Doctoral theses at NTNU, 2024:88

Susanne Skjervold Smeby Martinsen

Quality learning in undergraduate medical education

Improving teaching and assessment practices

Doctoral thesis

NTNU
Norwegian University of Science and Technology
Thesis for the Degree of
Philosophiae Doctor
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

Susanne Skjervold Smeby Martinsen

Quality learning in undergraduate medical education

Improving teaching and assessment practices

Thesis for the Degree of Philosophiae Doctor

Trondheim, March 2024

Norwegian University of Science and Technology
Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

NTNU
Norwegian University of
Science and Technology

NTNU
Norwegian University of Science and Technology

Thesis for the Degree of Philosophiae Doctor

Faculty of Medicine and Health Sciences
Department of Clinical and Molecular Medicine

ISBN 978-82-326-7770-2 (printed ver.)
ISBN 978-82-326-7769-6 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

Doctoral theses at NTNU, 2024:88

Printed by NTNU Grafisk senter

Undervisning- og vurderingspraksis for å fremme læring i legers utdanning

Det siste århundret er det gjort store fremskritt i forståelsen av hvordan mennesker lærer. Likevel er mye av vår praksis i medisinutdanningen langt fra integrert med denne forskningen. Klasseromsundervisning støtter seg fortsatt tungt på foredragsbaserte metoder, og klinisk undervisning mangler tilstrekkelig vurdering og tilbakemelding på praktiske ferdigheter. I tillegg reflekterer ikke eksamensinnholdet alltid de reelle utfordringene som nye leger vil møte i praksis. Dette svekker troverdigheten til vurderingene og deres mulighet til å fremme læring. Målsettingen med denne forskningen var å forbedre klasseromsundervisning, klinisk undervisning og vurderingspraksis på medisinstudiet ved Norges teknisk-naturvitenskapelige universitet (NTNU) i tråd med konstruktivistisk læringsteori.

I den første studien ble en modifisert form for teambasert læring (TBL) introdusert, kalt express TBL, som viste å engasjere studentene og øke deres tilfredshet og opplevelse av læring sammenlignet med tradisjonelle forelesninger. I en overkrysningsstudie med 105 tredjeårsstudenter viste metoden imidlertid ingen forbedring i prestasjon på eksamen sammenlignet med tradisjonelle forelesninger.

I den andre studien ble formative mini-Clinical Evaluation Exercise (mini-CEX)-vurderinger implementert under sykehuspraksis for femteårsstudenter. Studentene var positive til denne tilnærmingen og de fikk i større grad tilbakemelding på ferdigheter innen opptak av sykehistorie. I en randomisert kontrollert studie som inkluderte 38 femteårsstudenter fant vi ingen endringer i prestasjon på avsluttende skriftlig og muntlig-praktisk prøve.

I den tredje studien ble en ny metode for kvalitetssikring av eksamensoppgaver introdusert, ved bruk av fagfellevurdering av klinikere. Etter å ha gjennomgått 1353 oppgaver, anbefalte fagfellene at 20% av oppgavene burde endres eller fjernes fra eksamen. Det ble gjennomført endringer i 40% av de oppgavene som ikke ble godkjent, som tilsvarer nesten 10% av det totale antallet oppgaver som ble vurdert.

Målet med prosjektene var å forbedre undervisnings- og vurderingspraksis på medisinstudiet, med fokus på gjennomførbarhet. Selv om det viste seg vanskelig å påvise konkrete forbedringer i studentenes prestasjoner, har prosjektene reist viktige spørsmål som bør utforskes videre.

Navn kandidat:	Susanne Skjervold Smeby Martinsen
Institutt:	Institutt for klinisk og molekylær medisin (IKOM)
Veileder(e):	Tobias S. Slørdahl (hovedveileder), Børge Lillebo (biveileder), Erik Magnus Berntsen (biveileder), Vidar Gynnild (biveileder)
Finansieringskilde:	Fakultet for medisin og helsevitenskap, Norges teknisk-vitenskapelige universitet (NTNU), og NTNU Teaching Excellence.

Ovennevnte avhandling er funnet verdig til å forsvares offentlig
for graden PhD i Medisin og helsevitenskap.
Disputas finner sted ved NTNU torsdag 21. mars 2024, kl. 12.15.

Table of contents

Acknowledgements

The research in this thesis has been carried out at the Faculty of Medicine and Health Sciences at the Norwegian University of Science and Technology (NTNU), to which I am grateful for providing the funding for this PhD. The work has also been supported by grants from NTNU Teaching Excellence.

I started my PhD in August of 2016, just months within graduating from medical school at the same university. I spent three years as a PhD student, and continued the work alongside starting my career as a doctor, currently specialising in paediatrics. Now it is finally time to express my deep gratitude to those who have made completing this PhD possible.

First and foremost, I would like to thank my main supervisor Tobias for his unwavering support, patience, enthusiasm and optimism on everything from recruiting participants to publishing articles. Your (enviable) efficiency in reading manuscripts, providing feedback and replying to e-mails remains a complete mystery to me. I would also like to thank my co-supervisor Børge, who has been involved in the planning of all my research projects, and has impressed me with his knowledge and overview of the medical education literature. To my co-supervisors Erik and Vidar, I am grateful for your support and constructive feedback throughout. This thesis would not have been the same without our discussions.

I would also like to extend my sincerest appreciation to the students, doctors, patients and standardised patients who participated in this PhD, and shared their knowledge and experiences with me. A special thank you to the students at the Center of Assessment in Medical Education (CAME), for pulling off an impressive ‘examination’ to end my final project!

Finally, my deepest gratitude goes to my family and friends for their support and encouragement, not only during the past seven years. To my

parents for taking me abroad to experience a range of different educational settings, for encouraging my academic curiosity and for always believing in me. To my brother Karl Magnus for the shared memories and support along the way. But most importantly, to my husband Tobias and our daughter Mathilde, for your love and patience (and all those dinners). Without you, it would not have been possible.

Susanne Skjervold Smeby Martinsen,
Oslo, August 2023

List of papers

I. Smeby, S. S., Lillebo, B., Slørdahl, T. S., & Berntsen, E. M. (2020). Express Team-Based Learning (eTBL): **A Time-Efficient TBL Approach in Neuroradiology.** *Academic radiology*, *27*(2), 284-290.

II. [1]Martinsen, S. S. S., Espeland, T., Berg, E., Samstad, E., Lillebo, B., & Slørdahl, T. S. (2021). **Examining the educational impact of the mini-CEX: a randomised controlled study**. *BMC medical education*, *21*(1), 228.

[1]Please note that my surname has changed after my marriage.

III. Smeby, S. S., Lillebo, B., Gynnild, V., Samstad, E., Standal, R., Knobel, H., Vik, A., & Slørdahl, T. S. (2019). **Improving assessment quality in professional higher education: Could external peer review of items be the answer?**. *Cogent Medicine*, *6*(1), 1659746.

Abbreviations

CanMEDS	Canadian Medical Education Directives for Specialists
CEX	Clinical evaluation exercise
CRT	Cluster randomised trials
CT	Computed tomography
DOPS	Direct observation of procedural skills
IF-AT	Immediate feedback assessment technique
iRAT	Individual readiness assurance test
eTBL	Express-Team-based learning
IWF	Item-writing flaws
MCQ	Multiple choice question
MEQ	Modified essay questions
mini-CEX	Mini-Clinical Evaluation Exercise
MRI	Magnetic resonance imaging
NBME	National board of medical examiners
NSD	Norwegian Centre for Research Data
NTNU	Norwegian University of Science of Technology
OSCE	Oral structured clinical examination
OSLER	Objective structured long examination record
PBL	Problem-based learning
RAT	Readiness assurance test
RCT	Randomised controlled trial
STC	Systematic text condensation
tAPP	Team application
TBL	Team-based learning
tRAT	Team readiness assurance test
WBA	Workplace-based assessment
ZPD	Zone of proximal development

Summary in English

Although there have been significant advancements in our understanding of human learning over the last century, many educational practices within the medical field have not integrated educational research findings. Conventional classroom teaching still heavily depends on lecture-based techniques, while clinical teaching frequently lacks sufficient opportunities for assessment and feedback on clinical skills. Moreover, a considerable portion of assessment content falls short in accurately reflecting important and authentic clinical challenges that newly graduated doctors are likely to encounter in their daily practice. This compromises the validity of the assessments and its ability to fully capitalise on its potential impact on student learning.

The aim of this thesis was to develop and implement changes to classroom teaching, clinical teaching and assessment practices in the medical curriculum at the Norwegian University of Science and Technology (NTNU). The research was underpinned by constructivist learning theories and focused on evaluating effects on educational impact and assessment quality.

Paper I examines the implementation of a modified, time-efficient Team-based learning (TBL) approach called express TBL. It fostered high levels of engagement and students reported significantly higher satisfaction and learning when compared to traditional lectures. However, in a cross-over design with 105 third-year students, it showed no improvement in performance on a summative examination compared to traditional lectures.

Paper II examines the implementation of formative mini-Clinical Evaluation Exercise (mini-CEX) assessments during a clinical placement for 38 fifth-year students. Students were generally positive to their use, but in a randomised controlled trial (RCT) design, showed no overall improvement on direct observation, feedback or performance compared with ad-hoc

feedback. However, feedback on history taking skills were reported significantly more common in the intervention (mini-CEX) group.

Paper III examines a novel quality assurance procedure for in-house examinations: external double-blinded review of assessment items by clinicians. In all, 1353 items were reviewed by 170 external reviewers, of which 20% were either rejected or judged as needing revision. Content relevance, content accuracy and technical flaws emerged as the main reasons for not approving items. Following review and feedback, changes were made to 40% of disapproved items, which constitutes almost 10% of the total number of items that were reviewed. This study showed that external peer review is cost-effective and feasible in an in-house setting with fewer resources available, and can identify items that have the potential to significantly reduce the validity and educational impact of examinations.

Grounded in learning theory, these projects were designed with the aim of enhancing the quality of learning within the medical curriculum. Throughout this thesis, a central focus has been placed on feasibility to ensure that the proposed changes are viable within a realistic educational setting. Although the search for evidence demonstrating educational impact in terms of improved student outcomes presented challenges, it sparked numerous inquiries for future exploration.

Summary in Norwegian

Det er gjort betydelige fremskritt det siste århundret i vår forståelse av hvordan mennesker lærer. Likevel er mye av vår utdanningspraksis innen medisin langt fra integrert med utdanningsforskningen. Klasseromsundervisning støtter seg fortsatt tungt på foredragsbaserte metoder, mens klinisk undervisning ofte mangler tilstrekkelige muligheter for vurdering og tilbakemelding på kliniske ferdigheter. Videre faller en betydelig del av eksamensinnholdet kort i å gjenspeile viktige og autentiske kliniske utfordringer som nyutdannede leger vil møte i deres daglige praksis. Dette går på bekostning av validiteten av vurderingene, og deres evne til å utnytte potensialet for å fremme læring.

Målsettingen med denne avhandlingen var å utvikle og implementere endringer i klasseromsundervisning, klinisk undervisning og vurderingspraksis i medisinstudiet ved Norges teknisk-naturvitenskapelige universitet (NTNU). Forskningen støttet seg på konstruktivistisk læringsteori, med fokus på å evaluere effekten på læring og vurderingskvalitet.

I den første artikkelen presenteres en modifisert Team-basert lærings (TBL)-metode som kan gjennomføres på kortere tid (express TBL). Studien viste at eTBL aktiviserte studentene i høy grad, og studentene rapporterte betydelig høyere tilfredshet og læring sammenlignet med tradisjonelle forelesninger. I en overkrysningsstudie med 105 tredjeårsstudenter viste det imidlertid ingen forbedring i prestasjon på eksamen sammenlignet med tradisjonelle forelesninger.

I den andre artikkelen presenteres implementering av formative mini-Clinical Evaluation Exercise (mini-CEX)-vurderinger under sykehuspraksis for 38 femteårsstudenter. Studentene var generelt positive til mini-CEX, men i en randomisert kontrollert studie fant vi ingen endringer i direkte

observasjon, tilbakemelding eller prestasjon på avsluttende skriftlig og muntlig-praktisk prøve, sammenlignet med ad-hoc tilbakemeldinger. Imidlertid ble tilbakemeldinger på ferdigheter i anamneseopptak rapportert signifikant mer hyppig i intervensjonsgruppen (mini-CEX).

I den tredje artikkelen presenteres en ny metode for kvalitetssikring av eksamen gjennom fagfellevurdering av eksamensoppgaver av klinikere. I alt ble 1353 oppgaver fagfellevurdert av 170 klinikere, hvorav fagfellene ba om at 20% burde endres eller fjernes fra eksamen. Relevans, mangler eller feil i faglig innhold, og tekniske oppgavefeil var hovedårsakene til at fagfellene ikke godkjente oppgavene. Etter fagfellevurderingen ble det gjort endringer i 40% av underkjente oppgaver, som utgjør nesten 10% av det totale antallet oppgaver som ble vurdert. Studien viser at fagfellevurdering av klinikere er kostnadseffektivt og gjennomførbart, og kan identifisere oppgaver som kan betydelig redusere validiteten og læringseffekten av eksamen.

Målet med prosjektene var å forbedre kvaliteten på undervisning og vurderingspraksis i medisinstudiet. Gjennom hele prosessen har gjennomførbarhet vært et sentralt fokus for å sikre at de foreslåtte endringene er realistiske og praktisk gjennomførbare. Selv om det viste seg vanskelig å påvise konkrete endringer i studentenes prestasjoner, har prosjektene likevel inspirert til mange spørsmål som bør utforskes videre fremover.

1. Introduction

This thesis begins with an exploration of quality in medical education, drawing upon a historical review to elucidate the challenges encountered in today's undergraduate medical education. Delving into classroom teaching, clinical teaching, and assessment practices, the subsequent chapters shed light on the challenges faced within each domain. These key topics serve as building blocks for each of the three articles that form the core of this thesis: developing an active learning strategy for classroom teaching, implementing formative assessments in clinical placements, and a novel method for quality assurance of in-house examinations.

1.1 Quality in medical education

1.1.1 Defining and evaluating quality in medical education

Quality in education is a multi-interpretable concept, and different stakeholders have different perceptions about what it is (1). In medical education, students will see teaching which prepares them for their examinations and guides them to become great doctors (2 p. 101). University administrators will also see value for money and external accreditation. Likewise, teachers, employers, medical councils, funders and the general public will have their own perceptions about what quality medical education looks like. Although we all have an intuitive understanding of what quality comprises, it is often hard to articulate (3). It should therefore be of no surprise that the education literature lacks a unified concept and definition of quality (4).

In 2015, Schindler and colleagues identified four broad conceptualisations of quality in higher education through a review of the literature: quality as purposeful, exceptional, transformative, and accountable (4). Interestingly, they note that the conceptualisations are consistent with those described in the early 1990s, suggesting that the meaning of quality in higher education has remained relatively stable (3, 4).

Quality as purposeful is defined as conformance to a stated mission or a set of standards, including those defined by accrediting and regulatory bodies. Quality as exceptional is an achievement of distinction through the fulfilment of high standards (e.g., credibility, prestige, or ranking). Quality as transformative is a positive change in student learning and professional development. Lastly, quality as accountable refers to the accountability to stakeholders for the optimal use of resources and delivery of accurate educational products and services (e.g., student preparedness for employment, sufficiency of facilities, and focus on continuous improvement).

Donabedian's framework for evaluating quality in health services is useful and applicable also in higher and medical education (5, 6). Donabedian proposed that quality can be measured in three dimensions: structure, process and outcome. Structure refers to the conditions that must be in place in order to deliver quality education (i.e., funding and resources, facilities, human resources, design of the educational programme). Process refers to the implementation of the educational programme (i.e., the teaching and learning activities, learning environment, assessment and feedback practices). Finally, outcome refers to the effect of the education (i.e., what competencies students achieve, preparedness for employment).

Just as in the healthcare sector, the focus has shifted from assessing structures and processes, to defining and measuring outcomes, for reasons of cost and accountability (7). Given the growing scarcity of resources and rising costs, along with the need to enhance the quality of medical education and the competence of the doctor emerging from it, efficient resource utilisation and documenting outcomes are key. Quality structures and processes do not mean much in medical education if they do not lead to competent professionals.

This thesis aims to examine multiple facets of undergraduate medical education, including classroom teaching, clinical teaching, and assessment practices. By delving into these areas, I will explore the challenges they present and identify potential strategies for enhancing their quality in terms of improving student learning outcomes.

1.1.2 Medical education in a historical perspective

To present a comprehensive history of undergraduate medical education is beyond the scope of this introduction. Instead, the purpose of this summary is to present key trends in curricular design and assessment practices that have had significant implications for the quality of medical education. It serves as a foundation for understanding the challenges we face today.

In the mid-19th century, medical education in North-America was dominated by the propriety school model (8). These courses were superficial and brief, typically lasting two four-month terms of lectures, had low entrance requirements, teaching was almost exclusively didactic with no patient contact or laboratory experience, and they were run for-profit (8). Clinical training took the form of a one- to three-year apprenticeship with a private doctor, and the quality of learning depended wholly on the mentor's resources and experience (9). Without certification requirements, graduating doctors were marginally competent and the quality of their education generally poor (10 p. 11).

During the late 19th century, a number of university medical courses were making significant changes to their curricula amid the birth of experimental medicine in Europe (8). Increased admission requirements, longer curricula to cover new scientific subjects, the introduction of laboratory and clinical teaching, frequent testing and a focus on medical research were improving the quality of education. A number of schools, most famously John Hopkins, adopted the discipline-based curriculum, modelled after contemporary European medical schools (9). Faculty members were divided into discipline-specific departments (e.g., chemistry, anatomy, cardiology) and conducted both classroom and clinical teaching. The pedagogy of rote memorisation during the apprenticeship-model was now being replaced by new educational principles with the intention of creating independent thinkers and problem-solvers (9).

In 1910, Abraham Flexner published his famous report *Medical Education in the United States and Canada* for the Carnegie Foundation for the Advancement of Teaching, which would go on to influence medical education for over a century (11). Despite ongoing reform efforts, the state of medical education remained highly variable and generally inadequate when he embarked on his visits.

The model that Flexner proposed in order to improve the quality of medical education was based on that of John Hopkins, and consisted of rigorous entrance requirements and a university-based, research-oriented education with the scientific method of thinking as its foundation (10 p. 13). The structure of two years of basic sciences followed by two years of supervised clinical experience is still evident in North America today. Flexner also advised that that students spend most of their time in the laboratory and clinic, instead of in lectures (8). Although the educational ideas that Flexner presented in his report were not conceptually new, it brought the concerns about medical education to the attention of the general public (8). Within a decade, one-third of the schools had closed or merged with other schools, medical education became much more homogenous, and accreditation and licensing procedures were introduced (10 p. 13).

By the mid-20th century, the educational principles of the traditional discipline-based curriculum were being questioned, especially the segregated pre-clinical-clinical curricular structure and the vast amount of disjointed basic science information students were expected to learn (9). The lack of coordination between lecturers, failing to build on students' existing knowledge, and requiring students themselves to integrate basic science with clinical concepts, impeded learning in the discipline-based curriculum (10). Additionally, contrary to Flexner's advice of actively engaging students in learning, the pedagogical strategies employed in the discipline-based curriculum often relied heavily on lectures.

The organ-system-based curriculum, which had its origins at the Western Reserve School of Medicine during the early 1950s, attempted to mediate some of these disadvantages through organising teaching around body systems. The anatomy, physiology, pathophysiology (and in modern iterations, also the clinical signs and symptoms) within a single organ system are taught sequentially. However, when a doctor's primary task is to identify the underlying cause of a particular complaint among various organ systems or disease categories, this framework for conceptualising interconnected knowledge within a single organ system appears to be an artificial construct (9). This lack of authentic clinical context in which to learn was further undermined by an emphasis on formal scientific knowledge over clinical experience, and an assessment system that tended to focus on knowledge acquisition rather than clinical skills (10 p. 79).

During the late 1960s and early 1970s, the concept of problem-based learning (PBL) and the implementation of problem-based curricula gained prominence, with McMasters University and the University of New Mexico as well-known forerunners (10 pp. 80-81). In PBL, a clinical case serves as a stimulus for small-group discussions, aiming to actively involve students in self-directed learning and to share their knowledge with each other (10 pp.

80-81). PBL is based on the notion that learning is best facilitated within the authentic context of a patient or clinical case. This emphasis on contextual and active learning was influenced by contemporary cognitive science-based learning theories. However, like the previously mentioned curricular structures, the PBL model places greater emphasis on formal knowledge and clinical reasoning than on the development of patient care skills (10 p. 81). Furthermore, the assumption that PBL sessions would develop students' hypothetico-deductive reasoning skills (and consequently, their clinical problem-solving skills) irrespective of the clinical scenario, was soon disproven (9).

At the time, competence was seen as the sum of different components, for example 'knowledge', 'skills' and 'problem-solving' (12). These components were assumed to be independent of each other and relatively stable across situations, and so it was assumed that a trait such as 'problem-solving' could be learned and assessed independently of 'knowledge'. However, a number of research findings challenged this trait-model of competence. Most importantly, the consistent finding that the correlation of performance across different problems is low, more commonly known as 'content specificity' (13, 14). This implies that performance relies heavily on knowledge that is relevant to a specific problem, and that demonstrating knowledge for one problem does not automatically infer knowledge relevant for another (14).

An important shift occurred during the 1990s, when the trait-model of competence was replaced by the idea of competencies (15, 16). This concept of competence extended beyond the traditional focus on diagnostic problem-solving, and encompassed broader domains such as effective communication, professionalism, and system-based practice (17, 18). Medical education started to prioritise the development and assessment of these essential skills and attributes, alongside clinical knowledge. Today, competency frameworks such as the Canadian Medical Education Directives for Specialists (CanMEDS) and the UK's General Medical Council (Tomorrow's Doctors) form the basis of many medical education curricula worldwide (19). These frameworks offer a more comprehensive and multidimensional view of what it means to be a competent doctor, ensuring that they possess the diverse range of competencies required to meet the demands of modern healthcare effectively.

In 2010, marking a century since the Flexner report, the Carnegie Foundation released another report assessing the state of medical education in the United States (10). The challenges identified in this report remain relevant not only today, but also globally. The authors of the report observed that medical education tends to be rigid, excessively long, and lacking a learner-centred approach. Additionally, there is a disproportionate emphasis on memorisation of facts and disconnection between formal knowledge and practical learning experiences, limited teaching opportunities for clinical staff due to time constraints and inadequate support from hospitals for the educational mission.

This historical perspective shows us that the curricular models, pedagogical strategies and assessment practices in undergraduate medical education should be perceived as continuously evolving processes (10 p. 110). Despite significant advancement since the 19th century, such as standardisation of education, entry and certification requirements, as well as development of learning strategies and assessment practices, substantial challenges persist in undergraduate medical education. One important challenge is the implementation of evidence-based practices that take into account developments in the learning sciences, particularly in the face of rapid changes in medical practice and complexity of modern healthcare.

1.1.3 Learning theories: Constructivism, medical expertise and constructive alignment

Constructivism has given rise to theories of learning that have significant implications for medical education (20 p. 53). According to the constructivist perspective, learning takes place when learners actively construct the meaning of new knowledge based on their existing knowledge and past experiences (20 p. 53).

Constructivist theories of learning have their roots in the early 20th century with Piaget's theory of cognitive development (21). According to this theory, learning is a constructive process in which the learner builds cognitive schemata as a personal interpretation of his or her experience (22 p. 58). As individuals learn, these schemata become increasingly complex, modifying and expanding as new knowledge is integrated into the existing structures (20 p. 53).

Another influential figure in constructivist learning theory is Vygotsky, who introduced the concept of social constructivism, which emphasises the role of social interactions and cultural context in learning (22 p. 58, 23). Through these interactions, shared experiences and discussions, learners actively construct an understanding together that would not come about alone (24). He also introduced the notion of the zone of proximal development (ZPD), which refers to the gap between a learner's current level of ability and their potential level of ability with the support of more knowledgeable individuals (such as teachers or peers) (22 p. 58). Vygotsky believed that learning is maximised when learners are provided with appropriate scaffolding and guidance within their ZPD.

The development of medical expertise

Much of our understanding of learning in medicine is grounded in the principles of the constructivist perspective. Research into the development of medical expertise has shown that expertise develops with the ability to organise information in information-rich units, or schemas (25). Schemas can be constructed by combining a number of simpler elements into one, or by adding new elements into already constructed schemas. This implied that the existence of prior information makes it easier to store new information. This effective organisation of knowledge into schemas has the advantage of allowing the person to swiftly store new information, and retrieve relevant information when needed (26). Additionally, it greatly increases the capacity of the working memory, as even a highly complex schema can be treated as one unit (25).

The dominant theory of how expertise develops in medicine is that of illness script formation (27). The theory outlines several stages of development, and deviates from the early idea that expertise develops merely as a result of knowledge expansion (28). In the first stage, students form *causal networks* that link signs and symptoms of patients to the underlying pathophysiological concepts that they have learned (27). This process is time-consuming as they have not yet learned to recognise patterns of symptoms like experts do. In the second stage, the detailed *causal networks* become condensed into a smaller number of simplified, higher-level diagnostic concepts through a process termed *knowledge encapsulation.* For example, a patient with a high fever, tachycardia, hypotension and confusion may simply be recognised as having sepsis, without having to refer back to the complex pathophysiology. With increasing experience, the encapsulated knowledge is reorganised into structures known as *illness scripts*, which are governed more by *enabling conditions*, and less by the underlying physiology. Enabling conditions are features of a patient, like age, sex, ethnicity, comorbidities and risk factors, that are used to make a certain diagnosis more or less likely. Experienced doctors identify one or a few scripts in the process of solving a new problem, and match the information in the script to that of the patient.

Van der Vleuten summarises this view of expertise development as a "*transition from a conceptually rich and rational knowledge base (acquired from educational experience) to a non-analytical ability to recognize and handle situations efficiently and effectively (acquired from clinical experience)* (12). The aim of the medical education should be to support learners on this transition.

Implications for educational practice in medicine

Constructivist theories of learning have important implications for quality teaching and learning in medicine.

The central idea of constructivism is that knowledge is actively constructed by the learner. Thus, teaching cannot be viewed as the passive transmission of knowledge from enlightened to unenlightened (29). Instead of assuming a traditional role as instructor, teachers should take on the role of guides or facilitators, creating learning opportunities and offering scaffolding to support learners. Learning experiences should be authentic with regards to their professional development and include problems that are relevant and important to learners.

Teachers must consider learners' prior knowledge, as this is the basis upon which new knowledge is tested and built (29). To support students in the encapsulation process, biomedical and clinical sciences should be integrated and teaching contextualised (26, 27). They should be allowed to see many and varied patients, and to discuss and reflect upon them, to encourage illness script formation (27).

Learning is promoted through collaboration among students, and between students and teachers, and there is need for sufficient time for learners to actively build their knowledge, reflecting on the relationship between their experiences and previous ones (22 p. 59, 29).

Constructive alignment

The constructivist theories of learning have also had implications for design of curricula. Constructive alignment is a principle devised by Biggs during the 1990s to enhance the quality of learning, and represents a union between the constructivist understanding of learning and outcomes-based education (30). Constructive alignment describes how assessment tasks, as well as teaching and learning activities, align with the intended outcomes of the educational programme, in order that students achieve those outcomes more effectively (Figure 1) (31 p. 14).

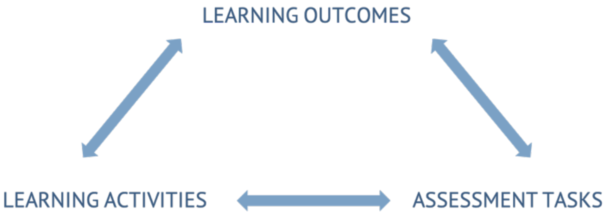


Figure 1 Biggs' constructive alignment (31)

The intended learning outcomes specify what, how and to what standard something should be learned as a result of engaging in the teaching and learning activities (31 p. 98). They are written from the students' perspective, indicating that the student is the focus, not the teacher. The verb (e.g., identify, explain, create) specifies the level of understanding required. Teaching and learning activities are designed to encourage students to achieve the intended learning outcome. In line with constructivist learning theories, activities need to engage students in activating the learning outcome verb. Finally, assessment tasks are constructed to assess whether students have achieved the intended learning outcomes. Following from this, assessment should be criterion-referenced (i.e., assessing whether a student's learning meets the intended outcomes) instead of norm-referenced (i.e., assessing how students compare with each other) (31 p. 106).

Constructive alignment aims to bridge the gap between a static body of declarative knowledge ('university' knowledge) and personal action ('professional' knowledge, functioning knowledge) that is so important in medical education and other professional education courses (31 p. 97). Biggs argues that this gap has traditionally been left to the students to do 'out there', but which is a job that should be done before graduation.

The 'hidden curriculum' is an often used metaphor for the discrepancy between formal statements of requirements, and what teachers actually expect, in terms of how they teach and what is rewarded through assessment. Snyder, who popularised the term in 1971, reported on the difference between the formal curriculum emphasising goals such as independent thinking and problem-solving, and students' experience that teaching and assessment in fact rewarded memorising facts and theories (32). Relating back to Biggs' constructive alignment: When learning outcomes and assessment tasks are not aligned, it is usually the assessment that prevails, emphasising its importance (33). As Boud simply stated: "*Every act of assessment gives a message to students about what they should be learning and how they should go about it"* (34). Constructive alignment capitalises on the effect of assessment on students' learning.

1.1.4 Evidence-based practice in medical education

The central argument of this thesis emphasises the importance of aligning teaching strategies and assessment practices in medical education with theories of learning and evidence-based educational practices. The aim is to ensure the delivery of high-quality undergraduate education that effectively prepares students for their future careers as doctors.

The idea of applying learning theories to medical education has been a topic of discussion for a long time. Even before the emergence of constructivist theories of learning, Flexner wrote: '*On the pedagogic side, modern medicine, like all scientific teaching, is characterized by activity. The student no longer merely watches, listens, memorizes; he* does.' (11 p. 53). Throughout the history of curricular innovations in medical education, the concept of providing a more active and authentic learning experience for students has been consistently present. These innovations aimed to move away from the passive transmission of knowledge and rote memorisation, and instead focus on engaging students in meaningful learning contexts. Despite this recognition, it is disheartening to find that many educational practices in the field of medicine are still not firmly rooted in educational research (35).

