous on paper do not always yield the anticipated effect when employed in daily practice. Therefore it is important that research in clinical education is a joint effort of clinicians and behavioural scientists/medical educationists. Last but not least, even though we have only solved a small part of the intricate puzzle of how postgraduate medical training can be optimised to suit the needs of trainees, supervisors and society, we managed to get a good idea of how assessment should be deployed in order to optimise postgraduate training.
A major limitation of this study is that we have only investigated the educational impact of assessment at Kirkpatrick level 1, or the perceptions of trainees and/or supervisors. Even though it is questionable whether randomised controlled studies are the only way to obtain high quality data, there is a great need for research on the impact of educational interventions that moves beyond Kirkpatrick level 1 and looks for a genuine effect on learning that is solely attributable to the assessment instrument under investigation. Furthermore, perhaps we should have performed the systematic review on ‘educational impact of assessment in postgraduate medical training’ in the first stages of the research project. This may have given us more guidance on how to set up studies investigating the educational impact of assessment. However, many studies that have been included in the review were only published after the research had commenced.

**Future research**

Future research should concentrate on establishing standards in the complex realm of clinical education and enabling and optimising feedback cycles. Additionally it is important to perform research with sufficiently large groups of trainees, who represent several different medical specialties. Moreover, most studies so far have been performed with volunteers. Since we have established that there is an individual susceptibility for feedback that determines whether a trainee and/or supervisor will initiate and complete a feedback cycle, it is well possible that volunteers are not representative of the whole group of postgraduate trainees and supervisors. In other words, it is uncertain whether positive study results can be extrapolated one to one to the realm of daily practice. On the other hand, it is probably unethical to perform randomised controlled studies during postgraduate training, especially when the time investments are disproportionate between groups or the stakes are high.

Finally, it is important to move beyond Kirkpatrick level I (stakeholders’ reactions) studies. Even though it is indisputable that user acceptability is pivotal for the successful implementation of educational intervention, once it is demonstrated which assessment instruments are able to achieve a significant educational impact it should be possible to motivate trainees and supervisors alike to get actively engaged in assessment for learning. However, evaluating assessment at Kirkpatrick level II and higher requires long-range studies and a considerable investment of time and money. The literature on postgraduate assessment is awash with level I studies because they are easy, cheap and the results are readily available. Meanwhile, trainees, supervisors and society deserve efficient and effective postgraduate medical training, to reduce costs and to ensure high quality patient care.

**Postgraduate medical education research consortium**

For all reasons mentioned above it is imperative that efforts and budgets are joined in a postgraduate medical education research consortium. This will make it possible to initiate multicentre trials that compare educational interventions to traditional training in several large cohorts.
We conclude this thesis with the following concrete suggestions for research topics that are worthwhile and realistic to explore:

**Formative knowledge assessment**

*Explore ways to introduce a gold standard*

- Relate formative knowledge assessment to summative knowledge assessment.
- Move from a norm-referenced formative standard (standard is set relative to the scores of peers) towards a criterion-referenced summative standard (a pre-set pass/fail standard is set before the test is taken).

*Explore ways to introduce a feedback cycle*

- Give individual feedback regarding the incorrect answers, and repeat the test several months later with a criterion-referenced standard.

**Formative workplace-based assessment**

*Is it possible to improve the educational impact of formative workplace-based assessment by enabling long-term trainee-supervisor relationships?*

- Compare training in clinical rotations with partial end goals, which are supervised by a single supervisor or a small group of subspecialists, to longer training periods with comprehensive end goals that are supervised by all faculty. The educational impact can be assessed, for example, by comparing performance on a knowledge or OSCE-station exam, multisource feedback, learning curves or time-to-competency interval.

*Which instrument is best positioned to sustainably improve communication skills?*

- Compare mini-CEX as used in daily practice to formative assessment of communicative skills involving standardised patients and behavioural scientists as assessors. Differences can be assessed by multisource feedback or patient questionnaires.

**Assessment of procedural skills**

*What is the most efficient way to teach procedural skills?*

- Compare the time-to-competency interval for training that is based on regular formative assessment to training as usual.
- Compare the time-to-competency interval for training that is based on regularly structured formative assessment to loosely structured formative assessment (feedback snapshots).
- Compare the time-to-competency interval for training a procedure with a single supervisor to being trained in a procedure by alternating supervisors.

*What gold standard should be used to assess competence in procedural skills?*

- Compare expert judgements to objective, structured checklists.
References


7. Dijksterhuis MGK, Schuwirth LWT, Braat DDM, Scheele F. An exploratory study into the impact and acceptability of formatively used progress testing in postgraduate obstetrics and gynaecology. Perspectives on Medical Education 2013;2(3):126-141.