This is as relevant for medical education in our local context of Norway, as anywhere else. Classroom teaching is still heavily lecture-based, there are few opportunities for evaluation and feedback on clinical skills, and assessment practices do not capitalise on their potential effects on student learning. These concerns are not limited to the observations of teachers and educational researchers alone. Medical students themselves, as evidenced by their responses in the Student Survey ('Studentbarometeret'), consistently express dissatisfaction with the quality of feedback and level of active learning experiences in their education (36).

This highlights the need for a more robust integration of educational research and practices in medical education, and one that takes feasibility of implementation into account. If the realities and constraints of the medical education setting are considered, we can begin to promote meaningful and sustainable improvements in the quality of education.

The following sections will highlight the specific challenges that persist in undergraduate medical education, pertaining to classroom teaching, clinical teaching, and assessment strategies. Each section will subsequently present a review of the literature, identifying research gaps which the three papers will attempt to address.

1.2 Classroom teaching

1.2.1 Challenges in classroom teaching

One of the important challenges facing medical education is curriculum overload. This is not new: concerns about overwhelming amounts of content and crowded curricula were raised over a century ago by Flexner (11). The rapid expansion of scientific and medical knowledge poses significant challenges not to what we can include, but what we can safely exclude from medical curricula. Additionally, the new competency frameworks highlight new expectations of doctors beyond scientific knowledge and problem-solving skills: communication skills, teamwork, ethics, patient safety and quality improvement, innovation, and cultural competence are just a few of the 'new' subjects that are being added.

Content overload is further exacerbated by the lack of integration of basic and clinical science, which is one of the main findings in the Carnegie report of 2010 (10). Early-stage medical students fail to see the relevance and clinical context of basic science when it is not linked to what they experience in clinical settings. Custers and colleagues found that when basic science is taught disassociated from patients, 30-50 percent is forgotten after two years (37). When students reach the clinical setting, they are required to reorganise this knowledge to a patient-centred clinical perspective (10 p. 28-29). Struggling with factual overload and the lack of integration, students are forced to memorise the content instead of appreciating the relationships between subjects and concepts, and how to apply them to clinical contexts.

Traditional, didactic lectures continue to be a key component of most medical curricula, despite research repeatedly showing that only a small percentage of the information delivered by lectures is retained (10 p. 92, 38, 39). This form of instruction assumes that the delivery of information, especially from an expert, leads to quality learning (40). However, as previously argued, it is the active processing of such information by the learner that leads to quality learning.

Kvernenes and Schei suggest three reasons for the continued use of lectures in medical education (6). First, challenged by the sheer amount of content to be covered, teachers are increasingly using lectures to *show* students what they need to learn instead of actually *teaching* the concepts. Second, students *like* lectures. It requires little of the student, and listening to a good and inspiring lecture can give the feeling of having learned. However, the lecturer's fluency can easily be mistaken for actual learning (41). Third, lectures give the teachers a certain amount of *control* over the learning activity, whereas collaboration, problem-solving and discussions (that would indeed support the learning process) means letting go of some of this control.

1.2.2 Active learning strategies: Team-based learning (TBL)

Quality education entails implementing teaching strategies that align with constructivist theories of learning, facilitating students' active processing of information. Active learning is an umbrella term that comprises of a variety of teaching and learning techniques that seek to shift the focus from the teacher to the learner, and promote learning through active engagement with the content (42). Active learning strategies in medical education usually require students to apply their knowledge to clinically relevant problems and encourages students to transfer knowledge to new situations (42). Group work also offer opportunities to learn skills such as communication and team skills (42). A number of different active learning strategies have been implemented in medical education, ranging from more elaborate forms such as PBL and team-based learning (TBL), to the use of active techniques in a more traditional lecture setting, such as brainstorming or use of student response systems.

TBL is an active learning strategy which is used extensively in health professions education. It was developed by professor Larry Michaelsen in the field of business during the early 1990s, in response to growing class sizes and the need for his students to face real-life problems of the business world (43, 44). One of the main advantages of TBL is that it is designed for large-group teaching, in contrast to other forms of active teaching strategies that require higher student to staff ratios.

The original application consists of three phases (Figure 2): (i) preparatory reading or other advance assignments; (ii) readiness assurance tests; and (iii) team application.

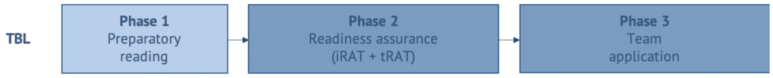


Figure 2 The phases of team-based learning (TBL). The light fields represent out-of-class preparations, and the darker fields represent in-class time. iRAT: individual readiness assurance test; tRAT: team readiness assurance test.

In Phase 1, students are set a list of reading material or other learning activities, which they complete in preparation for in-class work. Phase 2 starts with an individual readiness assurance test (iRAT), usually in the form of 10-20 MCQs which tests basic facts and concepts of the advance assignment.

Before receiving feedback on their performance, students retake the same test in teams of 5-7 students in order to reach a consensus on the answers (team readiness assurance test, tRAT). The tRAT is answered using an immediate feedback assessment technique (IF-AT), which can be in the form of a scratch card, so that students receive immediate feedback on whether an answer is correct or wrong, motivating students to collaborate until all answers are correct (45). In Phase 3, the teams apply their knowledge to solve problems they are likely to meet in the professional careers (team application, tAPP).

The tAPP follows four principles for effective problem design, known as the 4 S's principles (44, 45): problems should be significant for the students, problems should be the same for all teams, teams must make a specific choice and simultaneously report their answers. This ensures that students get immediate feedback on their answers, and that they are accountable to explain and defend their answers (44).

1.2.3 Theoretical grounding for team-based learning (TBL)

TBL adheres to the principles of constructivist learning theories and supports the constructive alignment of learning objectives, learning activities and assessment practices (45).

The preparatory learning activities give students opportunities to learn basic concepts and repeat previously covered material, supporting learning through the integration of new information into existing mental schemes (45). They also hold students accountable to come prepared and take responsibility for their own learning, which reduces the need to cover extensive amounts of content during class sessions, as students are already familiar with the material.

The in-class exercises ensure that students must engage with the material and process the information, instead of being passive receptors of information. Students compare their understanding to that of the group, and new connections are made by exposing these inconsistencies or misconceptions (45). The team application exercises are authentic and relevant to clinical practice which enables proper alignment with learning objectives and assessment practices that should support students' motivation to learn. Additionally, they give students a chance to integrate their formal knowledge with clinical experience, capitalising on the opportunities to reinforce connections between theory and practice.

Teamwork gives students a chance to reflect on the contributions by other team members, and receive feedback on their own in-puts (45). In line with social constructivist theory, learning takes place in the interaction with others as they develop a common understanding.

The teacher takes the role of facilitator, and uses his or her expertise in defining learning objectives, deciding preparatory learning material and developing appropriate problems for the teams to solve. This also means taking an active role in deciding what content is important and what can be safely left out, with reference to discussions on curriculum overload and constructive alignment. In class, the teacher's guidance (as well as from other team members) provide scaffolding for learners in line with constructivist learning theories.

1.2.4 Research on team-based learning (TBL)

TBL has been widely adopted by health professions education across the globe, which is also reflected in a tripling of research outputs on TBL in this field from 2001-2005 to 2011-2016 (46). The majority of published literature concerns undergraduate medical education (46).

A systematic review of the literature on TBL published in 2017, found that most articles related to learner reactions (46). On the whole, learners prefer TBL compared with more traditional forms of teaching (lectures being the most common comparison). They often highlight the active learning style, teamwork and the opportunity to discuss and apply their knowledge to relevant problems. However, a small number of studies showed that learners did not have an overall positive experience with TBL (47-49).

Learners also reported higher levels of engagement in TBL compared with traditional forms of teaching (46). Two studies undertook direct observations of learner engagement and found that most interactions were learner-to-learner (instead of learner-to-facilitator or learner-to-self), which reflects the amount of time spent in team discussions (50, 51).

There is a growing body of evidence that suggests TBL improves student outcomes in health professions education. A systematic review of the effectiveness of TBL on learning outcomes published in 2013 found that of the 14 studies included, seven showed significant increases in knowledge scores for the TBL group, four reported no difference and three showed improvement but did not comment on statistical significance (52). A number of studies find that TBL is more effective for the academically weaker students (46, 53-57).

Although this research supports the use of TBL, there are significant challenges to its implementation (58). The systematic review in 2017 previously mentioned, found that the highest number of TBL sessions reported was six, which could indicate challenges to a more widespread and sustained implementation (46). Teacher training, time spent preparing TBL material (including preparatory material, iRAT/tRAT, tAPP, as well as explanation for answers) and classroom management skills have all been identified as challenges to its implementation (58). After piloting TBL in our own medical programme in 2013, evaluations showed that TBL was perceived as time-consuming (one session lasting three 45-minute blocks) (59).

The structure of TBL lends itself to modifications based on the needs of learners and teachers, and many courses have implemented hybrid versions of TBL (46, 52, 60). Although most hybrid versions retain the readiness assurance phase, it is worth noting that this segment consumes a considerable amount of in-class time while primarily involving students in lower cognitive processes, focusing on basic facts and concepts (52). There are only a few studies on the effects of this phase. Carbrey and colleagues found that learners preferred completing the iRAT at home over the traditional in-class iRAT and tRAT, and found equivalent performance on a physiology test across the two methods (61). Similarly, Gopalan and colleagues found that although the iRAT helped increase subsequent tRAT scores, and decreased time of completion of the tRAT, examination scores were equivalent to the group who did not complete the iRAT.

Contrary to common intuition, a number of studies suggest that higher order learning is not enhanced by first building a foundation of factual knowledge (62, 63). In a study involving middle school and college students, it was found that performance on a delayed higher-order test saw improvement when participants were exposed to quizzes comprising higher-order material or a combination of higher-order and factual material (62). However, quizzes solely focused on factual material alone did not yield the same improvement. Similarly, in a study by McDaniel and colleagues, it was observed that quizzes that required the application of science principles resulted in improved examination performance for both definitional-type questions and application questions. However, quizzes centred solely around definitional questions did not exhibit the same benefits for application-based questions (63). Therefore, when aiming to foster complex learning, such as application of knowledge, engaging students in complex tasks may be more beneficial that beginning with basic facts and definitions.

The study in Paper I presents a modified and time-efficient TBL approach we have called express TBL (eTBL), which aims to facilitate the implementation of TBL in medical curricula. The readiness assurance phase is reduced to a short warm-up exercise, leaving the majority of in-class time to be spent on real-life complex problems (tAPP) and reducing in-class time to 45 minutes.

1.3 Clinical teaching

1.3.1 Challenges in clinical teaching

Clinical teaching has always been an important part of doctors' education. It gives students a chance to socialise into their professional roles, develop clinical reasoning and diagnostic skills, learn about management of patient care, and undergo attitudinal changes informed by their experiences with patients and their families, fellow students and future colleagues (64). Studies indicate that new doctors feel inadequately prepared for practice, but also suggest that early exposure to patients and quality clinical teaching make this transition easier (65).

Clinical teaching in undergraduate medical education, defined as a learning situation in which a patient is present or which takes place in a clinical setting, generally progresses along a trajectory from a more theoretical approach to authentic, workplace-based learning (6 pp. 168-169). Early clinical teaching often takes the form of patients in lectures, progressing through to bedside teaching and finally longer clinical placements. In parallel, the student's role transitions from observer to active participator.

As students' previous educational experiences have largely been centred in the classroom, learning in a clinical setting demands new ways of preparing for, engaging with and reflecting on activities that are both educational and practical (10 p. 42). It requires students to reorganise their knowledge base and convert skills for classroom learning to a clinical setting. Research on the transition from the non-clinical to clinical stage of undergraduate medical education show that students feel anxious and underprepared, but also motivated to learn from real patients and value using more active learning strategies (66-69).

Regrettably, the changing nature of the health services is putting an enormous strain on clinical teaching, in what is now referred to as the 'decline in clinical teaching' (70, 71). Students are spending less time with patients due to increased patient turnover with shorter stays, more efficient and specialised health services, increased workload for clinicians, increasing administrative demands and less protected teaching time. This challenges the ability of teachers to guide students' progression through the clinical years and give them feedback on their performance.

A number of studies document that students are seldom observed, and to an even lesser extent given feedback on, performing skills such as history taking or clinical examination (10 p. 107, 72-75). We have to a large degree relied on the notion that if students complete a specified number of procedures or weeks in a clinical placement, they will emerge competent without observation, feedback and guidance. In fact, participation has been found to be the second most common assessment method, accounting for 20% of all assessment in medical schools in the United States (71).

1.3.2 Feedback on performance

Formative feedback is defined as information shared with the learner intended to modify his or her thinking or behaviour in order to improve learning (76). The feedback should serve to close the gap between where students currently are, and where they aim to be, in terms of knowledge or skills (77). In order to do so, feedback should answer the three questions: Where am I going; how am I going; and where to next?

'Where am I going' relates to information about the learning goals and success criteria (77). 'How am I going' provides information on the progress made towards those goals, and ‘where to next’, about the subsequent steps that need to be taken to make progress or enhance learning. The categories are often referred to as ‘feed-up’, ‘feed-back’ and ‘feed-forward’. Effective feedback consists of two components: verification and elaboration (76, 78). Verification is the simple judgment of whether an answer is correct or incorrect, and elaboration guides the learner towards the correct answer. Elaborated feedback can range from simply reteaching the material, to error analyses focusing on specific errors or misconceptions (76). In its most complex form, the feedback provides verification, error flagging and strategic hints on how to proceed, known as informative tutoring.

Research into the efficacy of different types of feedback and the mechanisms that relate feedback to learning, is inconsistent and complex (76, 79, 80). Many studies show no or negative effects of feedback on learning (79). Feedback that does not take circumstances or context into account, is poorly informed, is not followed up, causes emotional distress, or provides grades coupled with low levels of specificity (i.e., vague feedback) tend to impede learning (76, 81).

Specific (or elaborated) feedback which is linked to the performance on a particular task, and which includes information beyond its accuracy, has been shown to be significantly more effective (78, 82). However, a study by Phye and Sanders nuanced this view: In an experiment testing feedback specificity and learning, they found that specific feedback was superior to general advice on a retention task, but showed no differences on a transfer task (76, 83). A related finding, is that feedback length or complexity is inversely related to both its ability to correct errors and enhance learning (76, 84).

All learners are different and certain characteristics have been shown to influence how they utilise and benefit from feedback. Low-achieving students or beginners may need more support and explicit guidance (directive feedback), whereas high-achieving or more motivated students may benefit more from feedback that challenges them (facilitative feedback, such as hints and cues) (76).

Another important feedback variable is timing. Feedback can be given either immediately following a task, or delayed. The research into the effects of timing on learning are mixed, but immediate feedback has been shown to give more efficient learning, whereas delayed feedback is associated with better transfer of learning (76).

1.3.3 Workplace-based assessments (WBAs): Mini-Clinical Evaluation Exercise (mini-CEX)

It is evident that if we are to improve learning through clinical teaching, we are in need of direct observation, assessment and feedback to learners in this context. In line with constructive alignment, learning objectives and assessment methods should be made clear to learners beyond just participation.

An often cited model for classifying methods of assessment is George Miller's pyramid of competence (Figure 3) (85). At the base of the pyramid is 'knows', representing the assessment of factual knowledge required in the practice of medicine. The next layer, 'knows how', refers to the assessment of reasoning and application of such knowledge to specific situations. Where the base layers assess knowledge, or cognitive skills, the two top layers are concerned with assessment of performance (85). The third layer, 'shows how', requires learners to demonstrate that they can use their knowledge under supervision, typically assessed through simulation of professional tasks. The topmost layer, 'does' is the assessment of actual behaviour when functioning independently in clinical practice, often using workplace-based assessments (WBAs). When situated in real clinical practice, assessments are more subjective and rely to a larger extent on holistic judgements, but are authentic and relevant to students' professional development (33).

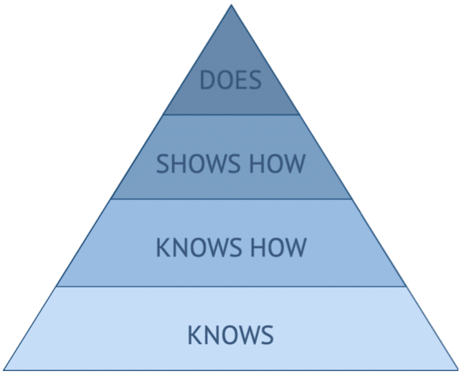


Figure 3 Miller's pyramid of clinical competence (85)

WBAs evaluate learners in a real-life clinical setting in terms of their development in clinical knowledge, skills and professionalism. Most of these observational assessment methods have a strong focus on feedback to the learner as an inherent part of the method itself, thereby situating their purpose towards the formative end of the spectrum. The mini-Clinical Evaluation Exercise (mini-CEX) is one of the most commonly used WBAs, and has been implemented in both undergraduate and postgraduate programmes since its introduction in 1995 (86, 87).

In the mini-CEX, trainees are observed and evaluated while performing clinical tasks (such as history taking and physical examination) in authentic patient encounters. This is usually followed by the trainee providing a summary of the encounter and which next steps he or she would take (e.g., a clinical diagnosis and management plan) (88). The mini-CEX is versatile and can be used for a wide range of clinical problems and workplace-based settings. Each assessment should take approximately 20-30 minutes, including observation, evaluation and feedback. This allows trainees to be evaluated multiple times with different patients, in different settings and by different assessors during their training period (88).

Trainees are assessed on six individual competencies that are important in high-quality patient care (history taking, physical examination, professionalism, clinical reasoning, counselling, and organisation and efficiency), as well as an overall score (86). Each competency is scored on a 9-point scale, where 1-3 is unsatisfactory, 4-6 is satisfactory and 7-9 is superior. The mini-CEX form used in the study in Paper II resembles the original, but also includes boxes for qualitative feedback (one for 'Especially Good' and one for 'Suggestions for Improvement'), and is provided in the Supplementary material.

1.3.4 Research on the mini-Clinical Evaluation Exercise (mini-CEX)

The mini-CEX was intended to be used in post-graduate assessments in general medicine. However, the mini-CEX form has been successfully changed and adapted to many local settings and contexts (87, 89-92).

There is a growing body of evidence for the reliability and validity of the mini-CEX as an assessment tool (87): Eight to ten encounters have been shown to yield acceptable reliability, the short time frame facilitates adequate domain sampling and the use of real patient encounters limit a number of common validity threats (93). However, as with other methods that involve ratings (including both performance tests and clinical observational methods), the mini-CEX is subject to classic rater errors such as severity/leniency, halo effects, central tendencies and restriction of the range (94 p. 46).

Despite the increasing use of the mini-CEX for formative purposes, there is limited evidence for its impact on learning. A systematic review on the educational impact of the mini-CEX published in 2018, found that the majority of studies report on effects on learner perceptions (95). Of these, most studies found that trainees were moderately or highly satisfied with the mini-CEX as a tool for learning.

Only three studies have investigated whether implementation of mini-CEX assessments can change learners' competence. Kim and colleagues compared mandatory formative mini-CEX to no or voluntary mini-CEX assessments during third-year clinical training, and found that failure rates were significantly lower on a summative clinical examination in the intervention group (96). Suhoyo and colleagues compared mandatory formative and summative mini-CEX assessments to the existing assessment programme (consisting of procedure lists, patient presentations and case reflections), and found that scores on a modified objective structured long examination record (OSLER) were significantly higher in the intervention group for internal medicine, but no significant difference was found in neurology (97). Since both of these studies adopt a sequential cohort design, it becomes challenging to establish causal relationships (95-97).

A third study by Karanth and colleagues used an experimental design, and compared performance on the traditional clinical evaluation exercise (CEX) between intervention and control group, and found a small statistically significant effect in favour of the intervention group (98). The intervention group underwent formative mini-CEX and direct observation of procedural skills (DOPS) assessments. However, it is unclear how students were assigned to groups, and the clinical teaching or assessments undertaken by the control group is not described.

The study in Paper II aims to address the limited research available on the educational outcomes of implementing formative mini-CEX assessments during clinical placements in undergraduate medical education. The study employs a randomised controlled design, which is notably absent in the existing literature on the mini-CEX.

1.4 Assessment practices

1.4.1 Challenges in assessment practices

Assessment is a key component of undergraduate medical education for several reasons. Summative assessments (i.e., assessment *of* learning) determine whether students have acquired the necessary knowledge, skills and attitudes to meet requirements for progression and certification. As such, it also serves as a quality assurance measure of the educational programme itself, ensuring that it maintains rigorous standards and produces competent doctors. Second, to guide students' learning process by setting clear expectations and standards for students to work towards, motivating students to engage actively in their education. Third, formative assessments (i.e., assessment *for* learning) offer students valuable feedback by highlighting areas where students need to improve and providing support and guidance on how to progress.

It is now widely recognised that no single assessment method can adequately assess all aspects of competence. The purpose of the assessment (i.e., which competencies we are trying to assess, whether the assessment is formative or summative), the stage within Miller's pyramid, and the resources available can all influence the choice of assessment methods (99). Multiple assessment methods must be chosen, in order to assess a breadth of competencies and compensate for the shortcomings of any one of the other methods (12). This purposeful selection of which competencies to assess, the specific assessment methods to use, and how these results are combined, is referred to as an assessment programme (100).

The concept of 'assessment drives learning' is widely cited in medical education literature, and continues to pose a significant challenge in developing assessment content and programmes. In face of an overwhelming curriculum, and particularly in the absence of useful learning outcomes, students resort to utilising previous examination questions to determine their learning focus and approach. Consequently, the assessment content serves as an indirect representation for students of what constitutes a competent doctor (6 p. 44). To encourage students towards meaningful and comprehensive learning, examinations should assess comprehension and application rather than mere reproduction of knowledge.

Assessment content must be relevant and reflect authentic clinical challenges students will encounter in their practice, if they are to be adequately prepared for their future roles as doctors. The assessment method and content should align with the stated learning objectives to ensure that they are being tested on knowledge and skills they are expected to acquire during their training. An additional challenge in health professions education is that knowledge and practices are constantly evolving, and assessment content needs to keep pace with the latest advancements and guidelines in the field.

1.4.2 Assessment methods

A wide variety of different assessment methods exist, and new methods or variations of well-known methods are frequently produced to suit different curricula and contexts. Written tests are useful for measuring cognitive knowledge, reflecting the 'knows' and 'knows how' levels of Miller's pyramid. Of these, multiple choice questions (MCQs) are the most frequently used.

MCQs consist of a stimulus question (stem) and a number of possible answers (options), of which one is correct and the others function as distractors (101 p. 657). The format generally requires less time than constructed-response formats, and MCQs can therefore test a large breadth of knowledge in a relatively short period of time, allowing for adequate sampling of the domain. They can easily be computer-scored and are therefore efficient for use in large groups of examinees.

In contrast to the traditional belief that only constructed-response items were able to test problem-solving ('knows how'), whereas selected-response items just tested factual knowledge ('knows'), it is now widely accepted that well-crafted MCQs are capable of measuring higher-cognitive knowledge (102). This follows the realisation that what is measured by an item is more determined by the stimulus format than by the response format (103). Stated simply, what is asked of the learner (the task) defines what is being measured to a larger degree than how the response is captured. However, in order to do this, MCQs must be constructed adhering to evidence-based principles (104).

1.4.3 Assessment quality

Recognising the significant impact of assessment on students' learning, both in terms of formative and summative evaluation, educators need to prioritise the pursuit of assessment quality. In 1996, Van der Vleuten proposed an equation to evaluate the utility of an assessment method (12):

Utility = Validity x Reliability x Educational impact x Acceptability x Cost

Utility is defined as the multiplicative function of these five components with their associated weights, all depending on the context and purpose of the assessment. The equation should be considered as a conceptual model where each component must be considered, and not as an actual equation with assigned numerical values. For example, educational impact should be weighted higher in formative assessments, and reliability higher in summative assessments. However, as Van der Vleuten argues, it has been deliberately designed as a multiplicative so that if one of the components is zero, the utility will also be zero (12). The five components later formed the basis for the Ottawa 2010 Conference consensus statement on criteria for good assessment (105).

Validity refers to whether a test accurately measures what it is intended to measure, and if its results can be meaningfully interpreted. The validity argument involves theoretical and empirical evidence from various sources to determine which inferences and actions based on the test results are reasonable (106). The Standards for Educational and Psychological Testing define five sources of validity evidence: content, response process, internal structure, relationship to other variables, and consequences (107). The extent of validity evidence required depends on the stakes of the assessment, with high-stakes tests needing a stronger and more detailed validity argument compared to low- and medium-stakes tests, such as formative assessments.

Messick identified two major threats to validity: construct underrepresentation and construct-irrelevant variance (108). Construct underrepresentation occurs when a test fails to adequately measure all aspects of the intended construct due to undersampling or biased sampling (109). This can happen if tests are too short or if items do not align with learning objectives, focusing on lower cognitive levels while learning objectives emphasise higher cognitive levels. On the other hand, construct-irrelevant variance introduces systematic errors unrelated to the construct being measured, leading to erroneous inferences and a systematic over- or underestimation of students' true test score (110, 111). This can affect both groups and individual examinees in different ways (111).

Reliability concerns random error in assessment data and how to quantify it (112). Although usually treated as separate entities, validity and reliability are closely related: reliability is a necessary, but not sufficient, source of validity evidence, as assessment data with a large component of random error will be meaningless for any use (113). Content-specificity is the primary factor influencing assessment reliability, and sufficient sampling is therefore necessary to produce reliable tests (114). Contrary to previous beliefs emphasising objectivity and standardisation, Van der Vleuten demonstrated that most assessment formats can achieve high reliability coefficients with adequate testing time and sampling (114). The required level of reliability depends on the test's purpose and importance (113).

Educational impact refers to an assessment method's influence on the learning process, and can be divided in pre-, post- and pure learning effects (115). Pre-assessment effects primarily relate to learning in summative assessment. Students' perception of the upcoming assessment influences their studying approach, guided by cues from lecturers, past exam papers, peers, and personal assessment experiences (116-118). Different assessment methods can influence how students prepare for an examination, with students recognising distinct cognitive processes assessed by various methods and adapting their strategies accordingly (119-122). Motivation to study is highest for moderately difficult material and for content important for future courses or work (118, 121). Post-assessment effects are more significant in formative assessments, and the impact of feedback is discussed in Chapter 1.3.2.

Pure learning effects, also known as the 'testing effect', demonstrate that combining studying with assessment greatly enhances learning (123). This has been consistently observed in laboratory settings using word lists

(124-126) and prose material (127, 128) since the 1970s. More recently, the concept has received attention in health professions education. A systematic review from 2018 on test-enhanced learning in health professions education found that the large majority of learning outcomes, including immediate learning outcomes, retention outcomes and transfer outcomes, favoured test-enhanced learning over studying (129). Recognising and leveraging the testing effect in both learning activities and assessment practices can greatly enhance learning.

Lastly, the aspect of acceptability revolves around stakeholders' endorsement of an assessment and its interpretation (130). Moreover, it is imperative to carefully consider the cost and feasibility of assessments to ensure their viability and practicality.

1.4.4 Quality assurance procedures

The preceding section explored the components of quality assessment, while this section will focus on essential aspects regarding its attainment. Quality assurance procedures around assessment such as item writing and item review is essential in ensuring quality assessment (131).

There are evidence-based principles for item writing to ensure valid and reliable assessments. Haladyna and colleagues have identified 31 principles for effective MCQs, with empirical support for about half of them (94 pp. 158-169, 104). They emphasise the importance of relevant and important item content aligned with learning objectives, testing higher-order cognitive levels, and avoiding clues or biases that may guide students to the correct answer. The stem should be concise but incorporate the information needed to answer the question, and options should be logical in ordering and homogenous in content and grammatical structure. Item-writing flaws (IWFs) is the term given to items that violate one or more of these principles. The principles are mostly studied for written tests, primarily selected-response items such as MCQs, but can also be generalised to constructed-response items.

Once an assessment item is created, and especially if its intended use is moderate- to high-stakes assessment, it should undergo a review and editing process to ensure adequate validity. Content review by experts, typically through a committee, is a common method to assess item quality. Reviewers evaluate the relevance of item content to the assessed field and document its alignment with test blueprint and learning objectives (132). Adherence to item writing guidelines and editorial style is also assessed, including considerations of language, grammar, and potential cultural biases. Collecting item performance statistics through pretesting is another method employed, often by embedding pretest items in actual examinations without affecting examinees' scores (132). Furthermore, a plan for fair and secure item administration and scoring should be established to maintain consistency among examinees (132). The final review of items occurs post-test through item analysis and feedback from examinees. Flawed items that may have evaded earlier reviews, such as incorrect answer keys or ambiguous wording, can be identified during this stage (133). Students should be encouraged to discuss, review and provide feedback on items before final scoring.

1.4.5 Research on assessment quality

The following review of the literature will centre on the quality of written assessments in medical education, which is the focus of Paper III. The existing literature mainly encompasses publications on the prevalence of item-writing flaws and their influence on student performance, guidelines for writing high-quality questions, and faculty development initiatives and review processes to improve the quality of items (134).

Numerous studies have documented the prevalence of poor item quality for in-house examinations in medical schools (135-142). These include items testing low cognitive levels, imprecise language, unfocused and negative stems, and other IWFs. In fact, Downing showed that around half of the items in basic science examinations for medical students were flawed, negatively impacting student achievement, introducing biased pass-fail decisions and limiting their educational value (141). Lack of motivation, time constraints, cost and logistics have been suggested as important barriers to writing quality items (134).

Item writer training is widely used in order to improve item quality, and there is evidence for its effect (134). Naeem and colleagues found a statistically significant increase in mean item quality scores after training (135). Abdulghani and colleagues found an improvement in psychometric properties and cognitive level of items post-training (136). In a study by Jozefowicz and colleagues, items by National Board of Medical Examiners (NBME)-trained writers were of significantly higher quality than writers without training (137). On the other hand, Iramaneerat showed that while there was a high satisfaction rate among participants in a series of short workshops, they did not result in statistically significant improvements in psychometric properties of items (143). The studies have a notable limitation in the form of small-scale interventions conducted over a short period of time and, in some cases, with voluntary enrollment. This limitation restricts the available evidence regarding the long-term effects and feasibility of training item writers in order to improve item quality.

Quality assurance procedures such as the use of guidelines for item writing and committee review that screens items and offer feedback to writers have also been reported to improve the quality of items (134). Wallach and colleagues found a significant increase in item quality after introducing guidelines and a committee review process (139). A similar study by Malau-Abduli and colleagues reported improved psychometric properties of items, including reliability and discrimination indices, after introducing a peer review process (144).

While we have evidence that low-quality items are common in in-house assessments, and that items can be improved through faculty development initiatives and item review processes, most studies focus on more technical aspects of item-writing (i.e., item formatting), and its effects on psychometric properties of the item or how well they align with established guidelines. Although several studies do report an increase in the number of items testing higher cognitive levels after undergoing review (136, 139), there is less focus in the literature on the relevance and importance of the actual test content, despite its consequence both to the validity of summative tests and its educational impact.

An additional challenge is the continuous demand for the production and review of new items that have not previously been used in examinations. This is especially relevant for institutions, like ours, that give students access to past examination papers. However, writing and reviewing items is resource-intensive, and new forms of quality assurance procedures must take feasibility into account. A review of the barriers and facilitators to writing quality items found that the sustained and wide-spread implementation of quality-assurance procedures was an area which lacked evidence (134). Many interventions are small and based on voluntary enrolment which makes it less likely to be useful long-term.

The objective of Paper III is to fill the research gap regarding the review of item content by presenting the findings of an external peer review of MCQs in an in-house setting. The review process was designed to accommodate the annual demand for new items with minimal resources, thereby taking feasibility into account.

2. Aims

The primary aim of this thesis was to develop and implement changes to teaching and assessment practices in the medical curriculum at the Norwegian University of Science and Technology (NTNU). The research was underpinned by constructivist learning theories and focused on evaluating effects on educational impact and assessment quality. This is done in the hope that improving student learning, and making valid decisions about their progress, will improve the quality of undergraduate medical education, and ultimately benefit patient health and safety. The following paragraphs describe the specific aims in each of the three papers.

Paper I: Development and evaluation of an active learning strategy for classroom teaching. TBL is an active learning strategy where students apply knowledge to solve authentic clinical problems, aligned with constructivist learning theories. While evidence supports its educational effect, challenges exist in its widespread implementation. Paper I presents the implementation of a modified and time-efficient TBL method called express TBL (eTBL), omitting the full RAT to focus in-class time on complex and authentic problems. **The study aims to examine eTBL's educational impact compared to traditional lectures in a neuroradiology course for third-year medical students, in addition to student perceptions of and engagement with eTBL.**

Paper II: Implementing systematic and structured formative assessments in clinical placements. Assessing and providing feedback to medical students during clinical placements hold significant potential. The mini-CEX offers an opportunity for active participation in authentic clinical scenarios, aligning with constructivist learning theories' emphasis on practical application and engagement. A number of studies have looked into the educational impact of the mini-CEX in terms of self-reported outcome measures, but few studies have measured its impact on performance. **This study compares mini-CEX assessments with traditional ad-hoc feedback during a 16-week clinical placement for fifth-year medical students, to examine its educational impact, effects on direct observation and feedback, and student perceptions of it as a formative tool.**

Paper III: A novel method for quality assurance of in-house examinations. Assessment content must be important and reflect authentic clinical challenges to enhance student learning and enable valid progress decisions. Quality assessment aligns content with learning objectives and activities, promoting meaningful learning. However, low-quality items are common in in-house undergraduate medical examinations, with limited focus on content quality assurance. **This study presents an analysis of the implementation of external peer review of MCQs by clinicians, aiming to examine how external review influences assessment quality. It explores to what extent clinicians consider MCQs acceptable for use in examinations, their feedback on items for revision or exclusion, and to what extent items are changed following review.**

3. Material and methods

3.1 Study setting

All three studies took place in the six-year undergraduate medical programme at NTNU, in Trondheim, Norway. The curriculum is an integrated and problem-based curriculum, with one oral and one written summative examination at the end of each academic year.

3.1.1 Curriculum

Years 1 and 2 cover most pre-clinical subjects, as well as an integrated clinical strand which includes patient communication, history taking and physical examination, as well as meeting patients in general practice on a regular basis. Years 3 and 4 cover most clinical subjects and paraclinical sciences, and students rotate through weekly clinical placements in the fields covered in their current term. The remaining teaching during the first four years is organised as mostly traditional lectures and weekly PBL sessions.

The structure of the clinical placement in Year 5 has been slightly changed, but at the time of study, it consisted of a 16-week clinical placement at one of the nine general hospitals in the region. It was divided into general medicine (7 weeks), general and orthopaedic surgery (7 weeks) and anaesthesia (2 weeks). All students were required to complete the same checklist of activities and procedures. The remaining term in Year 5 is spent conducting a research project. Year 6 consists of one term covering public health and primary care, which includes a 6-week placement in general practice, and a final term of summary and review before final examinations.

3.1.2 Assessment programme

Summative examinations are generally held at the end of each academic year, with the exception of Year 5 in which the assessment consists of a completed clinical placement checklist and research project. One further exception is that summative examinations are, for practical reasons, organised at the end of each term (twice yearly) in Years 4 and 6. Examinations are pass or fail, with a cut-off score of 65%.

Written examinations consist of 100-120 single best answer MCQs, and several modified essay questions (MEQ), with a total testing time of six hours. Oral examinations consist of an oral structured clinical examination (OSCE) for Years 1, 3 and 4. In Years 2 and 6 the oral examination take the form of long cases in which the students demonstrate history taking and physical examination of a simulated patient (Year 2) or two real patients (Year 6), and which in the latter is followed by a synthesis of their findings and formulation of a management plan.

Prior to the research project described in Paper III, the quality assurance procedures around test item development and administration was similar to that of the Maastricht model (145). The departments write items based on a blueprint, which are entered into a web-based item bank and reviewed by a multidisciplinary review committee for content, clarity and IWFs. Additionally, one or two senior students are asked to comment on the examination draft. Post-test analyses include item analysis and feedback from examinees, before the final scoring.

3.2 Overview of material and methods

Table 1 presents an overview of the aims, study design, materials and outcome measures of Papers I, II and III. The outcome measures have been divided into quantitative and qualitative components, which will be further explored in the next section on mixed methods research.

Table 1 Overview of material and methods.

	Paper I	*Paper II*	*Paper III*
Aims	To examine student perceptions of, and engagement fostered by, eTBL, and examine its educational impact compared with traditional lectures.	To examine student perceptions, effects on direct observation and feedback, and the educational impact of the mini-CEX compared with ad-hoc feedback.	To examine the effects of external review on assessment quality in terms of review decision and comments, and subsequent changes made to items.
Design	Experimental (2 x 2 cross-over study)	Experimental (RCT)	Case study
Materials	Data from summative examination scores of 105 students. Scores on a 17-item survey, including Student Self-Report of Engagement Measure, from 40 students.	Data from a summative OSCE and written test, and survey items, from 38 students. OSCE stations were filmed and later checklist-scored.	Data from internal and external review of 1353 MCQs.
Quantitative components	Statistical analysis of responses to a survey on student perceptions of eTBL and traditional lectures, and a survey on engagement during eTBL. Performance on assessment was compared between the groups (eTBL vs. traditional lecture).	Statistical analysis of responses to a survey on perceptions of the mini-CEX, and perceptions of direct observation and feedback during a clinical placement. Performance on assessments were compared between the groups (mini-CEX vs. no WBA).	Statistical analysis of reviewer decisions and subsequent changes made to items by item writers or the examination committee.
Qualitative components	n/a	Thematic analysis of written student comments on survey using STC.	Thematic analysis of reviewer comments using STC.
Reasoning	n/a	Qualitative components were used to gain insight into student perceptions (complementarity).	Qualitative components were used to gain insight into the reasoning behind reviewer decisions (expansion, complement-arity and triangulation).

eTBL: express Team-based learning; mini-CEX: mini-Clinical Evaluation Exercise; RCT: Randomised controlled trial; MCQ: Multiple choice question; OSCE: Oral Structured Clinical Examination; WBA: Workplace-based assessment; STC: Systematic text condensation.

3.2.1 Mixed methods research

In mixed methods research, quantitative and qualitative methods are intentionally combined to answer the research question (146). The research questions are often focused around understanding real-life contexts, where the quantitative methods enables the relationship between variables to be measured, and qualitative methods allow the topic to be explored in more depth. Using a pragmatic approach that often underpins mixed methods research, researchers are free to choose the methods that best meet their needs and purposes (147).

Greene and colleagues describe five major purposes for mixed methods research (148): (i) triangulation (seeking convergence and correspondence of results from different methods to increase validity); (ii) complementarity (seeking elaboration and illustration of the results from one methods with the results from the other method to increase meaningfulness); (iii) development (using the results from one method to inform the other method); (iv) initiation (seeking the discovery of paradoxes or contradictions to increase the depth of inquiry); and (v) expansion (seeking to extend the breadth of inquiry by using different methods).

In this thesis, a mixed methods approach has been used for Papers II and III, whereas Paper I utilises a purely quantitative approach. The quantitative data has been used to investigate magnitudes and relationships between variables. The qualitative data has been used to understand the experiences, perceptions and reasoning of the participants, recognising that an average is rarely representative of an entire group. An overview of the quantitative and qualitative components in Papers II and III can be found in Table 1.

In Paper III, the qualitative components (reviewer comments) have first and foremost been used for expansion and complementarity. That is, they allowed us to increase the scope of our study beyond the quantitative components of reviewer decision and whether changes were made to items. The comments illustrate and provide meaning to the statistical data obtained, letting us gain insight into *why* reviewers came to their decisions, and what this says about the assessment's quality. One could also argue that by cross-checking reviewer decision with the comments provided, the quantitative and qualitative data are triangulated against each other.

In Paper II, the qualitative components (written comments on perceptions of mini-CEX) have primarily been used for the purposes of complementarity. They provide depth and meaning to the fixed-response questions, which is especially important to uncover conflicting perceptions that are lost in average ratings.

3.2.2 Kirkpatrick's framework of educational outcomes

Kirkpatrick's framework is used to classify the educational outcomes of the interventions in Papers I and II (149). Although originally developed to measure outcomes of educational interventions in business, the framework has been widely implemented in medical education research (150).

The original model consists of four levels of outcome: learner reactions (on the learning experience), learning (changes in attitudes, knowledge and skills), behaviour (changes in practice and application of learning) and results (changes in organisational practice) (149). The framework has later been adapted for medical education research by Barr and colleagues (Table 2) (151). In this model, category two is modified to distinguish between modifications of attitudes and acquisition of knowledge and skills, and category four to distinguish between changes in organisational practice/delivery of care, and benefits to patients.

The medical education literature on educational outcomes have to a large degree focused on self-reported outcomes, that is learner reactions. However, experimental designs are increasingly being used to study educational impact on higher levels in the framework (152). It was important in this thesis not only to investigate learners' views and experiences (Level 1), but also document learning through their performance on assessments so as to study the acquisition of knowledge and skills (Level 2b).

In Paper I, learner reactions (Level 1) were examined through the use of a survey on perceptions of and their engagement during eTBL, and acquisition of knowledge (Level 2b) through performance on the summative written examination following the intervention. In Paper II, learner reactions (Level 1) were examined through the use of surveys on perceptions of the mini-CEX, direct observation and feedback. Performance on a written test (knowledge) and OSCE (skills) were used to examine effects on Level 2b.

Table 2 Kirkpatrick's framework of educational outcomes, adapted by Barr (151).

Level		Description
Level 1	Learner reactions	Learner views on the learning experience and satisfaction with the programme
Level 2a	Modification of attitudes	Changes in attitudes towards patients and their condition, circumstances and care
Level 2b	Acquisition of knowledge/skills	Acquisition of concepts, procedures and principles, thinking/problem-solving, psychomotor and social skills
Level 3	Changes in behaviour	Behavioural change transferred from the learning environment to the workplace.
Level 4a	Change in organisational practice	Changes in organisational practice or delivery of care, attributable to an educational programme
Level 4b	Benefits to patients/clients	Improvement in health and well-being of patients/clients

3.3 Paper I

The experimental study in Paper I is a description and analysis of the implementation of eTBL in neuroradiology teaching during Year 3. The aim was to examine its educational effect, student perceptions and to what extent it actively engages students.

3.3.1 Intervention

The phases of TBL and eTBL are shown in Figure 4. By reducing TBLs two-step method of readiness assurance (Phase 2) to a quick warm-up exercise of 10 MCQs that students answered individually, the in-class time for eTBL was shortened to 45 minutes (from 3 x 45-minute blocks for TBL). Most of the time was spent on application exercises (Phase 3), which were real clinical scenarios in neuroradiology and were solved in keeping with the 4S's principles.

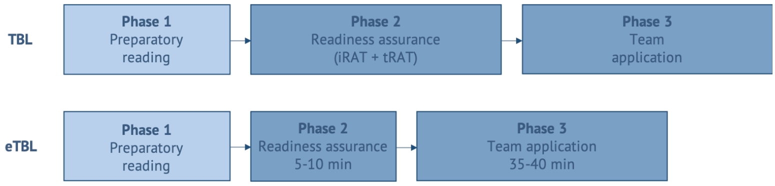


Figure 4 Phases of TBL and eTBL. The light fields represent out-of-class preparations, and the darker fields represent in-class time. iRAT: individual readiness assurance test; tRAT team readiness assurance test.

3.3.2 Participants

For the performance part, the participants were 105 third-year medical students at NTNU during the academic year of 2016/2017. The survey was answered by 40 third-year medical students in 2018.

3.3.3 Design and data collection

With the focus on trying to establish a causal relationship between the intervention and its educational impact, the study was conducted with an experimental design. In order for the study to be as realistic as possible, the intervention was imbedded in the curriculum and the summative examination was used as an outcome measure for performance. The students are already split into two groups at the start of the academic year, undertaking the same teaching but at different times. This setting leant itself to conduct a cross-over study, in which the groups change their respective arms (intervention or control) at a specific point during the study. This meant students did not need to be randomised and allocated into new groups, and also minimises the risk of confounding, as both intervention and control are measured on the same students.

The study in Paper I was conducted as a 2 x 2 cross-over study design (Figure 5). Neuroradiology during Year 3 had previously been taught in two 90-minute didactic lectures in computed tomography (CT) and magnetic resonance imaging (MRI) diagnostics. For the study, group 1 (*n* = 54) received teaching in CT diagnostics by a 90-minute didactic lecture and MRI diagnostics by a 45-minute eTBL session, during august 2016. Group 2 (*n* = 51) received teaching in CT diagnostics by a 45-minute eTBL session and MRI diagnostics by a 90-minute didactic lecture, during January 2017.

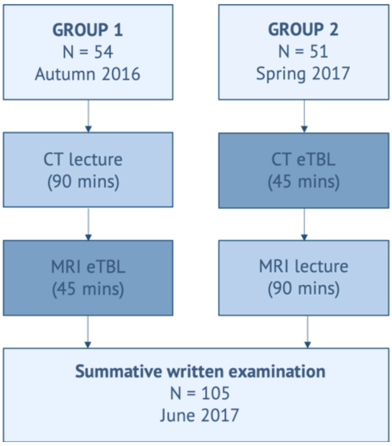


Figure 5 Study design: 2 x 2 cross-over study. Lecture: Traditional didactic lecture; eTBL: express Team-based learning; CT: Computed tomography; MRI: Magnetic resonance imaging.

Both groups (*n* = 105) sat the same summative written examination in June 2017. Neuroradiology was tested in one MEQ which consisted of seven sequential questions for a possible score of 10 points, accounting for 10% of the total score on the examination. The questions were divided into content covered in MRI teaching and content covered in CT teaching, which allowed for comparisons between the two groups. The MEQ was marked against a rubric by the item writer who was blinded as to what group the students belonged to. This accounted for educational impact at Level 2b (acquisition of knowledge) in Kirkpatrick's framework.

A 17-item survey was prepared to investigate self-reported measures of satisfaction and learning, accounting for educational impact at Level 1 in Kirkpatrick's framework. In order to investigate to what extent eTBL was able to foster active engagement in students, the survey included the nine-item Student Self-Report of Engagement Measure (153). This instrument has previously been validated against observed engagement and has shown good internal consistency (Cronbach's alpha of 0.84) (153).

3.4 Paper II

The experimental study in Paper II is a description and analysis of the implementation of mini-CEX assessments during the 16-week clinical placement for fifth-year students. The aim was to examine its educational impact, effects on direct observation and feedback, as well as student perceptions of the mini-CEX as a formative assessment tool.

3.4.1 Intervention

The students in the intervention group were expected to complete a minimum of eight formative mini-CEX assessments during their clinical placement. Since all participants and assessors had no prior experience with WBAs or the mini-CEX, a written guide and an introductory session including practical work was given. The students in the control group received ad-hoc feedback, as was the standard before mini-CEX assessments were introduced. The mini-CEX form used can be found in the Supplementary material, and is provided in both Norwegian and a translated version.

3.4.2 Participants

Six general hospitals in the region were enrolled. Participants were fifth-year medical students on their 16-week clinical placement in 2018. In total, 48 students were invited and 41 students consented to participate. Three students later withdrew because they were not able to attend outcome assessments, leaving a total of 38 participants, of which 19 students were in the intervention and 19 students were in the control group.

3.4.3 Design

We decided to conduct the study as a randomised controlled trial (RCT) in order to provide rigorous evidence about the relationship between the intervention and its educational impact.

The RCT conducted in Paper II uses cluster randomisation, in which hospitals, and not individual students, are allocated to intervention or control groups (Figure 6) (154). This was done for practical reasons and to avoid contamination, as doctors who received extra training in feedback using the mini-CEX could not be expected to treat individual students differently. For this reason, cluster randomised trials (CRTs) are often used for non-drug interventions, such as policy, service delivery and educational interventions (154).

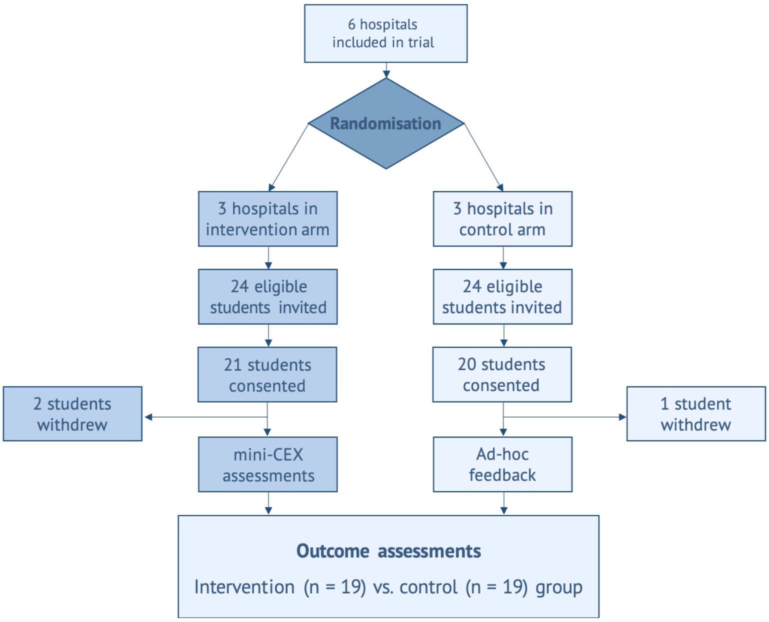

6 hospitals included in trial
Randomisation
3 hospitals in intervention arm
3 hospitals in control arm
24 eligible students invited
24 eligible students invited
21 students consented
20 students consented
2 students withdrew
1 student withdrew
mini-CEX assessments
Ad-hoc feedback
Outcome assessments
Intervention (n = 19) vs. control (n = 19) group

Figure 6 Study design: Cluster randomised trial. mini-CEX: mini-Clinical Evaluation Exercise.

One important implication of the CRT design, is that individuals within any one cluster are likely to be more homogenous than between clusters, and therefore more likely to respond to an intervention in a similar manner (155). This lack of independence leads to reduced statistical power compared with individual-level randomisation (155). In our study, we have chosen to base analyses on individual student-level data based on the assumption that because they choose the hospital for their clinical placement on the basis of a randomly assigned number, this ensures the necessary randomisation. However, this may have been an important factor in reducing the study's statistical power.

3.4.4 Data collection

The study follows a 'post-test only' design, in which the two groups are only tested on the outcome measure following the intervention. At the end of their clinical placement, all participants completed a survey (Kirkpatrick level 1), an 43-item MCQ test and a six-station OSCE (Kirkpatrick level 2b). Scores on the MCQ and OSCE test are presented both as observed scores, and scores controlled for baseline competence by using previous examination Z-scores. The assessments did not have any consequences for the students' progression. All outcome assessments were scored blinded to the intervention. In the case of the OSCE this was achieved by filming and later checklist-scoring the stations by two independent examiners.

The survey in Paper II consisted of two parts, and can be found in a translated English version in the Supplementary material. The first part was answered by both intervention and control group, and consisted of 40-Likert type questions and four free text questions. They were divided into three topics of interest: perceptions of feedback, perceptions of learning and confidence, and perceptions of motivation. The literature on feedback informed the design, and items were written adhering to best practices (156, 157). Cognitive interviews were also held with students who had recently finished their clinical placement, ensuring that questions were unambiguous and meaningful (158).

The second part of the survey consisted of eight Likert-type questions, four tick box questions and one free text question on perceptions of the mini-CEX, and were only answered by the intervention group. These questions were adapted with permission from Bindal and colleagues (159).

The qualitative data arising from the free text question in this part of the survey were analysed using systematic text condensation (STC, as described in Chapter 3.5.2) (160). The use of free text comments allowed the students to expand on their answers, as well as opening up for themes that were not necessarily covered by the fixed-response items. However, these types of data have limitations. Malterud points out that the context in which free text questions are asked, sends a strong message to the participants about what the researchers are after (161 pp. 204-205). Furthermore, leaving too little space for the answers, they can become short and meaningless, but leaving too much space can be daunting and may detract from any answer at all. Finally, written text leaves no room for further inquiry, which is a strength in interviews.

3.5 Paper III

The case study in Paper III is a description and analysis of a novel method of quality assurance of in-house examinations. The aim was to examine in what ways external peer review of MCQs by clinicians can influence assessment quality.

3.5.1 Intervention and data collection

The external peer review was first piloted for the end-of-year examinations in 2015, which comprised of 19 reviewers and 119 MCQs. The full scale project for the end-of-year examinations in 2016, which is described in Paper III, comprised of 170 reviewers and 1353 MCQs.

Junior doctors and general practitioners that did not write assessment items for the faculty were recruited as reviewers. The inclusion criteria were chosen so that reviewers represented the competence and work situation that medical students are being prepared for when newly qualified. Recruitment was done per e-mail and started among colleagues, and continued as snowball sampling with reviewers recommending their own colleagues. They received limited training on item writing and reviewing. There was no financial compensation, but they were given access to the university resources such as IT facilities.

The external peer review was added as an additional step in the quality assurance procedure already in place (Figure 7). Items had already been reviewed and approved by the multidisciplinary review committee before being subjected to external review. The external review was carried out double-blinded, so that reviewer and item writer did not know each other's identity.

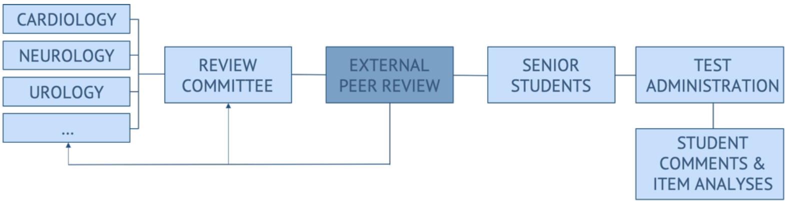


Figure 7 External peer review as part of the quality assurance procedure around examination development

External reviewers indicated whether an item should be approved, revised prior to use, or rejected. In the circumstance of the latter two, a comment explaining why had to be provided. Based on reviews, item writers decided whether to revise or delete the item, or leave it unchanged. If left unchanged, a comment explaining their decision had to be provided to the examination committee, who made the final decision on whether the item should be included in the examination. Review decision, reviewer comments and subsequent changes made to the item were registered.

3.5.2 Data analysis

The main part of the study focuses on a qualitative analysis of reviewer comments in order to gain insight into why reviewers judged items as rejected or needing review. Thematic analysis was chosen in order to identify themes and patterns of meaning within the qualitative data in relation to the research question.

Systematic text condensation (STC), as described by Malterud, is a method for thematic cross-case analysis of different types of qualitative data, including written text (160). The method is inspired by phenomenological ideas, and presents the experience of the participants as expressed by themselves, rather than exploring its possible underlying meaning.

The procedure follows several steps through decontextualization and recontextualization of the data: (i) reading through the entire material for an overall picture of the main themes; (ii) identifying and coding units of meaning; (iii) condensing and abstracting the meaning within each code group, and finally (iv) synthesising descriptions and concepts based on the condensates, making sure they reflect their original context (160).

For the study in Paper III, the analysis started by reading through all reviewers' comments to get an overall impression of the material. Meaningful text that represented the reviewers' reasons for not approving items was coded into main themes. Subsequently, subthemes within the main themes were identified, and the contents were condensed into artificial quotes representing the essence of each subtheme. Lastly, the content of each group was summarised in descriptions, and illustrated by selected quotes.

3.6 Statistics and data analysis

All quantitative data were systematised in Microsoft Excel and exported to IBM SPSS Statistics versions 24 and 25 (SPSS Inc., Chicago, IL, USA) for statistical analyses. The individual analyses are accounted for in each paper, and $p < 0.05$ was considered statistically significant throughout.

The qualitative data in Paper III (reviewer comments) were analysed 'manually' in Microsoft Word, whereas the qualitative data in Paper II (student comments in survey) were analysed using NVivo 11 (QRS International Pty Ltd., Melbourne, Australia).

3.7 Ethics

The studies in Paper II (project number: 56646) and Paper III (project number: 45229) were approved by the Norwegian Centre for Research Data (NSD). Approval for the study in Paper I was deemed unnecessary because only anonymous data was processed. In this study, student evaluations were anonymous and examination results were extracted and analysed anonymously, and in accordance with NSD guidelines.

4. Summary of work

This chapter presents a summary of the main finding from the three papers. The results are presented in more detail in each of the three papers.

4.1 Paper I

Smeby, S. S., Lillebo, B., Slørdahl, T. S., & Berntsen, E. M. (2020). Express Team-Based Learning (eTBL): A Time-Efficient TBL Approach in Neuroradiology. Academic radiology, 27(2), 284-290.

In this paper, we describe and analyse the implementation of a modified, time-efficient TBL approach (express TBL, eTBL) in neuroradiology for third-year medical students. In a cross-over study design, eTBL was compared with tradition lectures.

Student scores on the neuroradiology MEQ in the end-of-year examination were higher in the eTBL groups than lecture groups, but were not statistically significant. Median scores on MRI questions in the eTBL and lecture group were 2.5 and 2.0, respectively, but did not differ significantly ($p = 0.415$). Median scores on CT questions in the eTBL and lecture group were 4.0 and 3.5, respectively, but did not differ significantly ($p = 0.182$).

Students reported high levels of engagement during eTBL sessions, with a mean total score of 4.52 (out of a possible 5.00) on the Student Self-Report of Engagement Measure. Students indicated that they were more satisfied overall with eTBL than traditional lectures. They also rated eTBL higher than lectures on its ability to make difficult material comprehensible, ability to engage and give them feedback on their own knowledge. All comparisons based on student evaluations were statistically significant ($p < 0.001$).

Our study shows that eTBL fostered high levels of engagement and students reported significantly higher satisfaction and learning (Kirkpatrick level 1), but showed no improvement in performance on a summative examination (Kirkpatrick level 2b) compared with traditional lectures.

4.2 Paper II

Martinsen, S. S. S., Espeland, T., Berg, E., Samstad, E., Lillebo, B., & Slørdahl, T. S. (2021). Examining the educational impact of the mini-CEX: a randomised controlled study. BMC medical education, 21(1), 228.

Paper II describes the implementation of formative mini-CEX assessments during a 16-week clinical placement for fifth-year medical students. In a RCT design, mini-CEX assessments were compared with ad-hoc feedback.

Each participant in the intervention group completed a mean number of 8.4 mini-CEX assessments (range 8-10). The assessments covered a wide range of common clinical presentations and diagnoses. The majority (79%, 15/19) were positive or very positive to the use of mini-CEX assessments, and around 58% (11/19) found them useful or very useful in their clinical placement.

The analysis of the free-text comments in the survey, found that comments fell within two main themes: feedback and feasibility. Participants were divided in their perception of how useful the feedback had been, many commenting that they would have liked more constructive feedback on what could be improved. Some felt feedback from more experienced doctors to be more useful. However, several participants remarked on the value of the mini-CEX in terms of 'forcing' observation and feedback. Regarding feasibility, many participants felt that finding a time or suitable setting was challenging, and doctors were often too busy to conduct assessments. Some participants experienced that doctors did not know how to conduct assessments and give feedback.

Implementing formative mini-CEX assessments did not lead to reported increase of direct observation or feedback during the clinical placement overall, and were reported as infrequent in both groups. Statistical differences between the two groups were only found for two survey items: feedback on history taking was more commonly reported in the intervention group, and participants in the intervention group perceived their own ability to identify normal and abnormal findings higher than those in the control group.

There were no statistically significant differences between the two groups with regards to performance on the OSCE or written test. Observed mean scores on the OSCE were 3.4% higher in the intervention group, and when past OSCE Z-scores were controlled for, the difference between the group means decreased to 2.4%. Neither of these were statistically significant. Observed mean scores on the written test were 4.8% higher in the intervention group, and when past written examination Z-scores were controlled for, the difference between the group means decreased to 3.4%. Neither of these were statistically significant.

Our study shows that students were generally satisfied with mini-CEX assessments and their usefulness in learning (Kirkpatrick level 1). However, implementation did not lead to significant overall effects on direct observation, feedback, or performance (Kirkpatrick level 2) compared with ad-hoc feedback.

4.3 Paper III

Smeby, S. S., Lillebo, B., Gynnild, V., Samstad, E., Standal, R., Knobel, H., Vik, A., & Slørdahl, T. S. (2019). Improving assessment quality in professional higher education: Could external peer review of items be the answer?. Cogent Medicine, 6(1), 1659746.

In this paper, we describe and analyse the implementation of a novel quality assurance procedure suitable for in-house examinations: an external double-blinded peer review of MCQs by junior doctors and general practitioners. The review process was implemented in addition to the multi-disciplinary review committee already in place.

In all, 1353 items were reviewed by 170 external reviewers, of which 20% were either rejected or judged as needing revision by reviewers. Following review and feedback, changes were made to 40% of disapproved items, which constitutes almost 10% of the total number of MCQs that were reviewed.

Content relevance, content accuracy and technical flaws emerged as the main reasons for disapproving items. Content relevance refers to items that were flagged as unimportant, with inappropriate difficulty levels, irrelevant for clinical practice or only testing lower cognitive levels. Content accuracy refers to items that had content errors or were missing important information. Lastly, technical flaws in spelling, language or structure.

Our study shows that external peer review is cost-effective and feasible in an in-house setting, and can identify items that have the potential to significantly reduce the validity and educational impact of examinations.

5. Discussion

5.1 Applying constructivist learning theories to medical education

The overarching objective of this thesis was to develop and implement changes that were rooted in constructivist learning theories, and aimed to improve the quality of classroom teaching, clinical teaching and assessment practices. The fundamental goal of medical education is to facilitate students' learning, ensuring their preparedness to tackle the demands they will encounter as newly graduated professionals. To achieve this, teaching strategies and assessment practices must take into account the science of human learning, and incorporate evidence-based principles for curriculum design. Constructive alignment emerges as a logical foundation for achieving this objective, which draws upon the notion that students learn best when there is a clear connection between what they are expected to learn, how they are taught, and how their learning is assessed.

Teaching and learning activities must be designed to effectively facilitate the achievement of the learning outcomes that are set. In a constructivist approach, teaching strategies should encourage active learning, collaboration and inquiry, and enable continuous feedback and reflection. Despite this, medical students spend a vast amount of time in passive learning activities and seldom receive feedback on their knowledge or skills.

In Paper I, we developed and implemented a modified TBL approach, eTBL, to meet the challenges of passive learning and better align classroom teaching with learning objectives and assessment practices. The in-class exercises were constructed so as to integrate basic and clinical knowledge, and reflect authentic clinical scenarios that newly graduated doctors will meet. By using a validated self-report instrument, we showed that eTBL fostered high levels of engagement (mean total score of 4.52 out of a possible

5.00) (153). Students felt that they were actively involved, had contributed meaningfully, interacted with other students and paid attention during class. For active learning strategies to be successful, students need to be engaged with the material by interacting with each other and the instructor (162). Our findings are in line with other studies that report high levels of engagement for both full and modified implementations of TBL (57, 163-167).

Students reported that they were significantly more satisfied with eTBL than traditional lectures (Kirkpatrick level 1). Additionally, they rated eTBL significantly higher than lectures on its ability to make difficult material comprehensible, engage students and give them feedback on their own knowledge. The high levels of engagement and interaction reported in eTBL sessions may explain parts of their satisfaction. Previous research has shown that students are generally positive toward TBL, and emphasise the active learning style and interaction with other students (46). However, satisfaction may also be a result of simply introducing a different teaching strategy, and further research should see if these perceptions persist over time. Several studies have shown that learner satisfaction and how they perceive the usefulness of TBL decreases with time (47-49).

Just as in the classroom setting, clinical placements are arenas in which there is considerable potential to enhance learning. Their importance to medical education are highlighted by the fact that taking a medical history and performing a physical examination still remain the cornerstone of clinical practice (88). Despite this, procedure checklists and participation are often the only assessment criteria, and the lack of direct observation and feedback in clinical placements is well documented (168-170).

In Paper II, we implemented formative mini-CEX assessments during fifth-year clinical placements for medical students. This enabled students in the intervention arm to be observed, assessed and given feedback on their work in real-life clinical contexts. Used formatively in this manner, the mini-

CEX becomes a learning activity, enabling students to practice learning objectives under supervision, and which aligns well with the final year assessments. The clinical context places learning at the top of Miller's pyramid ('does') and aligns with constructive learning theories that emphasise students' need to apply their knowledge, receive feedback and reflect on their learning.

If formative mini-CEX assessments are to constructively align with learning objectives and assessments, they have to reflect a representative array of common clinical situations. In our study, almost all assessments were either history taking, clinical examinations or clinical case presentations, or a combination of the three. They were well spread across general medicine, general surgery and orthopaedics, and represented common patient complaints such as chest pain, shortness of breath, abdominal pain, fever and trauma. This is in line with other studies on the mini-CEX which have shown good content coverage of the cases and observations (86, 89, 171, 172).

The study showed that students were generally satisfied with mini-CEX assessments and their usefulness in terms of supporting learning (Kirkpatrick level 1). This is comparable to the literature on the mini-CEX, with a review article finding that all but one study reported trainee satisfaction from 6.0 to 8.8 on a 9-point Likert scale (172). Multiple participants in our study noted the significance of the mini-CEX in terms of its ability to 'force observation and feedback'. This emphasises that students perceive the lack of direct observation and feedback within their education, and find it challenging to request such opportunities. However, many participants commented that they would have liked more constructive feedback, and some felt that feedback from more experienced doctors would have been more useful. The nature of the feedback conversations, and the results on educational impact will be discussed in the next section.

Just as authenticity and relevance are important in learning activities within the classroom and clinical setting, they are equally essential when it comes to assessment practices. Ensuring that item content is relevant is perhaps the most important quality criterium, and means that they should reflect learning objectives and professional practice. It is necessary for ensuring the validity of assessments, and thereby making defendable decisions about students' progression or licensure. Just as important are the implications on the educational impact of assessments.

In Paper III, we describe and analyse the implementation of a novel quality assurance procedure suitable for in-house examinations: an external double-blinded peer review of MCQs by junior doctors and general practitioners. The review process was implemented after items had been approved by an internal multi-disciplinary review committee, and reviewers were primarily asked to consider the item's content.

In all, 1353 items were reviewed by 170 external reviewers, of which 20% were either rejected or judged as needing revision by reviewers. Following review and feedback, changes were made to 40% of disapproved items, which constitutes almost 10% of the total number of MCQs that were reviewed. Content relevance, content accuracy and technical flaws emerged as the main reasons for disapproving items.

Content relevance refers to items that were flagged as unimportant, with inappropriate difficulty levels, irrelevant for clinical practice or only testing lower cognitive levels. In in-house assessments, item relevance is strongly influenced by item writers' individual perceptions and experiences (173). Since item-writers frequently work as experienced clinicians or researchers, this often leads to the incorporation of trivial, detailed and specialised knowledge items in tests in undergraduate medical education (174). From medical expert theory, we know that these items contribute little to learning, do not support the encapsulation process or the formation of illness scrips, and do not aid students in fostering transfer of knowledge (26). Instead, they primarily assess rote memorisation of isolated facts, which directly contradicts the learning objectives that highlight comprehension of fundamental concepts and problem-solving. In an effort to reduce trivial content in progress tests, Janssen-Brandt and colleagues found that the use of a rubric to define item relevance altered the judgement on inclusion of that item in the test, by students, staff and item reviewers (174).

Content accuracy refers to items that had content errors or were missing important information. Although the number of items with errors, such as being based on outdated guidelines or classification systems, were small, they pose a significant threat to the validity of examinations and increases the likelihood of students learning erroneous information. With the rapid growth of medical knowledge, items that are stored in an item bank for later use quickly become outdated (175).

Junior doctors and general practitioners were recruited for the external review process. This selection was based on the rationale that if the reviewers represented generalists or clinicians at the early stages of their specialisation, it would increase the likelihood that examinees would share their conviction of the item's relevance to practice. When students perceive the assessments as meaningful and applicable to their future careers, they are more likely to invest time and effort in their learning. Consequently, the external review may serve to enhance both the acceptability of assessments by students, the validity of the inferences made from test scores and their educational impact. Furthermore, it provides an external source of data to defend decisions about progression or licensure.

It is important to note that while the study showed that clinicians identified a large number of items that have the potential to significantly reduce the validity and educational impact of examinations, its final effect on assessment quality comes from the changes that subsequently were made.

Although our study identified that many items were changed following review, the nature of these changes were not studied. In addition, its effects on psychometric measures, long-term effects on item quality and generalisability have yet to be decided.

In this way, the introduction of eTBL in classroom teaching, formative mini-CEX assessments in clinical teaching and an external peer-review of assessment content, have all served to incorporate constructivist learning theories and the concept of constructive alignment in the medical curriculum. However, is there evidence that it has improved student outcomes?

5.2 Student outcomes as a measure of quality

The other aim of this thesis has been to evaluate the effects of the interventions in terms of educational impact.

In Paper I, the educational impact of eTBL, compared with traditional lectures, were evaluated through student performance on the end-of-year written examination in a cross-over design (Kirkpatrick level 2b). Despite eTBL showing high levels of engagement and student satisfaction, there were no statistically significant differences in examination scores. This is in line with a systematic review of 14 studies in health professions education, in which seven studies showed improved knowledge scores in the TBL-group, but the remaining seven studies showed no differences (52). The lack of impact can be explained by several factors: the intervention was small with only one eTBL session in each arm of the study, students may have learned the material at other points in the curriculum, or may have compensated for the use of ineffective teaching methods through extensive preparations before summative examinations. Furthermore, written assessments fail to assess other aspects that TBL promote, such as teamwork and communication skills.

An interesting study by Rotgans and colleagues found that cognitive engagement fluctuates during a TBL session, but that students were significantly more engaged when working together during the tRAT and team application exercises (163). Additionally, they found that cognitive engagement was a significant predictor of performance on a subsequent knowledge test. In eTBL, the original version of TBL has been modified for time-saving reasons by leaving the tRAT out. However, the findings by Rotgans suggest that this may not be preferable in terms of either engagement or educational impact. However, two other studies have not been able to find the effects of tRAT on learning (61, 176).

In Paper II, the educational impact of formative mini-CEX assessments, compared with ad-hoc feedback in clinical placements, was evaluated by student performance on an OSCE and written test following the intervention. We showed that performance on both the OSCE and the written test were slightly higher in the intervention group, though not statistically significant. The absence of an effect may be explained by several factors. Firstly, the intervention may be too small to realistically expect significant differences. Secondly, the use of general outcome measures may have left a large part of the effect invisible, and the skills learned during assessments may not have been transferable to the outcome measures. Thirdly, and maybe most importantly, it is natural to think that the educational impact of the mini-CEX is heavily reliant on the quality of the feedback given.

Feedback can be a powerful phenomenon, especially when it is specific, positively angled and timely (77, 81). However, its effects on learning are inconsistent and complex, and it often fails to reach its potential (76, 79). This may be because feedback is delivered poorly and fails to provide the student with task-oriented information on how he or she is doing. In our study, the intervention group did not report higher frequency or quality of feedback during their clinical placement compared to the control group, despite having completed a mean of eight mini-CEX assessments. The free-text comments provided by participants in the intervention group show that perceptions of the usefulness of feedback from mini-CEX assessments varied, with some expressing disappointment in its limited utility and many highlighting the absence of constructive feedback. Furthermore, some participants reported instances where doctors lacked training in conducting assessments and delivering feedback. These findings suggest that there were clear limitations with regards to the implementation of the formative assessments, and that many of the feedback conversations did not follow evidence-based principles.

The research into feedback conversations in mini-CEX assessments have shown mixed results, but several studies indicate trainee dissatisfaction with its quality. A study in postgraduate medical education found that trainees valued the mini-CEX for facilitating specific feedback and appreciated its timeliness, and assessors found giving feedback easier with the mini-CEX (177). However, the same study reported that some assessors rated the mini-CEX without justifying their scores or discussing their feedback with the trainee. In another study, 27% of foundation programme doctors had 'rarely' or 'never' received feedback on their performance in the assessment, with only around one quarter viewing the mini-CEX as a useful means of gaining feedback (178). Inconsistent quality and depth of feedback, as well as a perception of bias and lack of honesty for assessments, were highlighted in a study on the use of the mini-CEX for general practice specialist trainees (179).

In addition to the quality of the feedback delivered, students need to actively respond to the feedback provided in order for assessments to effectively serve as a formative tool, thereby closing the feedback loop (180). Without this active utilisation of feedback to make improvements, neither students nor those giving feedback can know if it has been effective.

Receptivity to feedback has been shown to increase with authentic and relevant assessments, and by appropriate scaffolding and mentoring to aid the interpretation of feedback (181). On the other hand, several studies have identified barriers to the uptake and use of feedback in the context of summative assessment (181-183). Two studies suggest that although the purpose of assessments was formative, they were still perceived as summative by students (184, 185). These factors can have been instrumental in the uptake and use of feedback in our study.

Lastly, the relationship between the student and assessor has been shown to influence the impact of feedback. How learners perceive the credibility of the source clearly influences the impact of feedback: learners discount feedback to a larger extent from supervisors whom they feel lack clinical knowledge or experience (186). Consistent with this finding, some participants in our study reported that the feedback would have been more useful if it had come from more experienced doctors. Furthermore, many studies have shown that a trustful and long-term relationship between learner and supervisor (the 'educational alliance') that can secure continuity in assessment, feedback provision and its follow-up, is important for its effectiveness (187, 188). This is in stark contrast to how mini-CEX assessments were implemented in our study, where there was no way of securing this continuity and support.

Although Paper III did not investigate effects on student learning, we can explore how the enhanced validity of summative tests might impact learning. The challenges in feedback uptake in the context of summative assessments have been discussed previously, but it is the pre- and pure assessment effects that could be expected to have the most significant impact on learning (182). Students readily use past examination papers to guide their learning towards exams, making the content's alignment with learning objectives that emphasise problem-solving and understanding fundamental concepts, crucial in driving meaningful learning instead of rote memorisation. There is also evidence that how students perceive the purpose, credibility and value of assessments impact to what degree they facilitate learning (186, 189).

A significant part of this thesis concerns investigating educational outcomes using experimental and controlled designs. Despite strong theoretical grounding in constructivist learning theories, neither Paper I or Paper II showed improvements in student outcomes. This may well lead one to consider: are student outcomes good measures of quality in education?

The discussion on the use of experimental studies to examine learner outcomes in medical education is not a new one (152, 190-193). Many researchers highlight that education and learning are complex entities, consisting of multiple components that mutually impact each other in complex ways, highly dependent on the context (193). Some researchers have doubted the role of RCTs in medical education, stating that they are not feasible in educational research, and that 'treatment effects' in RCTs may potentially be lost in quantities of unexplained variance and may not even be detectible (190, 191).

In fact, 'no statistically significant difference' in educational interventions is the rule, rather than the exception, when comparing effectiveness of different curricula or teaching methods (194). One possible explanation was provided by Ten Cate when he highlighted that blinding in educational research is impossible: "*... indeed if they* were *unaware of the instruction, this is a sign of the ultimate failure of the intervention. And once aware, students may compensate for educational interventions"* (194). Medical students are well aware of what is happening to them, and are both highly motivated and well equipped to compensate for any curriculum (195). Already in 1976, McLeish stated that irrespective of differences in teaching methods, superior or inferior, the studying that students do for themselves in preparation for examinations will generally mean their scores are close to equal (196).

In this thesis (and many works before us), the assumption that the educational interventions act as the independent variable and the test to measure student outcome as the dependent variable, is probably too simple (194). Ten Cate argues that the test not only operates as a dependent variable, but also as an independent variable at the same time (194). The test is part of the curriculum, and drives learning in differential and unknowable ways in both the intervention and control groups. The true dependent variable is in fact what the student *does* in response to the interventions, for as previously argued, it is student activity and not the teaching itself that leads to learning. Therefore, the research question should in fact be: What student behaviour do these interventions promote?

5.3 Feasibility of quality improvement

The feasibility of interventions in medical education is decisive in its ability to improve quality. Even the most exceptional teaching strategies and assessment practices will yield limited impact on quality if they are too resource-intensive to be sustainably implemented.

In Paper I, TBL was modified to allow for a more time-efficient administration of TBL. Consequently, classroom hours were reduced from 90 to 45 minutes, and by reducing the RAT to a short warm-up exercise, more of the in-class time could be spent on problem-solving. Although this study cannot conclude on whether this modification improves implementation in the curriculum, it does address some of the challenges that have been identified with regards to its use (58).

In Paper II, mini-CEX assessments were chosen because feasibility has been identified as one of its strengths (86, 172). The assessments lend themselves to being used with a wide range of clinical problems and in different workplace settings, and resemble the interaction between trainees and supervisors in the clinical setting (172). We found that all participants met the pre-planned number of assessments, and completion times for evaluation and feedback were in line with the intention (86).

In our study, we chose to allow all doctors to carry out mini-CEX assessments, instead of placing this task on a few senior doctors. A shift of educational tasks to junior doctors in order to cope with increased workload for clinical staff has been successful elsewhere (197-199). Both intervention and control group indicated that feedback was most frequently given by junior doctors. However, some participants in the intervention group felt that feedback from more experienced doctors would be more useful. Additionally, both doctors and students were only given a short introduction to the assessments and the research project, which certainly eases implementation, but may have had detrimental effects on the quality of the delivery and uptake of feedback resulting from the mini-CEX assessments.

In Paper III, the implementation of the peer review process within an in-house setting was heavily influenced by feasibility considerations due to limited financial and staff resources. The number of items reviewed each year closely aligns with the number of MCQs required for our annual examinations and reassessments. Ensuring that each reviewer had a small and manageable workload, and that an IT solution was in place to facilitate remote work, was important for recruiting reviewers as they were not financially compensated. Additionally, the IT solution which supported the entire review process effectively reduced administrative costs.

5.4 Strengths and limitations

This thesis has a number of strengths. Firstly, it addresses significant and global challenges in undergraduate medical education, namely the need to integrate learning theory and evidence-based practices to address the issues of curriculum overload and passive learning in classroom teaching, the lack of direct observation and feedback in clinical teaching, and the importance of reviewing assessment content to ensure its relevance in in-house examinations. The research conducted within this thesis has been rigorous, employing several different methodologies that facilitate triangulation and enhance the robustness of the findings.

Several limitations also exist, many of which have already been discussed in the respective papers. All studies were conducted within a single institution, which restricts the extent to which the findings can be generalised. The studies on educational impact (Paper I and Paper II) were small and under-powered, increasing the risk of failing to detect true effects (type II-errors). Moreover, the interventions were likely too small, especially in Paper I, to expect educationally significant results. The limitations of using student outcomes as a quality indicator has already been discussed. Additional studies are needed to replicate our findings in other settings, and further research with longer-term studies is needed.

5.5 Implications for medical education and future research

Providing answers to the research questions outlined in this thesis has proven challenging: although there are results that indicate that the interventions have integrated learning theories and constructive alignment within the curriculum, their effects on learning remain unclear. Although our studies did not show a direct improvement in student outcomes as a result of TBL or mini-CEX assessments, it does not negate the substantial evidence that supports the impact on learning from active learning and feedback given the right conditions.

Although shortcomings in the research design leave us with limited insight into the educational impact of eTBL, the high levels of student engagement and satisfaction observed in eTBL sessions present compelling reasons for its wider implementation, not only within our institution but also in other educational settings. The modified TBL approach effectively addresses several challenges to TBL implementation identified in the literature (58). Future research should delve into the extent to which eTBL facilitates active learning integration in the classroom setting, particularly by investigating the teacher's role. This includes examining whether it reduces the time spent on preparing teaching material, whether it makes classroom management more attainable due to the shorter duration of classes, and whether it enhances the likelihood of teachers transitioning from traditional lectures to eTBL.

Once a more sustained implementation of eTBL is established, its effects on student learning can be studied. This research could investigate whether eTBL influences student behaviours in terms of out-of-class studying, and explore whether the inclusion of complex problems in eTBL shifts students' focus away from rote memorisation towards deep learning. Lastly, the implementation of eTBL may prompt a change in assessment content created by teachers, leading to a shift towards authentic problems and the assessment of higher-order cognitive skills, thereby aligning better with learning objectives.

In clinical teaching, many institutions are implementing mini-CEX assessments and other WBAs in order to improve the frequency and quality of feedback within their curriculums. The study in Paper II shows that simply implementing a WBA does not necessarily lead to more or better feedback, nor improved student outcomes. Most studies on the mini-CEX mention an orientation programme for familiarising the trainees and assessors with the tool, similar to the limited training that doctors received in our study (172). However, the qualitative results in our study indicate that there were significant shortcomings in the quality of feedback conversations. Training for both students and assessors have been shown to improve delivery and uptake of feedback, and institutions should therefore consider investing in staff development programmes alongside the implementation of WBAs (200, 201).

In light of the discussion on student outcomes, further research should explore feedback delivery and receptivity, and changes in behaviour following the implementation of formative assessments such as the mini-CEX. For example, do students view mini-CEX assessments as truly formative? What kinds of feedback are students getting, and is it actionable? What are the barriers and facilitators for the uptake of feedback following mini-CEX? Does continuity and mentorship facilitate the delivery and uptake of feedback?

The study presented in Paper III highlights the importance of conducting content review of in-house examinations. The findings demonstrate that clinicians can recognise items that have the potential to significantly reduce the validity and educational impact of locally developed examinations. Medical programmes should consider involving clinicians from outside the academic staff in the review of assessment content.

It is worth noting that item development and review procedures employed in national examinations are typically costly and go well beyond the resources available for in-house examinations. The external review process introduced in Paper III may offer a viable alternative within these limitations, and can prove to be feasible and beneficial for enhancing the quality of in-house examinations.

Future research should investigate what changes were made to items following peer review, and whether these indeed increased their quality. In addition, whether external peer review can affect measured such as reliability and item discrimination, and its long-term effects on item quality. It would also be of interest to examine whether stakeholder perceptions on the acceptability of assessments change following the introduction of external peer review. This goes for students, academic staff and item-writers, as well as patients and the wider public. Finally, this is a single-institution study, and whether its feasibility and findings replicate to other medical curricula and settings should be explored.

5.6 Concluding remarks

Despite the advancements in our understanding of human learning during the past century, many educational practices in the field of medicine are not yet firmly rooted in educational research. Traditional classroom teaching continues to rely heavily on lecture-based methods, while clinical teaching often lacks adequate opportunities for assessment and feedback on clinical skills. Furthermore, a significant portion of in-house assessment content fails to adequately reflect important and authentic clinical challenges, compromising its validity and failing to fully leverage its potential impact on student learning.

This thesis has made contributions to the development and implementations of innovative approaches in medical education. These include a modified TBL approach in classroom teaching, the incorporation of formative mini-CEX assessments in clinical teaching, and the introduction of an external peer review process for assessment items. Grounded in constructivist learning theories, these projects were designed with the primary aim of enhancing the quality of learning within the medical curriculum. Throughout this thesis, a central focus has been placed on feasibility to ensure that the proposed changes are viable within a realistic educational setting. Although the search for evidence demonstrating educational impact in terms of improved student outcomes presented challenges, it sparked numerous inquiries for future exploration.

6. References

1. Vroeijenstijn A. Quality assurance in medical education. Academic Medicine. 1995;70(7 Suppl):S59-67; discussion S8.

2. Kenwright DN, Wilkinson T. Quality in medical education. In: Swanwick T, Forrest K, O'Brien B, editors. Understanding Medical Education: Evidence, Theory, and Practice: Wiley-Blackwell; 2018. p. 101-10.

3. Harvey L, Green D. Defining quality. Assessment & evaluation in higher education. 1993;18(1):9-34.

4. Schindler L, Puls-Elvidge S, Welzant H, Crawford L. Definitions of Quality in Higher Education: A Synthesis of the Literature. Higher Learning Research Communications. 2015;5(3).

5. Donabedian A. Evaluating the Quality of Medical Care. The Milbank Memorial Fund Quarterly. 1966;44(3):166-206.

6. Kvernenes M, Schei E. Legers læring. Veileder i medisinsk pedagogikk. [Doctors' learning. A guide to medical pedagogy]. Oslo: Fagbokforlaget; 2022.

7. Dauphinee WD, Wood-Dauphinee S. The need for evidence in medical education: the development of best evidence medical education as an opportunity to inform, guide, and sustain medical education research. Academic Medicine. 2004;79(10):925-30.

8. Ludmerer KM. Commentary: understanding the Flexner report. Academic Medicine. 2010;85(2):193-6.

9. Papa FJ, Harasym PH. Medical curriculum reform in North America, 1765 to the present: a cognitive science perspective. Academic Medicine. 1999;74(2):154-64.

10. Cooke M, Irby DM, O'Brien BC. Educating Physicians: A Call for Reform of Medical School and Residency. San Francisco, CA: Jossey-Bass; 2010.

11. Flexner A. Medical Education in the United States and Canada. New York: Carnegie Foundation for the Advancement of Teaching; 1910.

12. van der Vleuten CPM. The assessment of professional competence: developments, research and practical implications. Advances in Health Sciences Education. 1996;1(1):41-67.

13. Norman G, Tugwell P, Feightner J, Muzzin LJ, Jacoby L. Knowledge and clinical problem-solving. Medical Education. 1985;19(5):344-56.

14. Schmidt HG, Norman GR, Boshuizen HP. A cognitive perspective on medical expertise: theory and implication. Academic medicine. 1990;65(10):611-21.

15. Schuwirth LW, van der Vleuten CPM. A history of assessment in medical education. Advances in Health Sciences Education. 2020;25(5):1045-56.
16. Hager P, Gonczi A. What is competence? Medical teacher. 1996;18(1):15-8.
17. Fernandez N, Dory V, Ste-Marie LG, Chaput M, Charlin B, Boucher A. Varying conceptions of competence: an analysis of how health sciences educators define competence. Medical education. 2012;46(4):357-65.
18. Wimmers PF. Developing Clinical Competence [Doctoral degree]. Rotterdam, Netherlands: Erasmus Universiteit Rotterdam; 2006.
19. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. Medical teacher. 2010;32(8):638-45.
20. Mann K, MacLeod A. Constructivism: learning theories and approaches to research. In: Cleland J, Durning SJ, editors. Researching Medical Education: John Wiley & Sons; 2015. p. 49-66.
21. Piaget J. The Child's Conception of the World. London: Routledge & Kegan Paul Ltd; 1929.
22. Kaufman DM. Teaching and learning in medical education: how theory can inform practice. In: Swanwick T, Forrest K, O'Brien B, editors. Understanding Medical Education: Evidence, Theory, and Practice: Wiley-Blackwell; 2018. p. 37-69.
23. Vygotsky LS, Cole M. Mind in society: Development of higher psychological processes: Harvard university press; 1978.
24. Brown JS, Collins A, Duguid P. Situated cognition and the culture of learning. In: Murphy P, Selinger M, Bourne J, Briggs M, editors. Subject Learning in the Primary Curriculum: Issues in English, science and mathematics. 18: Routledge; 1995. p. 301-19.
25. Van Merriënboer JJ, Sweller J. Cognitive load theory in health professional education: design principles and strategies. Medical education. 2010;44(1):85-93.
26. Schuwirth LW, van der Vleuten CPM. General overview of the theories used in assessment: AMEE Guide No. 57. Medical teacher. 2011;33(10):783-97.
27. Schmidt HG, Rikers RM. How expertise develops in medicine: knowledge encapsulation and illness script formation. Medical education. 2007;41(12):1133-9.
28. Chi MT, Feltovich PJ, Glaser R. Categorization and representation of physics problems by experts and novices. Cognitive science. 1981;5(2):121-52.
29. Hoover WA. The practice implications of constructivism. SEDL letter. 1996;9(3):1-2.

30. Biggs J. Enhancing teaching through constructive alignment. Higher Education. 1996;32(3):347-64.
31. Biggs J, Tang C. Teaching for Quality Learning at University: McGraw-hill education; 2011.
32. Snyder BR. The hidden curriculum. 1970.
33. van der Vleuten CPM, Sluijsmans D, Joosten-ten Brinke D. Competence assessment as learner support in education. Competence-based vocational and professional education: Springer; 2017. p. 607-30.
34. Boud D. Enhancing Learning Through Self-Assessment. London: Routledge; 1995. p. 39.
35. Norcini JJ, Banda SS. Increasing the quality and capacity of education: the challenge for the 21st century. Medical education. 2011;45(1):81-6.
36. Studentbarometeret [Student Survey] 2022 [Available from: https://www.studiebarometeret.no/no/student/studieprogram/1150_cmed-trondheim/1130_medisin-tromso/1120_prmedisin/.
37. Custers EJ. Long-term retention of basic science knowledge: a review study. Advances in Health Sciences Education. 2010;15(1):109-28.
38. Dunkin MJ. A review of research on lecturing. Higher education research and development. 1983;2(1):63-78.
39. Freeman S, Eddy SL, McDonough M, Smith MK, Okoroafor N, Jordt H, et al. Active learning increases student performance in science, engineering, and mathematics. Proceedings of the national academy of sciences. 2014;111(23):8410-5.
40. van der Vleuten CP, Driessen EW. What would happen to education if we take education evidence seriously? Perspectives on medical education. 2014;3:222-32.
41. Carpenter SK, Wilford MM, Kornell N, Mullaney KM. Appearances can be deceiving: Instructor fluency increases perceptions of learning without increasing actual learning. Psychonomic bulletin & review. 2013;20:1350-6.
42. Fornari A, Poznanski A. How-to guide for active learning: International Association of Medical Science Educators (IAMSE); 2015.
43. Michaelsen LK. Team learning: A comprehensive approach for harnessing the power of small groups in higher education. To improve the academy. 1992;11(1):107-22.
44. Parmelee D, Michaelsen LK, Cook S, Hudes PD. Team-based learning: a practical guide: AMEE guide no. 65. Medical teacher. 2012;34(5):e275-e87.
45. Hrynchak P, Batty H. The educational theory basis of team-based learning. Medical teacher. 2012;34(10):796-801.

46. Reimschisel T, Herring AL, Huang J, Minor TJ. A systematic review of the published literature on team-based learning in health professions education. Medical teacher. 2017;39(12):1227-37.
47. Fujikura T, Takeshita T, Homma H, Adachi K, Miyake K, Kudo M, et al. Team-based learning using an audience response system: a possible new strategy for interactive medical education. Journal of Nippon Medical School. 2013;80(1):63-9.
48. Zgheib NK, Simaan JA, Sabra R. Using team-based learning to teach pharmacology to second year medical students improves student performance. Medical teacher. 2010;32(2):130-5.
49. Moore-Davis TL, Schorn MN, Collins MR, Phillippi J, Holley S. Team-Based Learning for Midwifery Education. Journal of Midwifery & Women's Health. 2015;60(3):291-7.
50. Currey J, Oldland E, Considine J, Glanville D, Story I. Evaluation of postgraduate critical care nursing students' attitudes to, and engagement with, Team-Based Learning: A descriptive study. Intensive and Critical Care Nursing. 2015;31(1):19-28.
51. Feingold CE, Cobb MD, Arnold J. Student perceptions of team learning in nursing education. Journal of Nursing Education. 2008;47(5):214.
52. Fatmi M, Hartling L, Hillier T, Campbell S, Oswald AE. The effectiveness of team-based learning on learning outcomes in health professions education: BEME Guide No. 30. Medical teacher. 2013;35(12):e1608-e24.
53. Chung E-K, Rhee J-A, Baik Y-H. The effect of team-based learning in medical ethics education. Medical Teacher. 2009;31(11):1013-7.
54. Koles P, Nelson S, Stolfi A, Parmelee D, DeStephen D. Active learning in a year 2 pathology curriculum. Medical education. 2005;39(10):1045-55.
55. Koles PG, Stolfi A, Borges NJ, Nelson S, Parmelee DX. The impact of team-based learning on medical students' academic performance. Academic Medicine. 2010;85(11):1739-45.
56. Kang KA, Kim SJ, Oh J, Kim S, Lee MN. Effectiveness of simulation with team-based learning in newborn nursing care. Nursing & health sciences. 2016;18(2):262-9.
57. Tan NC, Kandiah N, Chan YH, Umapathi T, Lee SH, Tan K. A controlled study of team-based learning for undergraduate clinical neurology education. BMC medical education. 2011;11(1):91.
58. Burgess A, Matar E. Team-based learning (TBL): Theory, planning, practice, and implementation. In: Nestel D, Reedy G, McKenna L, Gough S, editors. Clinical Education for the Health Professions: Theory and Practice: Springer; 2020. p. 1-29.

59. Lillebo B, Slørdahl TS, Nordrum IS. Team-based learning: An active learning and instructor-led method [Teambasert læring–en studentaktiviserende og lærerstyrt undervisningsform]. Uniped. 2017;40(03):207-21.
60. Haidet P, Kubitz K, McCormack WTJJoeict. Analysis of the team-based learning literature: TBL comes of age. 2014;25(3-4):303.
61. Carbrey JM, Grochowski COC, Cawley J, Engle DLJJoeefhp. A comparison of the effectiveness of the team-based learning readiness assessments completed at home to those completed in class. 2015;12.
62. Agarwal P. Retrieval Practice & Bloom's Taxonomy: Do Students Need Fact Knowledge Before Higher Order Learning? Journal of Educational Psychology. 2019;111:189-209.
63. McDaniel M, Thomas R, Agarwal P, McDermott K, Roediger H. Quizzing in Middle-School Science: Successful Transfer Performance on Classroom Exams. Applied Cognitive Psychology. 2013;27:360-72.
64. Bailey JH, Rutledge B. The Educational Psychology of Clinical Training. The American Journal of the Medical Sciences. 2017;353(2):96-100.
65. Cameron A, Millar J, Szmidt N, Hanlon K, Cleland J. Can new doctors be prepared for practice? A review. The Clinical Teacher. 2014;11(3):188-92.
66. Prince KJ, Boshuizen HP, Van Der Vleuten CP, Scherpbier AJ. Students' opinions about their preparation for clinical practice. Medical education. 2005;39(7):704-12.
67. Moss F, McManus I. The anxieties of new clinical students. Medical Education. 1992;26(1):17-20.
68. Walker L, Haldane J, Alexander D. A medical curriculum: evaluation by final-year students. Medical Education. 1981;15(6):377-82.
69. Prince KJ, van De Wiel M, Scherpbier AJ, can der Vleuten CP, Boshuizen HP. A qualitative analysis of the transition from theory to practice in undergraduate training in a PBL-medical school. Advances in Health Sciences Education. 2000;5:105-16.
70. Peters M, ten Cate O. Bedside teaching in medical education: a literature review. Perspectives on Medical Education. 2014;3(2):76-88.
71. Blood AD, Farnan JM, Fitz-William W. Curriculum Changes and Trends 2010–2020: A Focused National Review Using the AAMC Curriculum Inventory and the LCME Annual Medical School Questionnaire Part II. Academic Medicine. 2020;95(9S).
72. Pulito AR, Donnelly MB, Plymale M, Mentzer J, Robert M. What do faculty observe of medical students' clinical performance? Teaching and learning in medicine. 2006;18(2):99-104.

73. Howley LD, Wilson WG. Direct Observation of Students during Clerkship Rotations: A Multiyear Descriptive Study. Academic Medicine. 2004;79(3).

74. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: The do’s, don’ts and don’t knows of direct observation of clinical skills in medical education. Perspectives on Medical Education. 2017:1-20.

75. Braend AM, Gran SF, Frich JC, Lindbaek M. Medical students’ clinical performance in general practice – Triangulating assessments from patients, teachers and students. Medical Teacher. 2010;32(4):333-9.

76. Shute VJ. Focus on formative feedback. Review of educational research. 2008;78(1):153-89.

77. Hattie J, Timperley H. The power of feedback. Review of educational research. 2007;77(1):81-112.

78. Kulhavy RW, Stock WA. Feedback in written instruction: The place of response certitude. Educational psychology review. 1989;1(4):279-308.

79. Kluger AN, DeNisi A. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological bulletin. 1996;119(2):254.

80. Azevedo R, Bernard RM. A meta-analysis of the effects of feedback in computer-based instruction. Journal of Educational Computing Research. 1995;13(2):111-27.

81. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do’s, don’ts and don’t knows of feedback for clinical education. Perspectives on medical education. 2015;4(6):284-99.

82. Bangert-Drowns RL, Kulik C-LC, Kulik JA, Morgan M. The instructional effect of feedback in test-like events. Review of educational research. 1991;61(2):213-38.

83. Phye GD, Sanders CE. Advice and feedback: Elements of practice for problem solving. Contemporary Educational Psychology. 1994;19(3):286-301.

84. Kulhavy RW, White MT, Topp BW, Chan AL, Adams J. Feedback complexity and corrective efficiency. Contemporary educational psychology. 1985;10(3):285-91.

85. Miller GEJAm. The assessment of clinical skills/competence/performance. 1990;65(9):S63-7.

86. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Annals of internal medicine. 1995;123(10):795-9.

87. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. Jama. 2009;302(12):1316-26.

88. Norcini J, Burch V. Workplace-based assessment as an educational tool: AMEE Guide No. 31. Medical teacher. 2007;29(9-10):855-71.

89. Alves de Lima A, Barrero C, Baratta S, Castillo Costa Y, Bortman G, Carabajales J, et al. Validity, reliability, feasibility and satisfaction of the Mini-Clinical Evaluation Exercise (Mini-CEX) for cardiology residency training. Medical teacher. 2007;29(8):785-90.

90. Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Medical education. 2008;42(4):364-73.

91. Prins SH, Brøndt SG, Malling B. Implementation of workplace-based assessment in general practice. Education for Primary Care. 2019;30(3):133-44.

92. Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. Academic Medicine. 2003;78(10):S33-S5.

93. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-Clinical Evaluation Exercise: a review of the research. Academic Medicine. 2010;85(9):1453-61.

94. Downing SM, Yudkowsky R. Assessment in Health Professions Education. New York: Routledge; 2009.

95. Lörwald AC, Lahner F-M, Nouns ZM, Berendonk C, Norcini J, Greif R, et al. The educational impact of Mini-Clinical Evaluation Exercise (Mini-CEX) and Direct Observation of Procedural Skills (DOPS) and its association with implementation: A systematic review and meta-analysis. PloS one. 2018;13(6).

96. Kim S, Willett LR, Noveck H, Patel MS, Walker JA, Terregino CA. Implementation of a mini-CEX requirement across all third-year clerkships. Teaching and learning in medicine. 2016;28(4):424-31.

97. Suhoyo Y, Schönrock-Adema J, Rahayu GR, Kuks JB, Cohen-Schotanus J. Meeting international standards: a cultural approach in implementing the mini-CEX effectively in Indonesian clerkships. Medical teacher. 2014;36(10):894-902.

98. Karanth K, Kanagasabai S, Ibrahim S, Najimuddin M, Marasinghe D, De S. Structured program for final-year undergraduate students to improve clinical skills to prepare for effective patient management. Internet J Gynecol Obstet. 2015;19(1).

99. Lockyer J, Carraccio C, Chan M-K, Hart D, Smee S, Touchie C, et al. Core principles of assessment in competency-based medical education. Medical Teacher. 2017;39(6):609-16.

100. Dijkstra J, van der Vleuten CPM, Schuwirth L. A new framework for designing programmes of assessment. Advances in health sciences education. 2010;15(3):379-93.

101. Downing SM. Assessment of knowledge with written test forms. International handbook of research in medical education: Springer; 2002. p. 647-72.

102. Palmer EJ, Devitt PG. Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. BMC medical education. 2007;7(1):1-7.

103. Schuwirth LW, van der Vleuten CPM. Different written assessment methods: what can be said about their strengths and weaknesses? Medical Education. 2004;38(9):974-9.

104. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education. 2002;15(3):309-33.

105. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Medical teacher. 2011;33(3):206-14.

106. Downing SM. Validity: on the meaningful interpretation of assessment data. Medical education. 2003;37(9):830-7.

107. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Standards for educational and psychological testing. Washington, DC: American Educational Research Association; 1999.

108. Messick S. Validity. In: Linn R, editor. Educational measurement, 3rd ed. The American Council on Education/Macmillan series on higher education.: American Council on Education; 1989. p. 13-103.

109. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Medical education. 2004;38(3):327-33.

110. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. Advances in Health Sciences Education. 2002;7(3):235-41.

111. Haladyna TM, Downing SM. Construct-irrelevant variance in high-stakes testing. Educational Measurement: Issues and Practice. 2004;23(1):17-27.

112. Tavakol M, Dennick R. Post-examination interpretation of objective test data: Monitoring and improving the quality of high-stakes examinations: AMEE Guide No. 66. Medical Teacher. 2012;34(3):e161-e75.

113. Downing SM. Reliability: on the reproducibility of assessment data. Medical education. 2004;38(9):1006-12.

114. van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: from methods to programmes. Medical education. 2005;39(3):309-17.

115. Dochy F, Segers M, Gijbels D. Assessment engineering: Breaking down barriers between teaching and learning, and assessment. Rethinking assessment in higher education: Routledge; 2007. p. 97-110.
116. Cilliers FJ, Schuwirth LW, Herman N, Adendorff HJ, van der Vleuten CPM. A model of the pre-assessment learning effects of summative assessment in medical education. Advances in Health Sciences Education. 2012;17(1):39-53.
117. Entwistle NJ, Entwistle A. Contrasting forms of understanding for degree examinations: the student experience and its implications. Higher education. 1991;22(3):205-27.
118. Van Etten S, Freebern G, Pressley M. College students’ beliefs about exam preparation. Contemporary Educational Psychology. 1997;22(2):192-212.
119. Sambell K, McDowell L. The construction of the hidden curriculum: messages and meanings in the assessment of student learning. Assessment & Evaluation in Higher Education. 1998;23(4):391-402.
120. Scouller K. The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. Higher Education. 1998;35(4):453-72.
121. Sambell K, McDowell L, Brown S. “But is it fair?”: an exploratory study of student perceptions of the consequential validity of assessment. Studies in educational evaluation. 1997;23(4):349-71.
122. Watkins D. Factors influencing the study methods of Australian tertiary students. Higher Education. 1982;11(4):369-80.
123. Karpicke JD, Roediger HL. The critical importance of retrieval for learning. science. 2008;319(5865):966-8.
124. Hogan RM, Kintsch W. Differential effects of study and test trials on long-term recognition and recall. Journal of Verbal Learning and Verbal Behavior. 1971;10(5):562-7.
125. McDaniel MA, Masson ME. Altering memory representations through retrieval. Journal of Experimental Psychology: Learning, Memory, and Cognition. 1985;11(2):371.
126. Carrier M, Pashler H. The influence of retrieval on retention. Memory & Cognition. 1992;20(6):633-42.
127. Glover JA. The" testing" phenomenon: Not gone but nearly forgotten. Journal of Educational Psychology. 1989;81(3):392.
128. Roediger III HL, Karpicke JD. Test-enhanced learning: Taking memory tests improves long-term retention. Psychological science. 2006;17(3):249-55.
129. Green ML, Moeller JJ, Spak JM. Test-enhanced learning in health professions education: a systematic review: BEME Guide No. 48. Medical teacher. 2018;40(4):337-50.

130. Norcini JJ, McKinley DW. Assessment methods in medical education. Teaching and teacher education. 2007;23(3):239-50.

131. van der Vleuten CPM, Schuwirth L, Scheele F, Driessen E, Hodges B. The assessment of professional competence: building blocks for theory development. Best Practice & Research Clinical Obstetrics & Gynaecology. 2010;24(6):703-19.

132. Downing SM, Haladyna TM. Test item development: Validity evidence from quality assurance procedures. Applied Measurement in Education. 1997;10(1):61-82.

133. Baranowski RA. Item editing and editorial review. In: Downing SM, Haladyna TM, editors. Handbook of test development: Routledge; 2011. p. 363-72.

134. Karthikeyan S, O'Connor E, Hu W. Barriers and facilitators to writing quality items for medical school assessments–a scoping review. BMC Medical Education. 2019;19:1-11.

135. Naeem N, van der Vleuten C, Alfaris EA. Faculty development on item writing substantially improves item quality. Advances in health sciences education. 2012;17:369-76.

136. Abdulghani HM, Ahmad F, Irshad M, Khalil MS, Al-Shaikh GK, Syed S, et al. Faculty development programs improve the quality of multiple choice questions items' writing. Scientific reports. 2015;5.

137. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Academic Medicine. 2002;77(2):156-61.

138. Abozaid H, Park YS, Tekian A. Peer review improves psychometric characteristics of multiple choice questions. Medical Teacher. 2017;39(sup1):S50-S4.

139. Wallach PM, Crespo L, Holtzman K, Galbraith R, Swanson D. Use of a committee review process to improve the quality of course examinations. Advances in Health Sciences Education. 2006;11(1):61-8.

140. Holsgrove G, Elzubeir M. Imprecise terms in UK medical multiple-choice questions: what examiners think they mean. Medical Education. 1998;32(4):343-50.

141. Downing SM. The effects of violating standard item writing principles on tests and students: the consequences of using flawed test items on achievement examinations in medical education. Advances in Health Sciences Education. 2005;10(2):133-43.

142. Downing SM. Construct-irrelevant Variance and Flawed Test Questions: Do Multiple-choice Item-writing Principles Make Any Difference? Academic Medicine. 2002;77(10).

143. Iramaneerat C. The impact of item writer training on item statistics of multiple-choice items for medical student examination. Siriraj Medical Journal. 2012;64(6):178-82.

144. Malau-Aduli BS, Zimitat C. Peer review improves the quality of MCQ examinations. Assessment & Evaluation in Higher Education. 2012;37(8):919-31.

145. Verhoeven B, Verwijnen G, Scherpbier A, Schuwirth L. Quality assurance in test construction: The approach of a multidisciplinary central test committee/Commentary. Education for Health. 1999;12(1):49.

146. Creswell JW, Klassen AC, Plano Clark VL, Smith KC. Best practices for mixed methods research in the health sciences. Bethesda (Maryland): National Institutes of Health. 2011;2013:541-5.

147. Creswell JW. Research design: Qualitative, quantitative, and mixed methods approaches - 2nd ed. Thousand Oaks, CA: Sage; 2003.

148. Greene JC, Caracelli VJ, Graham WF. Toward a conceptual framework for mixed-method evaluation designs. Educational evaluation and policy analysis. 1989;11(3):255-74.

149. Kirkpatrick DL. Evaluation of Training. In: Craig R, Bittel L, editors. Training and Development Handbook. New York: McGraw Hill; 1967.

150. Yardley S, Dornan T. Kirkpatrick's levels and education 'evidence'. Medical education. 2012;46(1):97-106.

151. Barr H, Freeth D, Hammick M, Koppel I, Reeves S. Evaluations of interprofessional education. London: United Kingdom Review of Health and Social Care. 2000.

152. Cook DA. Randomized controlled trials and meta-analysis in medical education: what role do they play? Medical teacher. 2012;34(6):468-73.

153. O'Malley KJ, Moran BJ, Haidet P, Seidel CL, Schneider V, Morgan RO, et al. Validation of an observation instrument for measuring student engagement in health professions settings. Evaluation & the health professions. 2003;26(1):86-103.

154. Hemming K, Eldridge S, Forbes G, Weijer C, Taljaard M. How to design efficient cluster randomised trials. bmj. 2017;358.

155. Campbell MK, Mollison J, Steen N, Grimshaw JM, Eccles M. Analysis of cluster randomized trials in primary care: a practical approach. Family Practice. 2000;17(2):192-6.

156. Artino Jr AR, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE Guide No. 87. Medical teacher. 2014;36(6):463-74.

157. Artino Jr AR, Gehlbach H, Durning SJ. AM last page: avoiding five common pitfalls of survey design. Academic Medicine. 2011;86(10):1327.

158. Collins D. Pretesting survey instruments: an overview of cognitive methods. Quality of life research. 2003;12(3):229-38.

159. Bindal T, Wall D, Goodyear HM. Trainee doctors’ views on workplace-based assessments: Are they just a tick box exercise? Medical teacher. 2011;33(11):919-27.

160. Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scandinavian journal of public health. 2012;40(8):795-805.

161. Malterud K. Qualitative research methods for medicine and health sciences [Kvalitative forskningsmetoder for medisin og helsefag]. 4th ed. Oslo: Universitetsforlaget; 2017.

162. Carini RM, Kuh GD, Klein SP. Student engagement and student learning: Testing the linkages. Research in higher education. 2006;47(1):1-32.

163. Rotgans JI, Schmidt HG, Rajalingam P, Hao JWY, Canning CA, Ferenczi MA, et al. How cognitive engagement fluctuates during a team-based learning session and how it predicts academic achievement. 2017:1-13.

164. Sharma A, Janke KK, Larson A, Peter WS. Understanding the early effects of team-based learning on student accountability and engagement using a three session TBL pilot. Currents in Pharmacy Teaching and Learning. 2017;9(5):802-7.

165. Balwan S, Fornari A, DiMarzio P, Verbsky J, Pekmezaris R, Stein J, et al. Use of team-based learning pedagogy for internal medicine ambulatory resident teaching. Journal of graduate medical education. 2015;7(4):643-8.

166. Kelly PA, Haidet P, Schneider V, Searle N, Seidel CL, Richards BF. A comparison of in-class learner engagement across lecture, problem-based learning, and team learning using the STROBE classroom observation tool. Teaching and learning in medicine. 2005;17(2):112-8.

167. Haidet P, Morgan RO, O'malley K, Moran BJ, Richards BF. A controlled trial of active versus passive learning strategies in a large group setting. Advances in Health Sciences Education. 2004;9(1):15-27.

168. Kassebaum DG, Eaglen RH. Shortcomings in the evaluation of students' clinical skills and behaviors in medical school. Academic medicine: journal of the Association of American Medical Colleges. 1999;74(7):842-9.

169. Kogan JR, Hauer KE. Brief report: Use of the mini-clinical evaluation exercise in internal medicine core clerkships. Journal of general internal medicine. 2006;21(5):501-2.

170. Daelmans H, Hoogenboom R, Donker A, Scherpbier A, Stehouwer C, van der Vleuten CPM. Effectiveness of clinical rotations as a learning environment for achieving competences. Medical teacher. 2004;26(4):305-12.

171. Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. Medical education. 2006;40(10):950-6.
172. Mortaz Hejri S, Jalili M, Masoomi R, Shirazi M, Nedjat S, Norcini J. The utility of mini-Clinical Evaluation Exercise in undergraduate and postgraduate medical education: A BEME review: BEME Guide No. 59. Medical Teacher. 2019:1-18.
173. Norcini J, Grosso L. The generalizability of ratings of item relevance. Applied Measurement in Education. 1998;11(4):301-9.
174. Janssen-Brandt XM, Muijtjens AM, Sluijsmans DM. Toward a better judgment of item relevance in progress testing. BMC medical education. 2017;17(1):151.
175. Sadaf S, Khan S, Ali SK. Tips for developing a valid and reliable bank of multiple choice questions (MCQs). Education for Health. 2012;25(3):195.
176. Gopalan C, Fox DJ, Gaebelein CJJAipe. Effect of an individual readiness assurance test on a team readiness assurance test in the team-based learning of physiology. 2013;37(1):61-4.
177. Weller JM, Jones A, Merry AF, Jolly B, Saunders D. Investigation of trainee and specialist reactions to the mini-Clinical Evaluation Exercise in anaesthesia: implications for implementation. British journal of anaesthesia. 2009;103(4):524-30.
178. Jackson D, Wall D. An evaluation of the use of the mini-CEX in the foundation programme. British Journal of Hospital Medicine (2005). 2010;71(10):584-8.
179. Sabey A, Harris M. Training in hospitals: what do GP specialist trainees think of workplace-based assessments? Education for Primary Care. 2011;22(2):90-9.
180. Sadler DR. Formative assessment and the design of instructional systems. Instructional science. 1989;18(2):119-44.
181. Harrison CJ, Könings KD, Dannefer EF, Schuwirth LWT, Wass V, van der Vleuten CPM. Factors influencing students' receptivity to formative feedback emerging from different assessment cultures. Perspectives on Medical Education. 2016;5(5):276-84.
182. Harrison CJ, Könings KD, Schuwirth L, Wass V, van der Vleuten CPM. Barriers to the uptake and use of feedback in the context of summative assessment. Advances in Health Sciences Education. 2015;20(1):229-45.
183. Harrison CJ, Könings KD, Molyneux A, Schuwirth LW, Wass V, van der Vleuten CP. Web-based feedback after summative assessment: how do students engage? Medical education. 2013;47(7):734-44.
184. Bok HGJ, Teunissen PW, Favier RP, Rietbroek NJ, Theyse LFH, Brommer H, et al. Programmatic assessment of competency-based

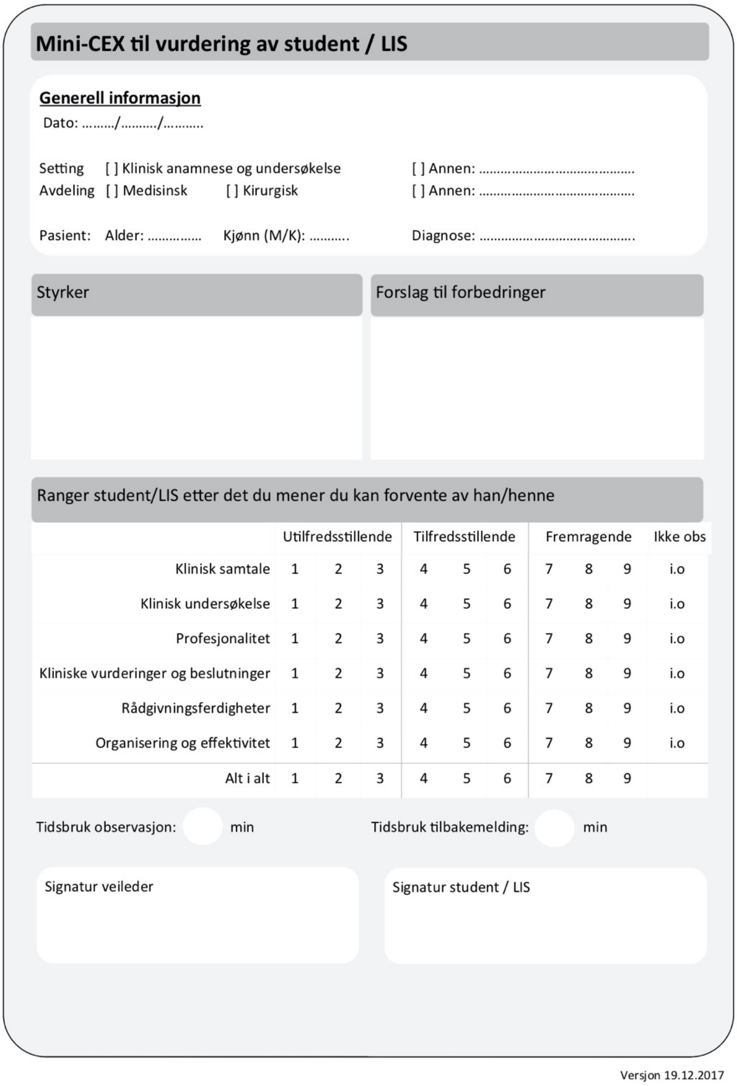
workplace learning: when theory meets practice. BMC Medical Education. 2013;13(1):123.
185. Heeneman S, Oudkerk Pool A, Schuwirth LW, van der Vleuten CP, Driessen EW. The impact of programmatic assessment on student learning: theory versus practice. Medical education. 2015;49(5):487-98.
186. Long S, Rodriguez C, St-Onge C, Tellier P-P, Torabi N, Young M. Factors affecting perceived credibility of assessment in medical education: A scoping review. Advances in Health Sciences Education. 2022;27(1):229-62.
187. Telio S, Ajjawi R, Regehr G. The "educational alliance" as a framework for reconceptualizing feedback in medical education. Academic Medicine. 2015;90(5):609-14.
188. Boud D. Feedback: ensuring that it leads to enhanced learning. The Clinical Teacher. 2015;12(1):3-7.
189. Ricci M, St-Onge C, Xiao J, Young M. Students as stakeholders in assessment: how students perceive the value of an assessment. Perspectives on Medical Education. 2018;7(6):352-61.
190. Thistlethwaite J, Davies H, Dornan T, Greenhalgh T, Hammick M, Scalese R. What is evidence? Reflections on the AMEE symposium, Vienna, August 2011. Medical Teacher. 2012;34(6):454-7.
191. Norman G. RCT= results confounded and trivial: the perils of grand educational experiments. Medical education. 2003;37(7):582-4.
192. Torgerson CJ. Educational research and randomised trials. Medical Education. 2002;36(11):1002-3.
193. Wong G, Greenhalgh T, Westhorp G, Pawson R. Realist methods in medical education research: what are they and what can they contribute? Medical education. 2012;46(1):89-96.
194. Ten Cate O. What happens to the student? The neglected variable in educational outcome research. Advances in Health Sciences Education. 2001;6(1):81-8.
195. Norman G. Best evidence medical education and the perversity of human as subjects. Adv Health Sci Educ Theory Pract. 2001;6:1-3.
196. McLeish J. The lecture method. Gage NL, editor: University of Chicago Press; 1976.
197. Hill AG, Yu TC, Barrow M, Hattie J. A systematic review of resident-as-teacher programmes. Medical Education. 2009;43(12):1129-40.
198. Ilgen JS, Takayesu JK, Bhatia K, Marsh RH, Shah S, Wilcox SR, et al. Back to the Bedside: The 8-year Evolution of a Resident-as-Teacher Rotation. The Journal of Emergency Medicine. 2011;41(2):190-5.
199. Dunne B, Smyth P, Furlong H, Rakovac-Tisdall A, Murphy D, Sreenan S. Interns as teachers of medical students: a pilot programme. Irish journal of medical science. 2011;180:211-4.

200. Liao K-C, Pu S-J, Liu M-S, Yang C-W, Kuo H-P. Development and implementation of a mini-Clinical Evaluation Exercise (mini-CEX) program to assess the clinical competencies of internal medicine residents: from faculty development to curriculum evaluation. BMC medical education. 2013;13:1-7.

201. Weston PS, Smith CA. The use of mini-CEX in UK foundation training six years following its introduction: lessons still to be learned and the benefit of formal teaching regarding its utility. Medical teacher. 2014;36(2):155-63.

7. Supplementary material

The mini-CEX form (Norwegian version)



Mini-CEX til vurdering av student / LIS

Generell informasjon

Dato: ………/………./……….

Setting [ ] Klinisk anamnese og undersøkelse [ ] Annen: ……………………………………
Avdeling [ ] Medisinsk [ ] Kirurgisk [ ] Annen: ……………………………………

Pasient: Alder: ………….. Kjønn (M/K): ………. Diagnose: ……………………………………….

Styrker

Forslag til forbedringer

Ranger student/LIS etter det du mener du kan forvente av han/henne

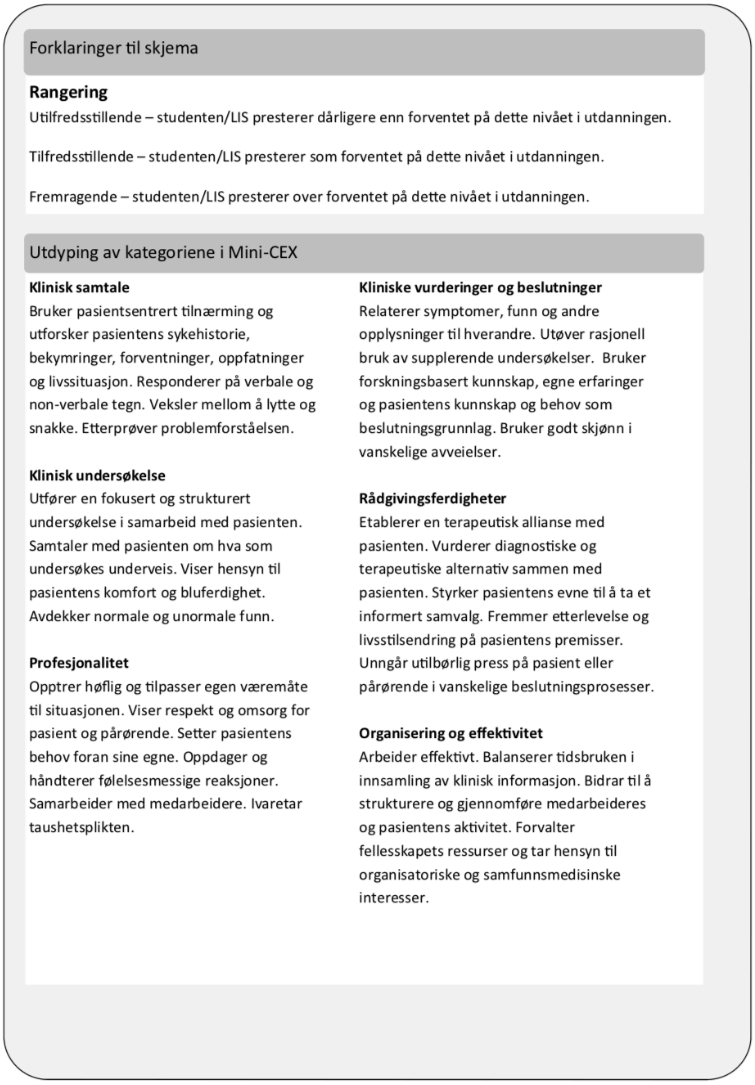
	Utilfredsstillende			Tilfredsstillende			Fremragende			Ikke obs
Klinisk samtale	1	2	3	4	5	6	7	8	9	i.o
Klinisk undersøkelse	1	2	3	4	5	6	7	8	9	i.o
Profesjonalitet	1	2	3	4	5	6	7	8	9	i.o
Kliniske vurderinger og beslutninger	1	2	3	4	5	6	7	8	9	i.o
Rådgivningsferdigheter	1	2	3	4	5	6	7	8	9	i.o
Organisering og effektivitet	1	2	3	4	5	6	7	8	9	i.o
Alt i alt	1	2	3	4	5	6	7	8	9	

Tidsbruk observasjon: ○ min Tidsbruk tilbakemelding: ○ min

Signatur veileder

Signatur student / LIS

Versjon 19.12.2017



Forklaringer til skjema

Rangering

Utilfredsstillende – studenten/LIS presterer dårligere enn forventet på dette nivået i utdanningen.

Tilfredsstillende – studenten/LIS presterer som forventet på dette nivået i utdanningen.

Fremragende – studenten/LIS presterer over forventet på dette nivået i utdanningen.

Utdyping av kategoriene i Mini-CEX

Klinisk samtale
Bruker pasientsentrert tilnærming og utforsker pasientens sykehistorie, bekymringer, forventninger, oppfatninger og livssituasjon. Responderer på verbale og non-verbale tegn. Veksler mellom å lytte og snakke. Etterprøver problemforståelsen.

Klinisk undersøkelse
Utfører en fokusert og strukturert undersøkelse i samarbeid med pasienten. Samtaler med pasienten om hva som undersøkes underveis. Viser hensyn til pasientens komfort og bluferdighet. Avdekker normale og unormale funn.

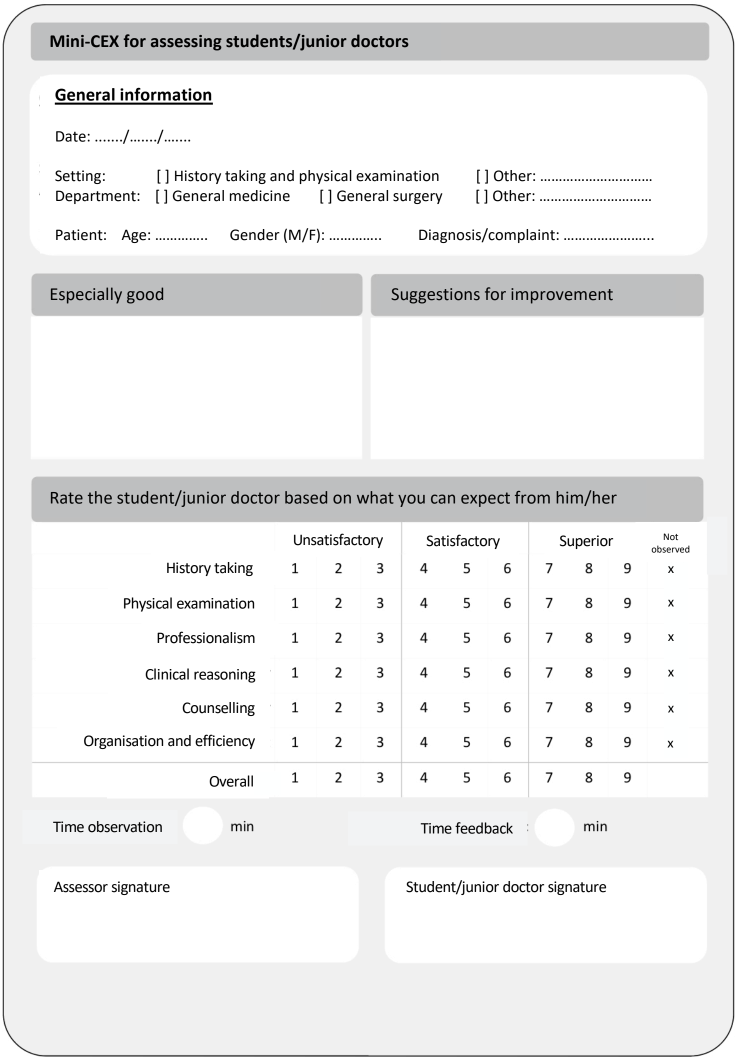
Profesjonalitet
Opptrer høflig og tilpasser egen væremåte til situasjonen. Viser respekt og omsorg for pasient og pårørende. Setter pasientens behov foran sine egne. Oppdager og håndterer følelsesmessige reaksjoner. Samarbeider med medarbeidere. Ivaretar taushetsplikten.

Kliniske vurderinger og beslutninger
Relaterer symptomer, funn og andre opplysninger til hverandre. Utøver rasjonell bruk av supplerende undersøkelser. Bruker forskningsbasert kunnskap, egne erfaringer og pasientens kunnskap og behov som beslutningsgrunnlag. Bruker godt skjønn i vanskelige avveielser.

Rådgivingsferdigheter
Etablerer en terapeutisk allianse med pasienten. Vurderer diagnostiske og terapeutiske alternativ sammen med pasienten. Styrker pasientens evne til å ta et informert samvalg. Fremmer etterlevelse og livsstilsendring på pasientens premisser. Unngår utilbørlig press på pasient eller pårørende i vanskelige beslutningsprosesser.

Organisering og effektivitet
Arbeider effektivt. Balanserer tidsbruken i innsamling av klinisk informasjon. Bidrar til å strukturere og gjennomføre medarbeideres og pasientens aktivitet. Forvalter fellesskapets ressurser og tar hensyn til organisatoriske og samfunnsmedisinske interesser.

The mini-CEX form (English version)



Mini-CEX for assessing students/junior doctors

General information

Date: ……./……./…….

Setting: [ ] History taking and physical examination [ ] Other: …………………………
Department: [ ] General medicine [ ] General surgery [ ] Other: …………………………

Patient: Age: …………. Gender (M/F): …………. Diagnosis/complaint: ……………………..

Especially good

Suggestions for improvement

Rate the student/junior doctor based on what you can expect from him/her

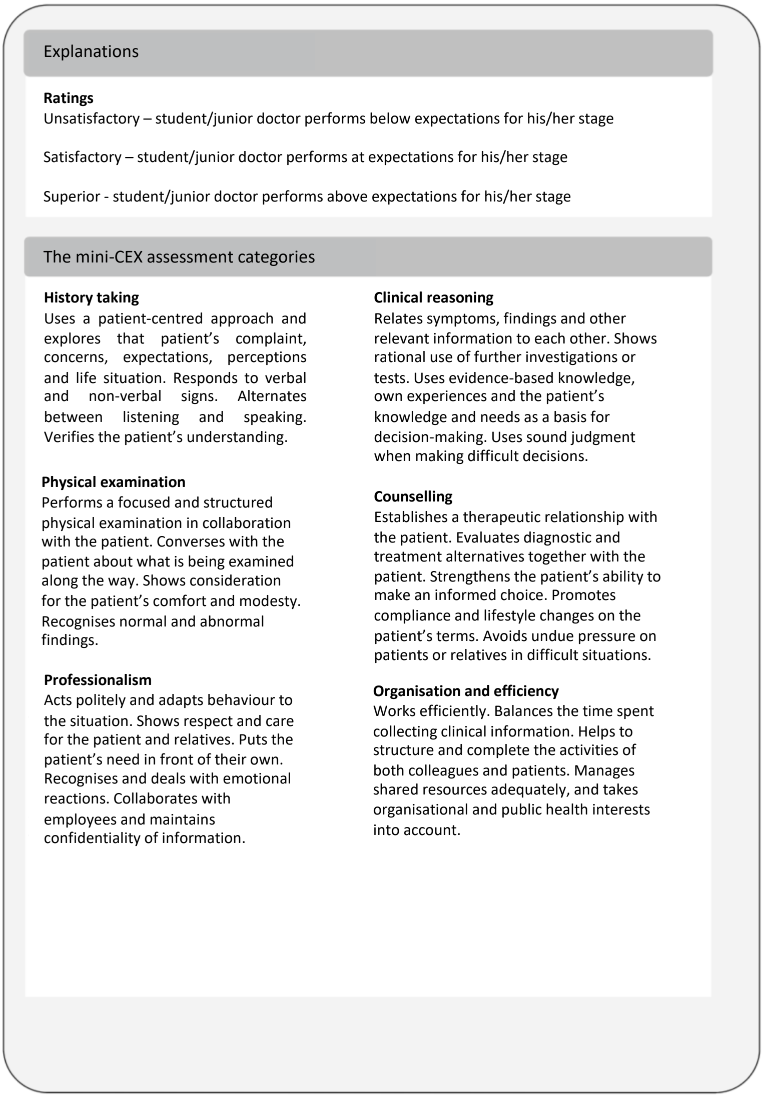
	Unsatisfactory			Satisfactory			Superior			Not observed
History taking	1	2	3	4	5	6	7	8	9	x
Physical examination	1	2	3	4	5	6	7	8	9	x
Professionalism	1	2	3	4	5	6	7	8	9	x
Clinical reasoning	1	2	3	4	5	6	7	8	9	x
Counselling	1	2	3	4	5	6	7	8	9	x
Organisation and efficiency	1	2	3	4	5	6	7	8	9	x
Overall	1	2	3	4	5	6	7	8	9	

Time observation ○ min

Time feedback ○ min

Assessor signature

Student/junior doctor signature



Explanations

Ratings

Unsatisfactory – student/junior doctor performs below expectations for his/her stage

Satisfactory – student/junior doctor performs at expectations for his/her stage

Superior - student/junior doctor performs above expectations for his/her stage

The mini-CEX assessment categories

History taking

Uses a patient-centred approach and explores that patient's complaint, concerns, expectations, perceptions and life situation. Responds to verbal and non-verbal signs. Alternates between listening and speaking. Verifies the patient's understanding.

Physical examination

Performs a focused and structured physical examination in collaboration with the patient. Converses with the patient about what is being examined along the way. Shows consideration for the patient's comfort and modesty. Recognises normal and abnormal findings.

Professionalism

Acts politely and adapts behaviour to the situation. Shows respect and care for the patient and relatives. Puts the patient's need in front of their own. Recognises and deals with emotional reactions. Collaborates with employees and maintains confidentiality of information.

Clinical reasoning

Relates symptoms, findings and other relevant information to each other. Shows rational use of further investigations or tests. Uses evidence-based knowledge, own experiences and the patient's knowledge and needs as a basis for decision-making. Uses sound judgment when making difficult decisions.

Counselling

Establishes a therapeutic relationship with the patient. Evaluates diagnostic and treatment alternatives together with the patient. Strengthens the patient's ability to make an informed choice. Promotes compliance and lifestyle changes on the patient's terms. Avoids undue pressure on patients or relatives in difficult situations.

Organisation and efficiency

Works efficiently. Balances the time spent collecting clinical information. Helps to structure and complete the activities of both colleagues and patients. Manages shared resources adequately, and takes organisational and public health interests into account.

Survey used in Paper III (English version)

Survey on mini-CEX (intervention group only)

Adapted with permission from authors from: Bindal, T., Wall, D., & Goodyear, H. M. (2011). Trainee doctors' views on workplace-based assessments: Are they just a tick box exercise?. Medical teacher, 33(11), 919-927.

1. How positive or negative were your expectations for the use of the mini-CEX during your clinical placement?
   Very negative – Negative – Neutral – Positive – Very positive

2. How confident are you that the mini-CEX is a true reflection of your capabilities?
   Very unconfident – Unconfident – Neutral – Confident – Very confident

3. How easy is it to find assessors to carry out the mini-CEX?
   Very difficult – Difficult – Neutral – Easy – Very easy

4. Who carried out your mini-CEXs? *Please circular for each of the following for frequency.*
   Never – Rarely – Sometimes – Often – Always
   Consultant
   Junior doctor (FY2/SpR)
   Junior doctor (FY1)
   Others *Please state* ………………

5. How are your mini-CEXs planned?
   Pre planned
   Ad hoc/on the job
   Retrospectively, form filled out retrospectively
   Retrospectively, patient I had previously spoken to/examined
   Other *Please state* ………….

6. On average, how soon after the assessment do you get feedback?

Oral feedback	*mini-CEX form completed*
Immediately	Immediately
< 30 minutes	< 30 minutes
< 2 hours	< 2 hours
Same day	Same day
Next day or later	Next day or later

7. Have you had a doctor refuse to carry out a Mini-CEX? Yes/No

If yes what are the reasons? *Please tick all that apply*

Too busy
Not confident
Not had training
Did not understand the form
Other *Please state* ………………

8. How useful are Mini-CEX assessments for helping with your medical training?
Very useless – Slightly useless – Neutral – Useful – Very useful

9. Please write down any comments about the use of Mini-CEX during your clinical placement.

Survey on feedback and learning (both groups)

Perceptions of feedback

1. To what degree have you received feedback and supervision from:
Never – Rarely – Sometimes – Often – Always

Consultant
Junior doctor (FY2/SpR)
Junior doctor (FY1)
Medical student
Others *Please state* ………………

2. To what degree have you received individual and specific feedback on:
   Never – Rarely – Sometimes – Often – Always
   History taking
   Physical examination
   Procedures
   Clinical reasoning
   Presenting findings/cases
   Medical record notes (admission, progress or discharge notes etc.)

3. To what extent do you agree with the following statements about feedback:
   Strongly disagree – Disagree – Neutral – Agree – Strongly agree
   Doctors or other health professionals directly observed my clinical skills working with patients.
   I often received positive feedback on what went well.
   I often received constructive, negative feedback on what could be improved.
   When something could be improved, I often received guidance on how to improve.
   I received feedback when working with a wide range of patients and complaints.
   I am very satisfied with the amount of feedback I have received during my clinical placement.
   I would very much have liked more feedback and supervision during my clinical placement.
   I am very satisfied with the quality of feedback I have received during my clinical placement.
   The feedback and supervision I have received has been very useful.
   Feedback during my clinical placement has led me to learn more.

4. Please write down any comments.

Perceptions of learning

1. To what extent do you agree with the following statements about what you have learned during your clinical placement?
   Strongly disagree – Disagree – Neutral – Agree – Strongly agree
   I have become much better at identifying key information in the history.
   I have become much more efficient in my history taking.
   I have become much better at employing patient centred clinical method in history taking.
   I have become much better at carrying out a structured clinical examination.
   I have become much more efficient in clinical examination.
   I have become much better at identifying normal and abnormal findings on clinical examination.
   I have become much better at carrying out the procedures on the skills list.
   I have become much better at suggesting differential diagnoses based on findings in the history and clinical examination.
   I have become much better at suggesting further investigations based on findings in the history and clinical examination.
   I have become much better at presenting cases.

2. To what extent do you agree with the following statements about what you master?
   Strongly disagree – Disagree – Neutral – Agree – Strongly agree
   I know which topics I master and which I need to spend more time learning.
   I know which clinical examinations I master and which I need to spend more time learning.
   I know which procedures I master and which I need to spend more time learning.

3. To what degree do you agree with the following statements about how confident you feel?
   Strongly disagree – Disagree – Neutral – Agree – Strongly agree
   I feel very confident in performing tasks that can be expected of a medical student who has finished his or her fifth year (provisional license to practice medicine).
   I feel very confident that I have learned enough to work with a provisional license.
   I am never afraid of asking for help from a more experienced colleague.
   I am never afraid of asking for feedback on my clinical skills.

4. Please write down any comments.

Perceptions of motivation and self-directed learning

1. To what degree do you agree with the following statements?
   Strongly disagree – Disagree – Neutral – Agree – Strongly agree
   I am motivated to meet/clerk patients.
   I am motivated to acquire medical knowledge (through reading, digital resources etc.)
   During my clinical placement, I regularly sought medical knowledge by myself.

2. Approximately how much time did you spend acquiring medical knowledge outside your clinical placement?
   Time in minutes for an average week.

3. Please write down any comments.

Paper I

Education

Express Team-Based Learning (eTBL): A Time-Efficient TBL Approach in Neuroradiology

Susanne Skjervold Smeby, MD, Børge Lillebo, MD, PhD, Tobias S. Slørdahl, MD, PhD, Erik Magnus Berntsen, MD, PhD

Abbreviations	
eTBL	express team-based learning
IF-AT	immediate feedback assessment technique
iRAT	individual readiness assurance test
MCQ	multiple choice question
MEQ	modified essay question
PBL	problem-based learning
RAT	readiness assurance tests
TBL	team-based learning
tRAT	team readiness assurance test
4 S's principle	Significant, Same, Specific, Simultaneous

Rationale and Objectives: Team-based learning (TBL) is a student-centred, teacher-directed instructional method that promotes active learning. The application phase of TBL stimulates group discussion and critical thinking, which could be useful for learning radiology. We designed and evaluated two modified TBL-sessions on computed tomography and magnetic resonance imaging diagnostics in neuroradiology. Our aim was to examine what effects engaging students in in-class team application tasks had on student learning.

Materials and Methods: A cross-over study was conducted, including 105 third-year medical students using two modified TBL sessions as the active learning intervention compared with two traditional lectures as a control. Student learning was assessed by results on the neuroradiology part of the end-of-year written examination. Student engagement and perceptions were assessed using the Student Self-Report of Engagement Measure and an additional four Likert-type items.

Results: There were no statistically significant differences in student scores on the examination. Students reported high levels of engagement, and reported being more satisfied overall with the TBL sessions than traditional lectures. Students rated the TBL sessions higher than lectures on ability to make difficult material comprehensible, ability to engage students and to give them feedback.

Conclusion: The modified TBL sessions halved in-class teaching time and by omitting the readiness assurance tests, there was more in-class time to focus on problem-solving of real clinical cases. Moreover, shorter sessions may ease implementation of TBL in the curriculum and allow for more frequent sessions. Students were more satisfied with eTBL than lectures, and reported high levels of engagement.

Key Words: Medical education; team-based learning; active learning; learning effect; student engagement.

INTRODUCTION

Team-based learning (TBL) is a student-centred instructional strategy that promotes active learning whilst maintaining a high student-teacher ratio (1). TBL was originally developed for business education, but is increasingly being used in both undergraduate and graduate medical education (2). It is well-suited to the rapidly growing field of medicine which demands that we educate life-long learners, and prepare students for the interprofessional and team-oriented field of practice (1). We believe it is especially well-suited for visual topics such as radiology, as it engages and facilitates group discussion of real-life complex radiological cases.

Acad Radiol 2020; 27:284–290

From the Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology, PO Box 8905, 7491 Trondheim, Norway (S.S.S., T.S.S.); Department of Circulation and Medical Imaging, Norwegian University of Science and Technology, Trondheim, Norway (B.L., E.M.B.); Clinic of Medicine and Rehabilitation, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway (B.L.); Department of Haematology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway (T.S.S.); Department of Radiology and Nuclear Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway (E.M.B.). Received March 26, 2019; revised April 25, 2019; accepted April 27, 2019. **Address correspondence to**: S.S. S. e-mail: susanne.s.smeby@ntnu.no

https://doi.org/10.1016/j.acra.2019.04.022

The original application of TBL consists of three phases (3). During the first phase, students do preparatory reading or other advance assignments before the TBL session. In the second phase, students complete an individual readiness assurance test (iRAT) that tests basic facts and concepts of the advance assignment, before retaking the same test in teams of 5–7 students (team readiness assurance test, tRAT). This test is answered using immediate feedback assessment technique (IF-AT), usually in the form of a scratch card, motivating the students to collaborate until all answers are correct (3). During the third phase (team application), the teams apply their knowledge to solve clinical problems that they are likely to meet in their professional careers. In line with TBL principles for effective problem design (4 S's principles), the problems should be significant for the students, the same for all teams, and the teams must make a specific choice and simultaneously report their answers (1,3). This ensures that students get immediate feedback and are accountable to explain and defend their answers (1).

There is a growing body of evidence suggesting that academic outcomes are as good or better with TBL compared to traditional teaching strategies (4,5). In a systematic review of 14 studies in health professions education, seven studies showed improved knowledge scores in the TBL group compared with a non-TBL group (4). No studies reported a decrease in scores for the TBL group. Learner attitudes toward TBL are generally positive, emphasising the active learning style and interaction with their peers (5).

By emphasising or skipping one or more of the phases of TBL, the method allows for flexibility in design (6). Although this variability can be a challenge in medical education research, it enables teachers to tailor TBL to course context and learner needs (7). After piloting TBL in its original format in our medical programme, evaluations showed that although students were positive, TBL was perceived as time-consuming with one session taking up a total of three 45-minute blocks (8). Many courses have tested modified versions of TBL and common to most of these is maintaining the RAT (9,10).

In this study, we describe and test a modified and time-efficient TBL method we have called Express TBL (eTBL). By omitting the full RAT, content learning was moved to out-of-class preparation, leaving in-class time to focus on problem solving of real-life complex cases. The aim of the study was to answer the following questions: Compared to traditional lectures, what effect does eTBL have on student learning assessed with a summative examination? How do students perceive this approach as compared to lectures? Does eTBL actively engage students?

MATERIALS AND METHODS

Study Setting

The six-year undergraduate medical programme at the Norwegian University of Science and Technology (NTNU) is integrated and problem-based, with one oral and one written summative examination at the end of each year. The third year covers 16 clinical specialties (including radiology) and four paraclinical sciences. Lectures and problem based learning (PBL) sessions are organized around weekly themes. In addition, students attend clinical rotations at the university hospital. Lectures are predominantly based on traditional didactic teaching, but students have previously had one TBL session in general pathology during their second year and several lecturers have started converting some of their lectures to TBL (8).

Study Design

The study was conducted during the academic year of 2016/2017 using a 2 × 2 cross-over study design (Fig 1). Neuroradiology, which had previously been taught in two 90-minute lectures in computed tomography (CT) and magnetic resonance imaging (MRI) diagnostics, was chosen for the intervention. Third-year medical students were divided into two

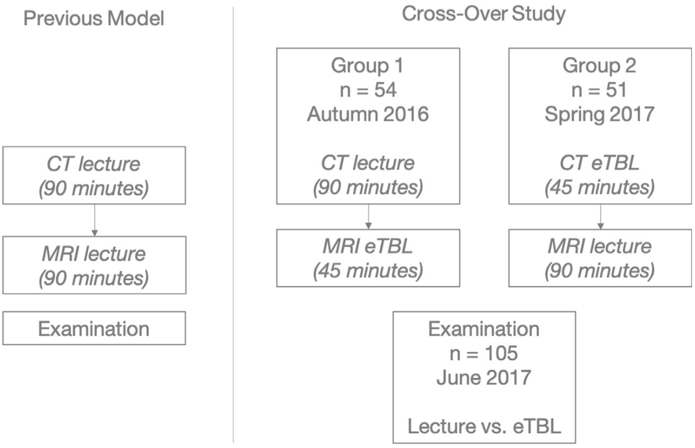


Figure 1. Study design. Lecture: Traditional didactic lecture. eTBL, express team-based learning.

groups at the start of the academic year. Group 1 ($n = 54$) undertook teaching in neuroradiology in August 2016. This group received teaching in CT diagnostics by a 90-minute didactic lecture and MRI diagnostics by a 45-minute eTBL session. Group 2 ($n = 51$) undertook teaching in neuroradiology in January 2017. This group received teaching in MRI diagnostics using a 90-minute didactic lecture and CT diagnostics by a 45-minute eTBL session. This cross-over design ensured that both groups experienced one traditional lecture and one eTBL session. Both groups ($n = 105$) sat for the same summative written examination in June 2017.

Intervention: eTBL

The different phases of eTBL are shown in Figure 2. One week prior to the eTBL sessions, students were sent preparatory reading consisting of a presentation on MRI physics, and a handout on CT and MRI sequences and common findings related to tumors, cerebrovascular and inflammatory diseases of the brain. Both groups received the same material. For eTBL sessions, students sat in teams according to their already established PBL groups, consisting of 6−8 students. Students are randomly assigned to PBL groups (corrected only for gender distribution) and the groups stay constant for each term.

The two-step method of readiness assurance (Phase 2) was reduced to a quick warm-up exercise of 10 multiple-choice questions (MCQs) that students answered individually using an online student response system (Kahoot! AS, Oslo) (11). Individual responses were selected over team responses for time-saving reasons. The majority of the time was spent on application exercises (Phase 3). The exercises were based on real clinical scenarios and included relevant information from the history and clinical examination, as well as CT or MRI images that students had to interpret. One eTBL session typically covered three clinical cases which were formatted as MCQs. In keeping with the 4 S's principles, all groups worked on the same problem and revealed their answers simultaneously. Groups were randomly picked to explain and defend their answers, and each clinical case ended with a summary by the teacher.

Student Performance

At the end of the academic year, students sat the same six-hour written examination, consisting of 100 MCQs and four modified essay questions (MEQs). All items are reviewed and approved by a multidisciplinary examination committee prior to use. The examination covered all subjects taught during the third year. Neuroradiology was tested in one MEQ which consisted of seven sequential questions for a possible total score of 10 points, accounting for 10% of the total score on the examination. The questions were written by the same teacher who held all lectures and eTBL sessions in neuroradiology. The questions were divided into: (a) content covered in MRI teaching (questions 2, 5, and 6) and (b) content covered in CT teaching (questions 3, 4, and 7). This allowed us to compare the two groups of students on how they scored on the two parts of the neuroradiology question. The MEQ was marked against a rubric by the item writer who was blinded for what group the students belonged to.

Student Evaluations

Student performance was our primary outcome measure, but after positive student feedback following eTBL in 2016/2017 we were interested in collecting student evaluations and measures of engagement. We prepared an anonymous 17-item survey for third-year students who had just completed an eTBL session in both CT and MRI diagnostics in the spring of 2018. The survey included the nine-item Student Self-Report of Engagement Measure (Table 1) which measures engagement using a five-point Likert scale (1 = strongly disagree, 5 = strongly agree) (12). Students were also asked to rate eTBL and traditional lectures on a five point Likert scale with regards to overall satisfaction (1 = very dissatisfied, 5 = very satisfied), ability to make difficult material comprehensible, ability to engage, and perception of receiving feedback on their own knowledge (1 = to a very small extent, 5 = to a very large extent).

Statistical Analysis

Statistical analysis was performed using IBM SPSS Statistics 24. Mann-Whitney *U* tests were used to compare student scores on the neuroradiology MEQ for CT and MRI questions separately. Wilcoxon signed-rank tests were used to compare student evaluations of eTBL and traditional lectures. Nonparametric tests were used because scores were not normally distributed. Two-tailed significance was set at $p < 0.05$.

Ethical Considerations

The TBL sessions and lectures were organized as noncompulsory learning activities and student evaluations were anonymous. Examination results were extracted and analyzed anonymously. In accordance with the Norwegian Center for Research Data (NSD) guidelines, approval for this study was deemed unnecessary because only anonymous data was processed.

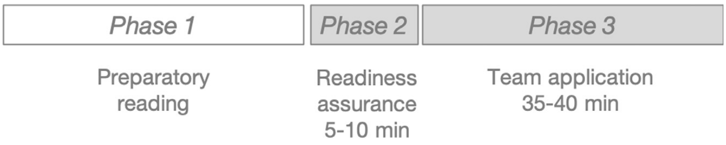


Figure 2. Phases of eTBL. The white field represents out-of-class preparations and gray fields represent in-class time.

TABLE 1. Student Self-Report of Engagement Measure in eTBL group. Response Categories for Items Ranged From 1 (Strongly Disagree), 2 (Disagree), 3 (Neither Agree Nor Disagree), 4 (Agree), to 5 (Strongly Agree)

Item	Mean Scores (SD)
1. I contributed meaningfully to class discussions today.	4.35 (0.80)
2. I was not paying attention most of the time in class.[a]	4.95 (0.22)
3. I contributed my fair share to class discussions.	4.48 (0.75)
4. I participated in class discussions today.	4.60 (0.63)
5. I talked in class with other students about class material.	4.73 (0.51)
6. I was mostly a passive learner in class today.[a]	4.40 (0.87)
7. I paid attention most of the time in class.	4.73 (0.72)
8. I was mostly an active learner in class today.	4.30 (0.99)
9. Most students were actively involved in class today.	4.15 (0.95)
Mean total score	4.52 (0.49)

[a] Denotes items that were reverse scored. Total score was calculated by reverse scoring items 2 and 6, and averaging the nine items.

RESULTS

Student Performance

Figure 3 shows a box plot of the median, quartile, and range of scores on CT and MRI questions in the neuroradiology MEQ in the end-of-year examination based on teaching method. Mann-Whitney *U* tests were conducted to compare student performance. Median scores on MRI questions (questions 2, 5, and 6, maximum score 4.0) in the lecture and eTBL group were 2.0 and 2.5, respectively, and did not differ significantly ($U = 1255$, $p = 0.415$). Median scores on CT questions (questions 3, 4 and 7, maximum score 4.0) in the lecture and eTBL group were 3.5 and 4.0, respectively, and did not differ significantly ($U = 1191$, $p = 0.182$).

Student Evaluations

Of 41 students who participated in the eTBL session in 2018, 40 completed the student evaluation. The Student Self-Report of Engagement Measure (Table 1) showed that students reported high levels of engagement, with a mean total score of 4.52 (12).

Figure 4 shows median scores on Likert-type items comparing student evaluations of eTBL and lectures. For better legibility, a bar chart was constructed instead of a box plot. A Wilcoxon signed-rank test indicated that students were more satisfied overall with eTBL (*Mdn* = 5) than traditional lectures (*Mdn* = 3, $Z = 4.96$, $p < 0.001$). Students rated eTBL higher than lectures on ability to make difficult material comprehensible (*Mdn* = 4 vs. 3, $Z = 4.57$, $p < 0.001$) and on its ability to engage students (*Mdn* = 5.0 vs. 3.0, $Z = 5.10$, $p < 0.001$). Additionally, students perceived eTBL superior to traditional lectures on ability to give them feedback on their own knowledge (*Mdn* = 4.5 vs. 2, $Z = 5.17$, $p < 0.001$).

DISCUSSION

With more institutions adopting TBL, it is necessary to understand how greater efficiencies can be gained from the method. In this study, we used a cross-over design to explore the educational effects of a modified and time-efficient TBL method in neuroradiology. Traditional lectures were chosen as the control because passive teaching methods still constitute the majority of teaching in undergraduate medical education (13). Results showed that there were no statistically significant differences in student performance on the end-of-year examinations based on teaching method. This is consistent with other studies in medical and health professions education where examination or test results remain the same after implementation of TBL, or when compared to traditional lectures (5,14−17). However, a meta-analysis of findings from 17 studies across a variety of different fields at undergraduate and graduate level, found a moderate positive effect of TBL on content knowledge (18).

The lack of impact of eTBL compared to traditional lectures in this study may be explained by several factors. First, students may have learned content at other points in the curriculum, for example in clinical teaching or through the preparatory reading material that was given to all students. Second, the intervention was small with only one eTBL session per group. Third, end-of-year summative examinations may be a poor measure of effectiveness. Medical students spend an extensive amount of time preparing for examinations, probably compensating for the use of ineffective teaching methods (19). Additionally, written examinations fail to assess other aspects that TBL aims to promote, such as teamwork and communication skills (1). Finally, a number of studies have indicated that the impact of TBL seems to be largest for academically weaker students (6,20−22). In the studies of Kang et al, and Koles et al, the lowest quartile was the only group who showed a significant improvement in test scores (20,21).

The purpose of the RAT is to link advance preparations to the application exercises, and when done well, is said to give effective content coverage, better teamworking skills, and insight about the value of diverse input (23). After a search of the literature, there seem to be few studies that examine the claims made about the RAT. A study by Rotgans et al examined how cognitive engagement fluctuates during a TBL session, and found that students are significantly more engaged when working together during the tRAT and application exercises (24). Although this does not validate all claims made about the readiness assurance procedure, it indicates that the tRAT is able to foster student engagement. Two studies have examined the RAT's effect on students' knowledge of the material. Carbrey et al found that performance on a knowledge test after traditional in-class RATs was equal to having learners complete the iRAT at home

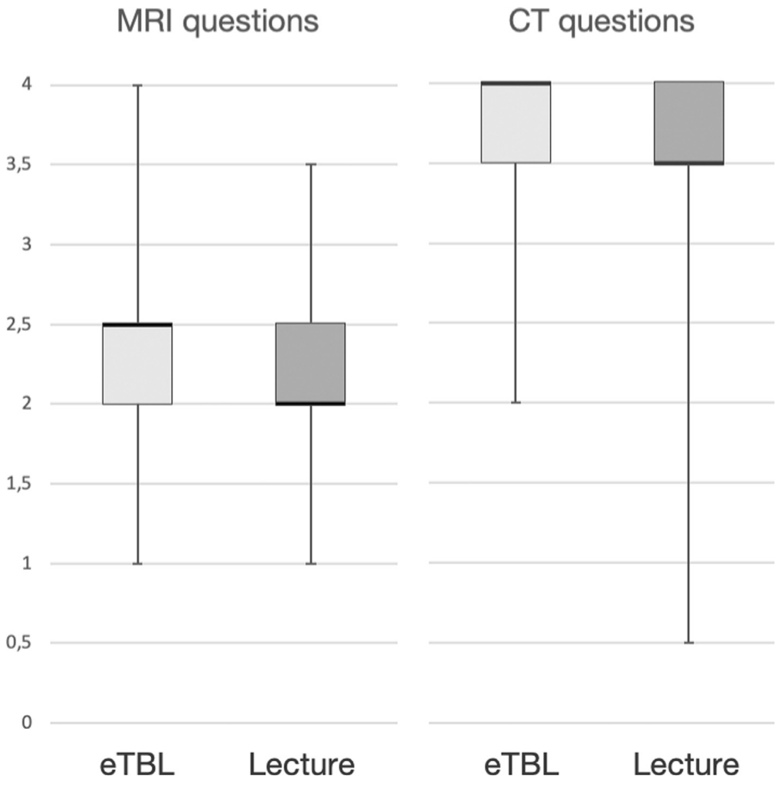


Figure 3. Box plot of scores on neuroradiology MEQ. The graphs display the median, quartiles and range of scores by teaching method on the two parts of the MEQ: MRI questions (maximum score 4.0) and CT questions (maximum score 4.0). CT, computed tomography; MEQ, modified essay question; MRI, magnetic resonance imaging.

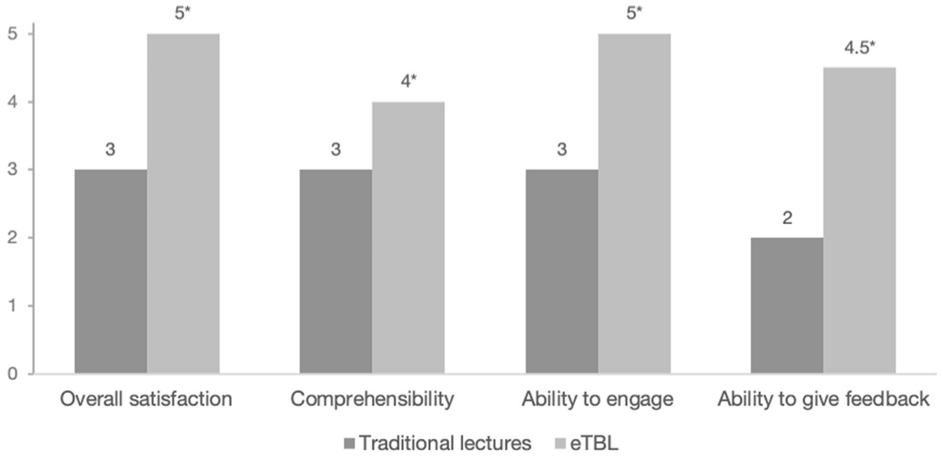


Figure 4. Student evaluations in eTBL group. The graph displays median scores on Likert-type items relating to overall satisfaction with traditional lectures and eTBL, their ability to make difficult material comprehensible, engage students and perception of feedback. *Note:* * indicates statistically significant differences at $p < 0.001$.

without a tRAT (25). Another study by Gopalan et al found that although the iRAT helps teams earn higher tRAT scores, it does not affect students' examination scores (26).

The eTBL method skips full administration of the RAT, allowing for a more time-efficient administration of TBL. Our study cannot determine whether the RAT has additional effects on teamwork skills or knowledge that students do not gain through the application exercises. However, curriculum overload has long been recognized as a challenge in medical education, and lecture hours cannot be expanded in parallel with the rapid growth of biomedical knowledge (27). By delivering content in the eTBL format, classroom hours were reduced from 90 to 45 minutes, and by reducing the RAT to a short warm-up exercise, there was more in-class time to focus on problem solving. Shorter sessions may also ease implementation of TBL in the curriculum and allow for more frequent sessions. However, studies have found that when students are taught using overly contextualized knowledge, they may have issues with transferring that knowledge to other situations (28). Therefore, we must be careful that problem-solving is an application of what the student has learned, and not the only way that the information is presented to them.

The secondary goal of this study was to document student engagement, and student opinions of eTBL compared with traditional lectures. Several studies link student engagement to positive learning outcomes such as critical thinking and grades (29). Student engagement was measured by a nine-item self-report instrument which has previously shown good internal consistency (Cronbach's alpha of 0.84) (30). Validity evidence is further provided by a similar pattern of results between the self-report instrument and levels of observed engagement (12). In our study, students reported high levels of engagement during eTBL, with a mean total score of 4.52. This is in line with other studies reporting high levels of engagement both for full and modified implementations of TBL (6,24,31−34). Using the same self-report instrument, Sharma et al found that ratings of engagement were higher during TBL than during traditional lectures for five of the measures (31). Although not surprising, the ability of TBL and eTBL to foster active learning in a large-group setting makes it attractive compared to other forms of active teaching strategies that have lower student to staff ratios.

In this study, students reported that they were significantly more satisfied with eTBL than traditional lectures. This is in line with the literature previously discussed, with student attitudes toward TBL being generally positive (5). Further research is needed to see if this persists over time, as several studies have shown that learner satisfaction and perception of the usefulness of TBL decreases with time (16,35,36). Students rated eTBL significantly higher than lectures on ability to make difficult material comprehensible, on its ability to engage students and its ability to give them feedback on their own knowledge. Interestingly, the greatest difference between eTBL and traditional lectures was students' perception of receiving feedback on their own knowledge. In eTBL students receive feedback from performance on application exercises, peers, and staff. This finding is in contrast with the hypotheses that students are unable to recognize feedback and therefore give poor feedback ratings (37). This deserves further study, to confirm the finding and to clarify which aspects of eTBL students perceive as feedback and whether this supports self-directed learning.

CONCLUSION

Introducing eTBL in a neuroradiology course halved in-class teaching time, and by reducing the RAT to a short warm-up exercise, there was more in-class time to focus on problem-solving. Shorter sessions may ease implementation of TBL in the curriculum and allow for more frequent sessions. This study showed no difference in student performance when comparing eTBL to lectures, but students reported high levels of engagement, and they were more satisfied with eTBL compared to lectures. Additionally, students rated eTBL higher on its ability to make difficult material comprehensible and its ability to give them feedback on their own knowledge.

REFERENCES

1. Parmelee D, Michaelsen LK, Cook S, et al. Team-based learning: a practical guide: AMEE guide no. 65. Med Teach 2012; 34:e275–e287.
2. Michaelsen LK. Team learning: A comprehensive approach for harnessing the power of small groups in higher education, 11. Lincoln: University of Nebraska 1992. p. 107–122.
3. Hrynchak P, Batty H. The educational theory basis of team-based learning. Med Teach 2012; 34:796–801.
4. Fatmi M, Hartling L, Hillier T, et al. The effectiveness of team-based learning on learning outcomes in health professions education: BEME Guide No. 30. Med Teach 2013; 35:e1608–e1624.
5. Reimschisel T, Herring AL, Huang J, et al. A systematic review of the published literature on team-based learning in health professions education. Med Teach 2017; 39:1227–1237.
6. Tan NC, Kandiah N, Chan YH, et al. A controlled study of team-based learning for undergraduate clinical neurology education. BMC Med Educ 2011; 11:91.
7. Haidet P, Levine RE, Parmelee DX, et al. Perspective: guidelines for reporting team-based learning activities in the medical and health sciences education literature. Acad Med 2012; 87:292–299.
8. Lillebo B, Slørdahl TS, Nordrum IS. Team-based learning − a student-activating and teacher-directed learning activity. Uniped 2017; 40(03): 207–221.
9. Burgess AW, McGregor DM, Mellis CM. Applying established guidelines to team-based learning programs in medical schools: a systematic review. Acad Med 2014; 89:678.
10. Haidet P, Kubitz K, McCormack WT. Analysis of the team-based learning literature: TBL comes of age. J Excell Coll Teach 2014; 25:303.
11. Kahoot! AS. Kahoot! Oslo Available at: https://kahoot.com.
12. O'Malley KJ, Moran BJ, Haidet P, et al. Validation of an observation instrument for measuring student engagement in health professions settings. Eval Health Prof 2003; 26:86–103.
13. Association of American Medical Colleges. Use of instructional methods by US and Canadian medical schools, AAMC curriculum inventory 2014-2015; Available at: https://www.aamc.org/initiatives/cir/456458/ci07.html.
14. Nieder GL, Parmelee DX, Stolfi A, et al. Team−based learning in a medical gross anatomy and embryology course. Clin Anat 2005; 18:56–63.
15. Yang L-H, Jiang L-Y, Xu B, et al. Evaluating team-based, lecture-based, and hybrid learning methods for neurology clerkship in China: a method-comparison study. BMC Med Educ 2014; 14:98.
16. Moore−Davis TL, Schorn MN, Collins MR, et al. Team−based learning for midwifery education. J Midwifery Womens Health 2015; 60:291–297.

17. Bleske B, Remington T, Wells T, et al. A randomized crossover comparison between team-based learning and lecture format on long-term learning outcomes. 2018;6:81.
18. Swanson E, McCulley LV, Osman DJ, et al. The effect of team-based learning on content knowledge: A meta-analysis. Act Learn High Educ 2019; 20:39–50.
19. Ten Cate O. What happens to the student? The neglected variable in educational outcome research. Adv Health Sci Educ 2001; 6:81–88.
20. Koles P, Nelson S, Stolfi A, et al. Active learning in a year 2 pathology curriculum. Med Educ 2005; 39:1045–1055.
21. Kang KA, Kim SJ, Oh J, et al. Effectiveness of simulation with team−based learning in newborn nursing care. Nurs Health Sci 2016; 18:262–269.
22. Chung E-K, Rhee J-A, Baik Y-H. The effect of team-based learning in medical ethics education. Med Teach 2009; 31:1013–1017.
23. Parmelee DX, LK Michaelsen. Twelve tips for doing effective team-based learning (TBL). Med Teach 2010; 32:118–122.
24. Rotgans JI, Schmidt HG, Rajalingam P, et al. How cognitive engagement fluctuates during a team-based learning session and how it predicts academic achievement. Adv Health Sci Educ 2017; 23:1–13.
25. Carbrey JM, Grochowski COC, Cawley J, et al. A comparison of the effectiveness of the team-based learning readiness assessments completed at home to those completed in class. J Educ Eval Health Prof 2015; 12:34.
26. Gopalan C, Fox DJ, Gaebelein CJ. Effect of an individual readiness assurance test on a team readiness assurance test in the team-based learning of physiology. Adv Physiol Educ 2013; 37:61–64.
27. Parsell GJ, Bligh J. The changing context of undergraduate medical education. Postgrad Med J 1995; 71:397–403.
28. Bransford JD, Brown AL, Cocking RR. How people learn: brain, mind, experience, and school: Expanded edition. Washington, D. C.: National Academies Press 2000.
29. Carini RM, Kuh GD, Klein SP. Student engagement and student learning: Testing the linkages. Res High Educ 2006; 47:1–32.
30. Haidet P, O'malley KJ, Richards B. An initial experience with “team learning” in medical education. Acad Med 2002; 77:40–44.
31. Sharma A, Janke KK, Larson A, et al. Understanding the early effects of team-based learning on student accountability and engagement using a three session TBL pilot. Curr Pharm Teach Learn 2017; 9:802–807.
32. Balwan S, Fornari A, DiMarzio P, et al. Use of team-based learning pedagogy for internal medicine ambulatory resident teaching. J Grad Med Educ 2015; 7:643–648.
33. Kelly PA, Haidet P, Schneider V, et al. A comparison of in-class learner engagement across lecture, problem-based learning, and team learning using the STROBE classroom observation tool. Teach Learn Med 2005; 17:112–118.
34. Haidet P, Morgan RO, O'malley K, et al. A controlled trial of active versus passive learning strategies in a large group setting. Adv Health Sci Educ 2004; 9:15–27.
35. Zgheib NK, Simaan JA, Sabra R. Using team-based learning to teach pharmacology to second year medical students improves student performance. Med Teach 2010; 32:130–135.
36. Fujikura T, Takeshita T, Homma H, et al. Team-based learning using an audience response system: a possible new strategy for interactive medical education. J Nippon Med Sch 2013; 80:63–69.
37. Boud D, Molloy E. Introduction. Feedback in higher and professional education: understanding it and doing it well. Routledge 2013. p. 13.

SUPPLEMENTARY MATERIALS

Supplementary material associated with this article can be found in the online version at doi: 10.1016/j.acra.2019.04.022.

Paper II

Martinsen *et al. BMC Medical Education* (2021) 21:228
https://doi.org/10.1186/s12909-021-02670-3

BMC Medical Education

RESEARCH Open Access

Examining the educational impact of the mini-CEX: a randomised controlled study

Susanne Skjervold Smeby Martinsen[1*], Torvald Espeland[2,3], Erik Andreas Rye Berg[2,4], Eivind Samstad[1,5], Børge Lillebo[2,6] and Tobias S. Slørdahl[1,7]

Abstract

Background: The purpose of this study is to evaluate the mini-Clinical Evaluation Exercise (mini-CEX) as a formative assessment tool among undergraduate medical students, in terms of student perceptions, effects on direct observation and feedback, and educational impact.

Methods: Cluster randomised study of 38 fifth-year medical students during a 16-week clinical placement. Hospitals were randomised to provide a minimum of 8 mini-CEXs per student (intervention arm) or continue with ad-hoc feedback (control arm). After finishing their clinical placement, students completed an Objective Structured Clinical Examination (OSCE), a written test and a survey.

Results: All participants in the intervention group completed the pre-planned number of assessments, and 60% found them to be useful during their clinical placement. Overall, there were no statistically significant differences between groups in reported quantity or quality of direct observation and feedback. Observed mean scores were marginally higher on the OSCE and written test in the intervention group, but not statistically significant.

Conclusions: There is considerable potential in assessing medical students during clinical placements and routine practice, but the educational impact of formative assessments remains mostly unknown. This study contributes with a robust study design, and may serve as a basis for future research.

Keywords: Medical education research, Formative assessment, Feedback, Workplace-based assessment

Background

Along with the adoption of competency-based education programmes, there has been increasing emphasis on workplace-based assessments (WBAs) in medical education [1, 2]. WBAs are assessments that assess clinical competence and professional behaviour in everyday practice. As WBAs require direct observation of trainees in the workplace, they also provide opportunities for feedback, and are therefore increasingly being used as methods of formative assessment [3].

The mini-Clinical Evaluation Exercise (mini-CEX) is one of the most commonly used WBAs, and since its introduction in 1995 has been implemented in both undergraduate and postgraduate programmes worldwide [1, 4–7]. Trainees are observed and evaluated while performing a history or physical examination, followed by structured feedback [3, 8]. The mini-CEX can be used with a wide range of clinical problems and workplace settings, allowing trainees to receive feedback from different supervisors [3]. The mini-CEX evaluates multiple competencies that are important in high-quality care [3].

The mini-CEX remains among the most studied WBAs with regards to reliability and validity as an assessment tool [1]. Research has shown that acceptable reliability can be achieved with eight to ten encounters,

* Correspondence: susanne.s.smeby@ntnu.no
[1]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
Full list of author information is available at the end of the article

but the exact number will naturally vary with the stakes and purpose of the assessment [9]. The close correspondence between assessment and practice setting limits validity threats such as construct-irrelevant variance and construct underrepresentation [9]. There are also consistent findings of positive correlations with other assessment outcomes, including high-stakes national specialty examinations [7, 9–12]. Additionally, a number of studies report higher scores with each year of postgraduate training or improvement in scores throughout the academic year [4, 8, 9, 13, 14]. However, concerns have been raised against the scoring component of the mini-CEX [9]. These are primarily rater leniency, high intercorrelations on the individual competencies, and limited research into the effects of rater training.

Evidence is limited for its consequential validity as a formative assessment tool. As the mini-CEX and other WBAs are increasingly being used for providing feedback to trainees in order to support learning and development, research into the impact on educational outcomes would constitute an important source of validity [15]. A systematic review of the educational impact of the mini-CEX found that the majority of articles presented effects on learner perceptions [15]. Only two articles reported on acquisition of knowledge and skills, and demonstrated positive effects on trainee performance in summative clinical examinations [16, 17]. However, as these studies were sequential cohort studies, drawing conclusions concerning causality is difficult.

The aim of this study was to compare mini-CEX assessments with traditional ad-hoc feedback in order to examine its educational impact, effects on direct observation and feedback, as well as student perceptions of the mini-CEX as a formative assessment tool.

Methods

Study design

We conducted a cluster randomised controlled trial with two groups and blinded outcome assessment. A cluster trial design was chosen to avoid contamination (doctors who received extra training in assessment and feedback using the mini-CEX could not be expected to treat individual students differently), as well as for practical purposes.

Study setting

The six-year undergraduate medical programme at the Norwegian University of Science and Technology (NTNU) is integrated and problem-based. Students cover most clinical subjects in Years 3 and 4. The following year, they complete a 16-week clinical placement at one of the general hospitals in the region, during which this study took place in 2018. This undergraduate setting was chosen as it allows for better standardisation of what is learned during these weeks, and made organising post-study assessments easier.

The clinical placement includes general medicine (7 weeks), general and orthopaedic surgery (7 weeks) and anaesthesia (2 weeks), and all students are required to complete the same checklist of activities and procedures. Prior to this study, feedback had not been formalised in WBAs and was given on an ad-hoc basis. That is, immediate feedback given by doctors or other health professionals while working with students, or prompted by students asking for feedback or help.

Participants and randomisation

Six of the nine general hospitals in the region were enrolled in the study (Fig. 1). The six hospitals were allocated in a 1:1 ratio to give feedback using mini-CEX assessments (intervention arm) or continue with ad-hoc feedback (control arm), using a simple randomisation procedure by means of drawing lots. Student participation was voluntary and there were no exclusion criteria. All participants provided written consent. The study was approved by the Norwegian Centre for Research Data (project number: 56646).

Forty-eight students were invited by email, and of these, 41 students consented to participate. Three students later withdrew from the trial because they were unable to attend outcome assessments, leaving 19 students in the intervention group and 19 students in the control group that were included in the analyses.

Intervention

Participants in the intervention group were asked to complete a minimum of eight formative mini-CEX assessments. They were encouraged to choose patients with a wide range of clinical problems and assessors with different levels of training (newly qualified doctors to consultants). Apart from mini-CEX assessments, no other changes were made to their clinical placement. The amount of time spent in clinical practice, and requirements with regards to checklist activities and procedures remained the same between the groups.

The assessment part of the mini-CEX consists of six competencies and one overall score [13]. Each competency is scored on a nine-point rating scale. The feedback part consists of one box for ‘Especially Good’ and one for ‘Suggestions for Improvement’.

All participants and assessors were naïve to the mini-CEX. Thus, a 45-min session was held for doctors in each intervention hospital. It emphasised the importance of direct observation and effective feedback. Using a video recording, doctors completed a mini-CEX assessment, followed by a plenary discussion. A written guide was also provided.

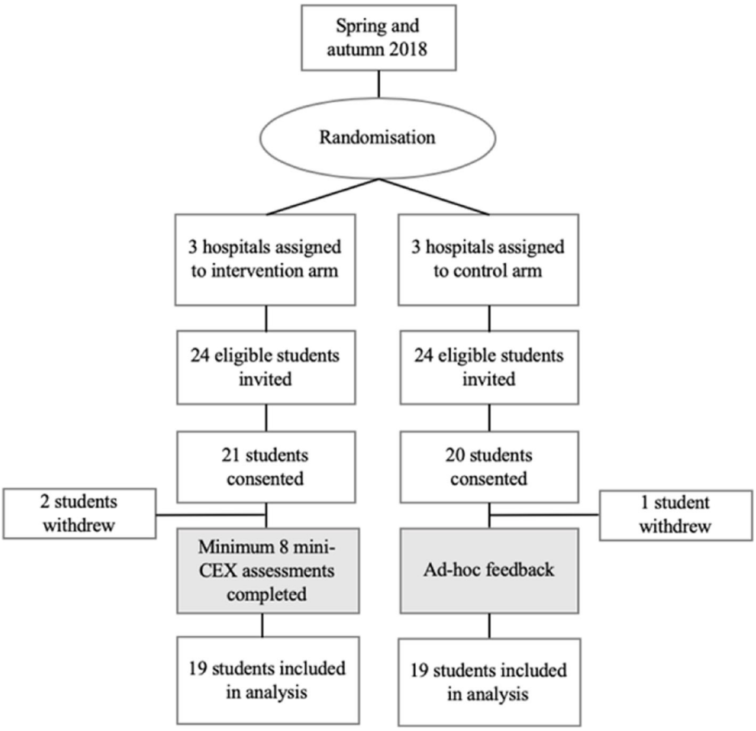

Spring and autumn 2018
Randomisation
3 hospitals assigned to intervention arm
3 hospitals assigned to control arm
24 eligible students invited
24 eligible students invited
21 students consented
20 students consented
2 students withdrew
1 student withdrew
Minimum 8 mini-CEX assessments completed
Ad-hoc feedback
19 students included in analysis
19 students included in analysis

Fig. 1 Flow chart of randomised controlled study. Mini-CEX: mini-Clinical Evaluation Exercise

Students in both groups were given a presentation of the study aims and outcome assessments, in addition to written material included in the invitation email. Students in the intervention group were also given the same introduction to the mini-CEX as was held for the doctors in the intervention hospitals.

Outcome measures

At the end of the clinical placement, all participants completed a survey, a written test and an Objective Structured Clinical Examination (OSCE). These assessment methods were chosen because they are familiar to students from the university's assessment programme, but were held separately and did not have any consequences for the students' progression.

The OSCE consisted of six eight-minute stations (Table 3). Station topics were chosen based on common patient presentations to emergency departments (i.e., chest pain, dyspnoea, fever, abdominal pain, limb injury and neurological symptoms). All stations were drafted by the first author, and reviewed and edited by content experts. Standardised patients were trained in each specific clinical scenario, and remained the same throughout the study. The stations were filmed and later checklist-scored by two independent examiners, blinded to the intervention.

The written test consisted of 43 single best answer multiple choice questions (MCQs). Most items were selected from previously used examination items, with item difficulty of 0.20–0.80 and item discrimination index above 0.20. Tests were corrected without negative marking or corrections-for-guessing [18].

The first part of the survey was answered by both groups, and consisted of 40 Likert-type questions and 4 free text questions divided into three sections: (a) perceptions of feedback, (b) perceptions of learning and confidence, and (c) perceptions of motivation. A review of the literature on feedback, especially the work of Hattie and Timperley, informed the design [19]. Items were constructed adhering to best practices for item-writing and item-design [20]. To ensure that questions were unambiguous and meaningful, cognitive interviews utilising the probing method were held with students who had recently completed their clinical placement [21].

The second part of the survey was answered only by the intervention group and comprised of 13 items on perceptions of the mini-CEX, adapted with permission from Bindal and colleagues [22]. There were eight Likert-type questions, four tick box questions and one free text question.

Statistical analyses

Analyses of student learning and perceptions were based on individual student-level data, rather than on the cluster-level summarised data. Students select their hospital on the basis of a randomly assigned number which ensures some degree of randomisation. Data from previous examinations indicated that a total of 17 students in each arm for the OSCE and 29 students in each arm for the written test, were needed to achieve an 80% power to detect a 5% difference in test scores with a 2-sided 0.05 significance level.

One-way analysis of variance (ANOVA) was used to compare intervention and control group mean scores on the OSCE and written test. Since the trial was cluster randomised, a second analysis was performed using a one-way analysis of covariance (ANCOVA), controlling for previous examination scores to account for baseline differences in students' clinical skills and knowledge. For the OSCE, mean Z-scores of the three previous summative OSCEs in Years 3 and 4 were used as the covariate. For the written test, mean Z-scores of the three previous summative written examinations in Years 3 and 4 were used as the covariate.

Interrater reliability for the two examiners on the OSCE was calculated using a two-way random absolute agreement intraclass correlation (ICC2,2) [23]. The reliability of the total score was calculated based on the mean of the two examiners' scores using Cronbach's alpha. Reliability of the written test was calculated using the Kuder-Richardson Formula 20. Item difficulty was given by the proportion of students who answered individual items correctly, and item discrimination by the Point-Biserial Correlation.

The first part of the survey was divided into one question (seniority of doctors providing feedback) and five scales (quantity of feedback, quality of feedback, learning, confidence, and motivation) consisting of 3–11 items. Three items were removed to improve internal consistency of scales, which were calculated using Cronbach's alpha. Mann-Whitney U-tests were used to compare groups because of non-normality of data. Correction for multiple comparisons was not performed on the basis that this study is considered preliminary, and all comparisons were planned ahead and reported in their entirety. Statistical analyses were performed using IBM SPSS Statistics 25 (SPSS Inc., Chicago, IL, USA).

Free text answers on the second part of the survey (mini-CEX) were analysed using Systematic Text Condensation (STC) according to Malterud's description [24]. NVivo 11 (QSR International Pty Ltd., Melbourne, Australia) was used to conduct the analysis.

Results

Characteristics of mini-CEX encounters

A total of 160 mini-CEX assessments were collected and analysed (Table 1). Each participant completed a mean number of 8.4 mini-CEX assessments (standard deviation 0.8; range 8–10). Of the 160 encounters, 54% occurred in general medicine, 43% in general surgery and orthopaedics, and 3% in anaesthesiology. For additional characteristics, see Additional file 1.

Student perceptions of mini-CEX assessments

The majority (79%, 15/19) of participants in the intervention group were positive or very positive to the use of mini-CEX assessments during their clinical placement (Table 2). About 58% (11/19) of participants found mini-CEX assessments useful or very useful in their clinical placement. Only three participants found the assessments useless.

A minority of the participants reported that a doctor had refused to do an assessment. Reasons were being too busy (100%, 4/4), and lack of training and not being confident in how to perform assessments (25%, 1/4).

Ninety-five percent (18/19) of participants gave free text comments on the use of mini-CEX. Analysis of comments fell within two main themes, each with two subthemes: *Feedback* (usefulness of feedback, forcing observation and feedback) and *feasibility* (difficulty conducting assessments, need for assessor training).

Table 1 Characteristics of mini-CEX assessments

	Frequency (% of total)[a]	Mean (SD)	Range
Assessment			
History taking	117 (73.1)	7.55 (1.19)	3–9
Physical examination	113 (70.6)	7.40 (1.26)	3–9
Professionalism	158 (98.8)	8.04 (1.00)	5–9
Clinical reasoning	77 (48.1)	7.44 (1.27)	3–9
Counselling	50 (31.3)	7.50 (1.28)	4–9
Organisation/efficiency	128 (80.0)	7.34 (1.36)	3–9
Overall	114 (71.3)	7.71 (0.99)	5–9
Feedback			
Especially good	135 (83.8)		
Suggestions for improvement	112 (70.0)		
Time			
Observation (minutes)	149 (93.1)	19.8 (14.7)	2.0–90.0
Feedback (minutes)	140 (87.5)	5.6 (4.5)	0–30.0
Total no. of mini-CEXs	160 (100.0)		

Note: [a] denotes the number of mini-CEX forms (and percent of total number of forms) on which each competency, feedback or time spent was recorded.

Table 2 Responses to survey on mini-CEX assessments

	Mean (SD)
Expectations for the use of mini-CEX[a]	4.2 (0.9)
Confidence that mini-CEX is a true reflection of your abilities[b]	2.9 (1.0)
Ease of finding doctors to conduct mini-CEX[c]	3.2 (0.7)
Usefulness of mini-CEX in clinical placement[d]	3.5 (1.0)
	N (% of respondents)
Planning of mini-CEX	
Pre-planned	13 (68.4)
Ad hoc/on the job	6 (31.6)
Retrospective	0 (0.0)
Time taken after mini-CEX to receive feedback	
Immediately	9 (47.4)
< 30 min	9 (47.4)
< 2 h	1 (5.3)
> 2 h	0 (0.0)
Time taken after mini-CEX to receive form	
Immediately	9 (47.4)
< 30 min	10 (52.6)
< 2 h	0 (0.0)
> 2 h	0 (0.0)
Doctor refuse to carry out mini-CEX	
Yes	4 (21.1)
No	15 (78.9)

[a] *1* Very negative, *2* negative, *3* neutral, *4* positive, *5* very positive
[b] *1* Very unconfident, *2* unconfident, *3* neutral, *4* confident, *5* very confident
[c] *1* Very difficult, *2* difficult, *3* neutral, *4* easy, *5* very easy
[d] *1* Very useless, *2* useless, *3* neutral, *4* useful, *5* very useful

Usefulness of feedback

Participants were divided in their perception of the usefulness of mini-CEX assessments. Some commented that feedback had been very valuable for their learning and development, and wished they had done more assessments. Others commented that feedback had been less useful than expected. Many participants commented that they would have liked to receive more constructive feedback on what could be improved: "*I have found [assessments] very useful, especially when assessors have taken the time to give both positive and constructive feedback. Some assessors did not come up with any suggestions for improvement, whereby it loses its purpose.*" Some participants felt that feedback from more experienced doctors, such as specialty registrars and consultants, was or would have been more useful.

Forcing observation and feedback

Some participants remarked on the value of mini-CEX assessments in terms of 'forcing' observation and feedback: "*Mini-CEX assessments are a fantastic way of 'forcing' doctors to observe you conducting a clinical examination or history.*" One participant also commented that assessments made asking for constructive feedback easier, because it was part of the form.

Difficulty conducting assessments

Many participants felt that finding a time or suitable clinical setting was challenging, especially as assessors were often too busy. Some participants pointed out that ease of conducting assessments varied between departments, medicine being easier than surgery. Some participants stated they would have liked doctors to suggest performing mini-CEX from time to time.

Need for assessor training

Some participants experienced that doctors did not have training in how to conduct assessments and give feedback.

Impact on clinical skills (OSCE) and knowledge (written test)

Characteristics of the OSCE are presented in Table 3. Mean total score based on the average of the two

Table 3 Characteristics of OSCE

Station	Topic	Skills assessed	Total score possible	Examiner 1 mean raw score (SD)	Examiner 2 mean raw score (SD)	Cronbach's alpha if item deleted[a]
1	Febrile neutropenia	H, CR	30	18.9 (2.5)	18.0 (2.4)	0.64
2	Ruptured AAA	PE, CR	22	20.4 (2.7)	20.5 (2.9)	0.64
3	Transient ischaemic attack	PE, CR	32	24.0 (3.2)	24.2 (2.6)	0.65
4	Tachycardia-induced myopathy	H, CR	30	14.1 (1.8)	14.1 (1.8)	0.67
5	Pulmonary embolism	H, CR	32	17.6 (2.7)	17.2 (3.2)	0.63
6	Osteoarthritis of the hip	PE, CR	32	20.9 (4.1)	22.4 (3.3)	0.68
						Cronbach's alpha[a]
Total			178	115.8 (10.9)	116.4 (9.9)	0.69

AAA abdominal aortic aneurysm, *H* history taking, *PE* physical examination, *CR* clinical reasoning
[a]Cronbach's alpha calculation based on the mean of the two examiner scores

examiners' scores was 116.1 (65.2%). Mean percentage scores on stations ranged from 61.5% (Station 1) to 75.3% (Station 3). Interrater reliability was found to be 0.92 and Cronbach's alpha was 0.69 for total test scores. For the written test, the mean total score was 21.8 (50.8%) and reliability (KR-20) was 0.44. Mean item difficulty was 0.51 and mean item discrimination (point-biserial correlation) was 0.20.

Table 4 compares mean percentage scores on the OSCE and written test between the intervention and control group. Observed mean scores on the OSCE were 3.4% higher in the intervention group. When past OSCE Z-scores were controlled for, the difference between the group means decreased to 2.4%. Neither of these were statistically significant.

Observed mean scores on the written test were 4.8% higher in the intervention group. When past written examination Z-scores were controlled for, the difference between the group means decreased to 3.4%. Neither of these were statistically significant.

Perceptions of direct observation, feedback and learning

Both groups reported that doctors in their first year of training most frequently provided feedback and supervision. More experienced junior doctors and consultants provided feedback to a lesser extent.

Table 5 presents a summary of survey items and scales. There was good internal consistency in the data looking at the entire scale with a Cronbach's alpha of 0.84. There were no statistically significant differences between the two groups with respect to the five scales. Statistically significant differences were found for only two of the survey items: feedback on history taking was more commonly reported in the intervention group, and students in the intervention group perceived their own ability to identify normal and abnormal findings higher than those in the control group.

Discussion

In this study, formative mini-CEX assessments were compared to traditional ad-hoc feedback to examine student perceptions and effects on direct observation, feedback and learning outcomes. Students were positive towards the use of mini-CEX, and most found them helpful for their learning. We found no differences between the groups with regards to direct observation, feedback or learning outcome.

Implementation of formative mini-CEX assessments in an undergraduate clinical placement was feasible, and all

Table 4 Comparison of mean percentage scores on OSCE and written test between intervention and control group

	N	Observed mean % score (SD)	ANOVA	Adjusted mean % score (SE)	ANCOVA
OSCE					
Intervention	19	0.669 (0.053)	F = 3.603, *p* = 0.066	0.664 (0.012)[a]	F = 1.884, *p* = 0.179[b]
Control	19	0.635 (0.056)		0.640 (0.012)[a]	
Written test					
Intervention	19	0.532 (0.090)	F = 2.674, *p* = 0.111	0.525 (0.020)[c]	F = 1.395, *p* = 0.245[d]
Control	19	0.484 (0.094)		0.491 (0.020)[c]	

[a]Adjustments based on mean Z-scores of past OSCE = 0.102; [b]Homogeneity of regression tested and not significant: F = 0.088, p > 0.05; [c]Adjustments based on mean Z-scores of past written examinations = 0.029; [d]Homogeneity of regression tested and not significant: F = 0.552, *p* > 0.05

Table 5 Survey scales with comparisons of mean scores between intervention and control group

Scale	Cronbach's alpha	Intervention group, mean (SD)	Control group, mean (SD)	Mann-Whitney U test
Quantity of feedback	**0.61**	**2.5 (0.4)**	**2.4 (0.5)**	$\boldsymbol{p = 0.39}$
History taking[1]		3.0 (0.6)	2.2 (0.7)	$p < 0.01$*
Physical examination[1]		2.8 (0.6)	2.5 (0.6)	$p = 0.15$
Procedures[1]		3.0 (0.7)	3.0 (0.7)	$p = 0.84$
Clinical reasoning[1]		2.4 (0.7)	2.7 (0.7)	$p = 0.21$
Presenting findings/cases[1]		2.3 (0.7)	1.9 (0.9)	p = 0.21
Satisfaction with amount of feedback[2]		2.5 (0.9)	2.5 (1.0)	$p = 0.77$
Would have liked more feedback[a, 2]		1.4 (0.5)	1.7 (0.9)	$p = 0.37$
Quality of feedback[2]	**0.75**	**3.1 (0.6)**	**3.3 (0.6)**	$\boldsymbol{p = 0.64}$
Direct observation		2.3 (0.9)	2.7 (1.0)	$p = 0.16$
Positive feedback		3.7 (0.7)	3.2 (0.9)	$p = 0.08$
Constructive, negative feedback		2.8 (0.9)	2.7 (0.7)	$p = 0.71$
Guidance on how to improve		3.3 (0.9)	3.4 (0.8)	$p = 0.73$
Wide range of patients		3.0 (1.2)	3.0 (1.0)	$p = 0.86$
Quality of feedback		3.0 (0.9)	3.3 (0.9)	$p = 0.44$
Usefulness of feedback		3.6 (1.0)	4.0 (0.9)	$p = 0.28$
Feedback made me learn more		3.5 (0.9)	4.1 (1.0)	$p = 0.09$
Learning[2]	**0.64**	**3.9 (0.3)**	**3.8 (0.4)**	$\boldsymbol{p = 0.58}$
Identifying key information in the history		4.1 (0.5)	3.8 (0.8)	$p = 0.25$
Efficiency in history taking		4.2 (0.7)	4.1 (0.8)	$p = 0.75$
Structured clinical examination		4.2 (0.9)	4.0 (0.7)	p = 0.25
Efficiency in clinical examination		4.2 (0.6)	4.1 (0.8)	p = 0.86
Identifying normal and abnormal findings		4.2 (0.6)	3.5 (0.7)	$p = 0.02$*
Carrying out procedures		3.8 (0.7)	3.6 (1.0)	$p = 0.43$
Suggesting differential diagnoses		3.5 (0.7)	3.7 (0.9)	$p = 0.27$
Suggesting further investigations		3.8 (0.4)	3.9 (0.7)	$p = 0.56$
Knowing which topics that I master		3.4 (0.6)	3.6 (0.9)	$p = 0.34$
Knowing which examinations that I master		3.8 (0.4)	3.7 (0.9)	$p = 1.00$
Knowing which procedures that I master		3.9 (0.5)	4.1 (0.6)	p = 0.34
Confidence[2]	**0.74**	**3.6 (0.6)**	**3.7 (0.7)**	**p = 0.84**
Not afraid of asking for help		4.2 (0.6)	4.4 (0.6)	$p = 0.35$
Not afraid of asking for feedback		3.7 (0.9)	3.6 (0.9)	p = 0.77
Confidence in performing tasks expected of a fifth-year medical student		3.2 (1.0)	3.2 (0.8)	$p = 0.89$
Confidence in having learned enough		3.3 (0.9)	3.4 (1.1)	p = 0.75
Motivation[2]	**0.30**	**3.6 (0.6)**	**3.5 (0.5)**	$\boldsymbol{p = 0.23}$
Motivation to meet/clerk patient		4.1 (0.8)	3.9 (0.7)	$p = 0.49$
Motivation to learn		3.8 (0.9)	3.6 (1.0)	p = 0.64
Regularly sought medical knowledge		3.1 (0.8)	2.8 (0.8)	p = 0.44

[1] *1* never, *2* rarely, *3* sometimes, *4* often, *5* always

[2] *1* strongly disagree, *2* disagree, *3* neutral, *4* agree, *5* strongly agree

Note: [a] denotes item that was reverse scored; * denotes items where difference was statistically significant at $p < 0.05$

participants met the pre-planned number of assessments. Assessments were completed in a mean of approximately 25 min, 20 min for observation and 5–6 min for feedback, which is in line with both the intention and the published research [8, 25]. The assessments covered a wide range of common clinical problems, and all participants met the pre-planned requirement of eight mini-CEX encounters. This is higher than completion rates reported in most other studies, with a recent systematic review finding mixed results but rates generally above 50% [5, 7, 13, 25]. This may be explained by several factors. Firstly, our study took place in an undergraduate setting, where doctors are already used to supporting students when seeing patients. Secondly, a small number of students per hospital and allowing all doctors to carry out assessments, thereby minimising workload per doctor. Thirdly, our participants typically spent seven weeks in the same rotation, which may have contributed to facilitating assessments. Short rotations have been found to make assessments and meaningful feedback more challenging, as trainees and supervisors do not get to know each other [26].

Despite the high completion rate, many participants commented that finding a time or suitable clinical setting was challenging, and assessors were often perceived to be busy. Feasibility issues relating to time constraints have been identified in numerous other studies [22, 26–28]. However, it is encouraging to see that only four participants reported that a doctor had refused to do an assessment. Previous recommendations for facilitating implementation of WBAs have emphasised the need for ensuring the necessary resources, including time and faculty development [26].

Student perceptions

Most students were positive to the use of mini-CEX assessments and found them useful during their clinical placement. Participants recognised the importance of constructive feedback, and would have liked more feedback on areas of improvement. While most studies show that trainees value feedback and find assessments useful [4, 5, 29]; others found that trainees regard WBAs as a tick-box exercise or a waste of time [22, 30]. We did not find the latter in our study, possibly explained by the voluntary inclusion and emphasis on the assessments' formative nature.

A number of participants did not feel confident that the mini-CEX assessments gave a true reflection of their capabilities. Similar results among junior doctors have been described previously [22]. This could reflect the students' perception that feedback was limited, or a need to train assessors for accurate scoring. Previous research has shown that raters seldom use the full nine-point scale and leniency in scoring is common, which is also the case in our study [9].

Effects on direct observation and feedback

Implementing formative mini-CEX assessments did not lead to reported increase of direct observation or feedback overall. Direct observation of clinical skills was reported as infrequent in both groups, and the majority were not satisfied with the amount of feedback they received. This may be explained by different expectations to or perceptions of what constitutes direct observation and feedback. The intervention group, having been introduced to the mini-CEX both through theory and practice, may have expected more of their feedback conversations in terms of both quantity and quality. In order to study the genuine difference, field studies are needed.

However, feedback on history taking was reported significantly more common in the intervention group. This is encouraging, as concerns have been raised over supervisors basing their assessments of trainees' clinical skills on proxy information, such as inferring history takings skills based on the case presentation [31, 32]. Some participants highlighted the mini-CEX's value in terms of 'forcing' observation and feedback, and this may be especially relevant for more time-consuming skills such as history taking.

Both groups indicated that junior doctors most frequently provided supervision and feedback, and some participants felt that feedback from more experienced doctors would be more useful. We know from previous research that credibility is an important determinant of how impactful feedback is [33, 34]. This includes trainees' perceptions of supervisor characteristics such as experience [34]. However, this must be weighed against feasibility aspects. If direct observation and feedback can only be given by experienced doctors, workload on the few increases, and less experienced doctors are deprived of situations in which they can develop their skills as supervisors. This should also be supported by robust faculty development to improve their skills as educators.

Educational impact

Educational impact can be classified according to Kirkpatrick's framework, later adapted for medical education research by Barr and colleagues [35, 36]. In this study, we have presented both self-reported outcome measures (Kirkpatrick level 1) and impact on performance (Kirkpatrick level 2b). We found that for self-reported improvement in performing key tasks, such as history taking and clinical examination, there was no statistically significant difference between the groups overall. Interestingly though, the intervention group perceived their ability to identify normal and abnormal findings significantly higher than the control group. This may indicate that students use mini-CEX assessments as learning situations in which their clinical findings can be verified by a more experienced doctor. In this case, there is a recognised knowledge gap from the student's point of view, and feedback given is both specific and actionable, and therefore more likely to be effective [37].

Performance on the OSCE and written test found slightly higher scores in the intervention group, though not statistically significant. This contrasts two previous studies that have shown positive effects on trainee performance, although none of these were randomised controlled studies [16, 17].

The inconsistent findings may be explained by several factors. Firstly, all studies have used general outcome measures, which may have left a large proportion of the effect invisible [25]. Secondly, it is logical to think that educational impact of the mini-CEX depends heavily on the quality of the feedback conversation following the assessment. Although we have little data with regards to the content in these conversations, we found that positive feedback was provided on over 80% of forms and suggestions for improvement in 70% of forms. The quality of feedback provided on WBA forms was the topic of a study by Vivekananda-Schmidt and colleagues, who found that only around 40% of forms contained free-text comments and goal-oriented feedback to support trainee development was uncommon [38]. Further research into the efficacy of formative mini-CEXs should also consider the quality of feedback conversations and its impact on learning.

Strengths and weaknesses

There are several limitations to our study. The study is small and the effect size of approximately one standard deviation may be too large to be realistically expected of the intervention. Regrettably, we were not able to include the number of participants needed to achieve adequate power to evaluate the written test, as we did not have resources available to include additional hospitals in the study. The results from the written test are further limited by low reliability, most probably as a consequence of few items. Another limitation related to the analyses is that the increase in error across multiple comparisons was not controlled, but we consider the research preliminary and encourage replication of its findings. Additionally, generalisability may be limited by the study being a single-institution study. However, we believe that including both general medicine and surgery, as well as multiple hospitals, strengthen the generalisability of our findings. This is, to our knowledge, the first randomised controlled study of the effects of mini-CEX on direct observation, feedback and educational impact. The study included both self-reported and objective data on performance. Performance data was controlled for baseline competence in the form of scores from previous examinations, and scoring was blinded as to what group the participants belonged to.

Conclusions

There is still considerable potential in assessing medical students during clinical placements and in routine practice, but the educational impact of formative assessments remains mostly unknown. We found that the mini-CEX is feasible and students are generally positive towards their use. However, we found no measurable effects with regards to overall feedback, or performance on summative tests. This study contributes to the ongoing discussion with a robust study design, and may serve as a basis for future research.

Abbreviations

WBA: Workplace-based assessment; mini-CEX: Mini-Clinical Evaluation Exercise; OSCE: Objective Structured Clinical Examination; MCQ: Multiple Choice Questions; ANOVA: One-way analysis of variance; ANCOVA: One-way analysis of covariance; STC: Systematic Text Condensation

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12909-021-02670-3.

Additional file 1. Characteristics of mini-CEX encounters.

Acknowledgements

We gratefully thank the standardised patients: Karin Wisløff, Rolf-Ulrik Wisløff, Brit Lysholm, Oljana Parten, Stian Vist Solbakken and Kjell Ivar Aune. We also thank Mona Dalland Stormo for helping us with examination data.

Authors' contributions

S.M., B.L. and T.S. conceived and designed the study. S.M., T.E., E.B. and E.S. collected data for the study. S.M. performed the data analysis, and T.E., E.B., E.S., B.L. and T.S. contributed to data interpretation. S.M. drafted the manuscript. All authors contributed in critical revision of the manuscript and approved the final version to be published.

Funding

The study was supported by the Norwegian University of Science and Technology (NTNU) programme for teaching excellence (NTNU Teaching Excellence).

Availability of data and materials

The datasets used and analysed in this study are available from the corresponding author on reasonable request.

Declarations

Ethics approval and consent to participate

All participants provided written and informed consent. The methods used in this study were carried out in accordance with relevant guidelines and regulations. The study was approved by the Norwegian Centre for Research Data (project number: 56646). The Norwegian Centre for Research Data acts as the national ethics committee for research projects which do not involve patients or health data.

Consent for publication

All individuals thanked under the heading 'Acknowledgements' have provided written consent for their names being mentioned. Otherwise, the manuscript contains no individual person's data.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [2]Department of Circulation and Medical Imaging, Norwegian University of Science and Technology (NTNU), Trondheim, Norway. [3]Clinic of Cardiology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [4]Clinic of Thoracic and Occupational Medicine, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway. [5]Clinic of Medicine and Rehabilitation, Ålesund Hospital, Møre og Romsdal Hospital Trust, Ålesund, Norway. [6]Clinic of Medicine and Rehabilitation, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway. [7]Department of Haematology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

Received: 6 December 2020 Accepted: 14 April 2021
Published online: 21 April 2021

References

1. Kogan JR, Holmboe ES, Hauer KE. Tools for direct observation and assessment of clinical skills of medical trainees: a systematic review. Jama. 2009;302(12):1316–26. https://doi.org/10.1001/jama.2009.1365.
2. Frank JR, Snell LS, Cate OT, Holmboe ES, Carraccio C, Swing SR, et al. Competency-based medical education: theory to practice. Med Teach. 2010; 32(8):638–45. https://doi.org/10.3109/0142159X.2010.501190.
3. Norcini J, Burch V. AMEE guide 31: workplace-based assessments as an educational tool. Med Teach. 2007;29(9):855–71. https://doi.org/10.1080/01421590701775453.
4. Alves de Lima A, Barrero C, Baratta S, Castillo Costa Y, Bortman G, Carabajales J, et al. Validity, reliability, feasibility and satisfaction of the mini-clinical evaluation exercise (mini-CEX) for cardiology residency training. Med Teach. 2007;29(8):785–90. https://doi.org/10.1080/01421590701352261.
5. Wilkinson JR, Crossley JG, Wragg A, Mills P, Cowan G, Wade W. Implementing workplace-based assessment across the medical specialties in the United Kingdom. Med Educ. 2008;42(4):364–73. https://doi.org/10.1111/j.1365-2923.2008.03010.x.
6. Prins SH, Brøndt SG, Malling B. Implementation of workplace-based assessment in general practice. Educ Prim Care. 2019;30(3):133–44. https://doi.org/10.1080/14739879.2019.1588788.
7. Kogan JR, Bellini LM, Shea JA. Feasibility, reliability, and validity of the mini-clinical evaluation exercise (mCEX) in a medicine core clerkship. Acad Med. 2003;78(10):S33–S5. https://doi.org/10.1097/00001888-200310001-00011.
8. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med. 1995;123(10):795–9. https://doi.org/10.7326/0003-4819-123-10-199511150-00008.
9. Hawkins RE, Margolis MJ, Durning SJ, Norcini JJ. Constructing a validity argument for the mini-clinical evaluation exercise: a review of the research. Acad Med. 2010; 85(9):1453–61. https://doi.org/10.1097/ACM.0b013e3181eac3e6.
10. Durning SJ, Cation LJ, Markert RJ, Pangaro LN. Assessing the reliability and validity of the mini—clinical evaluation exercise for internal medicine residency training. Acad Med. 2002;77(9):900–4. https://doi.org/10.1097/00001888-200209000-00020.
11. Boulet JR, McKinley DW, Norcini JJ, Whelan GP. Assessing the comparability of standardized patient and physician evaluations of clinical skills. Adv Health Sci Educ. 2002;7(2):85–97. https://doi.org/10.1023/A:1015750009235.
12. Hatala R, Ainslie M, Kassen BO, Mackie I, Roberts JM. Assessing the mini-clinical evaluation exercise in comparison to a national specialty examination. Med Educ. 2006;40(10):950–6. https://doi.org/10.1111/j.1365-2929.2006.02566.x.
13. Norcini JJ, Blank LL, Duffy FD, Fortna GS. The mini-CEX: a method for assessing clinical skills. Ann Intern Med. 2003;138(6):476–81. https://doi.org/10.7326/0003-4819-138-6-200303180-00012.
14. Wiles C, Dawson K, Hughes T, Llewelyn J, Morris H, Pickersgill T, et al. Clinical skills evaluation of trainees in a neurology department. Clin Med. 2007;7(4):365–9. https://doi.org/10.7861/clinmedicine.7-4-365.
15. Lörwald AC, Lahner F-M, Nouns ZM, Berendonk C, Norcini J, Greif R, et al. The educational impact of Mini-Clinical Evaluation Exercise (Mini-CEX) and Direct Observation of Procedural Skills (DOPS) and its association with implementation: A systematic review and meta-analysis. PloS one. 2018; 13(6):e0198009.
16. Kim S, Willett LR, Noveck H, Patel MS, Walker JA, Terregino CA. Implementation of a mini-CEX requirement across all third-year clerkships. Teach Learn Med. 2016;28(4):424–31. https://doi.org/10.1080/10401334.2016.1165682.
17. Suhoyo Y, Schönrock-Adema J, Rahayu GR, Kuks JB, Cohen-Schotanus J. Meeting international standards: a cultural approach in implementing the mini-CEX effectively in Indonesian clerkships. Med Teach. 2014;36(10):894–902. https://doi.org/10.3109/0142159X.2014.917160.
18. Downing SM. Guessing on selected-response examinations. Med Educ. 2003;37(8):670–1. https://doi.org/10.1046/j.1365-2923.2003.01585.x.
19. Hattie J, Timperley H. The power of feedback. Rev Educ Res. 2007;77(1):81–112. https://doi.org/10.3102/003465430298487.
20. Artino AR Jr, La Rochelle JS, Dezee KJ, Gehlbach H. Developing questionnaires for educational research: AMEE guide no. 87. Med Teach. 2014;36(6):463–74. https://doi.org/10.3109/0142159X.2014.889814.
21. Collins D. Pretesting survey instruments: an overview of cognitive methods. Qual Life Res. 2003;12(3):229–38. https://doi.org/10.1023/A:1023254226592.
22. Bindal T, Wall D, Goodyear HM. Trainee doctors' views on workplace-based assessments: are they just a tick box exercise? Med Teach. 2011;33(11):919–27. https://doi.org/10.3109/0142159X.2011.558140.
23. Landers RN. Computing intraclass correlations (ICC) as estimates of interrater reliability in SPSS. Winnower. 2015;2:e143518.
24. Malterud K. Systematic text condensation: a strategy for qualitative analysis. Scand J Public Health. 2012;40(8):795–805. https://doi.org/10.1177/1403494812465030.
25. Mortaz Hejri S, Jalili M, Masoomi R, Shirazi M, Nedjat S, Norcini J. The utility of mini-Clinical Evaluation Exercise in undergraduate and postgraduate medical education: A BEME review: BEME Guide No. 59. Med Teach. 2020;42(2):125–42.
26. Lörwald AC, Lahner F-M, Mooser B, Perrig M, Widmer MK, Greif R, et al. Influences on the implementation of mini-CEX and DOPS for postgraduate medical trainees' learning: a grounded theory study. Med Teach. 2019;41(4): 448–56. https://doi.org/10.1080/0142159X.2018.1497784.
27. Weston PS, Smith CA. The use of mini-CEX in UK foundation training six years following its introduction: lessons still to be learned and the benefit of formal teaching regarding its utility. Med Teach. 2014;36(2):155–63. https://doi.org/10.3109/0142159X.2013.836267.
28. Davies H, Archer J, Southgate L, Norcini J. Initial evaluation of the first year of the foundation assessment Programme. Med Educ. 2009;43(1):74–81. https://doi.org/10.1111/j.1365-2923.2008.03249.x.
29. Weller JM, Jolly B, Misur M, Merry A, Jones A, Crossley JM, et al. Mini-clinical evaluation exercise in anaesthesia training. Br J Anaesth. 2009;102(5):633–41. https://doi.org/10.1093/bja/aep055.
30. Sabey A, Harris M. Training in hospitals: what do GP specialist trainees think of workplace-based assessments? Educ Prim Care. 2011;22(2):90–9. https://doi.org/10.1080/14739879.2011.11493974.
31. Kogan JR, Hatala R, Hauer KE, Holmboe E. Guidelines: the do's, don'ts and don't knows of direct observation of clinical skills in medical education. Perspect Med Educ. 2017:1–20.
32. Pulito AR, Donnelly MB, Plymale M, Mentzer J, Robert M. What do faculty observe of medical students' clinical performance? Teach Learn Med. 2006; 18(2):99–104. https://doi.org/10.1207/s15328015tlm1802_2.
33. Sargeant J, Mann K, Ferrier S. Exploring family physicians' reactions to multisource feedback: perceptions of credibility and usefulness. Med Educ. 2005;39(5):497–504. https://doi.org/10.1111/j.1365-2929.2005.02124.x.
34. Bing-You RG, Paterson J, Levine MA. Feedback falling on deaf ears: residents' receptivity to feedback tempered by sender credibility. Med Teach. 1997;19(1):40–4. https://doi.org/10.3109/01421599709019346.
35. Kirkpatrick DL. Evaluation of training. In: Craig R, Bittel L, editors. Training and development handbook. New York: McGraw Hill; 1967.
36. Barr H, Freeth D, Hammick M, Koppel I, Reeves S. Evaluations of interprofessional education. London: United Kingdom Review of Health and Social Care; 2000.
37. Lefroy J, Watling C, Teunissen PW, Brand P. Guidelines: the do's, don'ts and don't knows of feedback for clinical education. Perspect Med Educ. 2015; 4(6):284–99. https://doi.org/10.1007/s40037-015-0231-7.
38. Vivekananda-Schmidt P, MacKillop L, Crossley J, Wade W. Do assessor comments on a multi-source feedback instrument provide learner-centred feedback? Med Educ. 2013;47(11):1080–8. https://doi.org/10.1111/medu.12249.

Publisher's Note

Paper III

cogent medicine

EVIDENCE-BASED MEDICINE & MEDICAL INFORMATICS | RESEARCH ARTICLE

Improving assessment quality in professional higher education: Could external peer review of items be the answer?

Susanne Skjervold Smeby[1*], Børge Lillebo[2,3], Vidar Gynnild[4], Eivind Samstad[1,5], Rune Standal[6], Heidi Knobel[1,7], Anne Vik[8,9] and Tobias S. Slørdahl[1,10]

Received: 18 January 2019
Accepted: 20 August 2019
First Published: 27 August 2019

*Corresponding author: Susanne Skjervold Smeby, Olav Kyrres gate 9, PLUS-senteret, Medisinsk-teknisk forskningssenter, Trondheim 7030, Norway
E-mail: susanne.s.smeby@ntnu.no

Reviewing editor:
Anne Spurkland, University of Oslo, Norway.

Additional information is available at the end of the article

Abstract: Summative assessment in professional higher education is important for student learning and making sound decisions about advancement and certification. Despite rigorous pre-test quality assurance procedures, problematic assessment items are always discovered post-test. This article examines the implementation of external peer review of items by clinicians in a six-year undergraduate medical programme. The purpose of the article is to identify to what extent clinicians consider multiple choice items to be acceptable for use in examinations, and what comments they provide on items they believe should be revised or not be used at all. 170 clinicians were recruited and reviewed 1353 multiple choice questions. Results showed that one out of five items reviewed were not approved. There were three main reasons for not approving items: (i) relevance of item content, (ii) accuracy of item content and (iii) technical item writing flaws. The article provides insight into a promising quality assurance procedure suitable for in-house examinations in professional higher education.

Subjects: Higher Education; Assessment; Medical Education

Keywords: assessment quality; item relevance; medical education; multiple choice questions

1. Introduction

Professional higher education strives to teach students the competencies they will need in their future professions. This encompasses both subject-specific and generic competencies that prepare students for the complex problems of today's workplace, as well as life-long learning and development (Baartman, Bastiaens, Kirschner, & Van der Vleuten, 2006; Van der Vleuten, Schuwirth,

ABOUT THE AUTHOR

Susanne Skjervold Smeby is a medical doctor and PhD-student at the Norwegian University of Science and Technology. Her research interests include assessment quality and the relation between assessment and learning.

PUBLIC INTEREST STATEMENT

In professional higher education, the link between test content and professional practice is especially important to make sound certification decisions. This research uses clinicians to review examination content in a medical undergraduate programme. One in five examination questions were not approved by clinicians, and relevance and accuracy of content were two of the main reasons. Consulting practitioners in the field may lead to more relevant and accurate content, increasing the validity of examinations.

Scheele, Driessen, & Hodges, 2010). Developing high quality summative assessment is important for both student learning and sound advancement decisions. In the field of medicine, both employers and patients rely on medical schools' ability to certify that students have the knowledge, skills and attitudes necessary to practice medicine safely.

Summative assessment in undergraduate medical education can be in-house examinations prepared by academic staff involved in teaching, or national examinations generally prepared by licensing organisations. Test items in national examinations are usually written and extensively reviewed by subject-specific test committees trained in item writing. National examinations also provide an arena for relevant stakeholders to engage in the process of assessment design, content and standards for entry into practice (Melnick, 2009). Such measures typically result in high quality assessment, but they come at a high cost (Melnick, 2009). Although the use of national licensing examinations in medicine is likely to increase, the majority of examinations are in-house (Swanson & Roberts, 2016). Therefore, developing quality assurance procedures that can be implemented for in-house settings with fewer resources is important.

Written assessments make up a large part of assessments in medical education, along with assessments that cover other important competencies, such as communication, professionalism and clinical skills. Despite the fact that multiple choice questions (MCQs) have many advantages and disadvantages, they remain the most frequently used assessment method in medicine (Wallach, Crespo, Holtzman, Galbraith, & Swanson, 2006). They are efficient for use in large groups of examinees as they can be administered in a relatively short period of time and are easily computer scored (Downing & Yudkowsky, 2009). Additionally, MCQs can test a large breadth of knowledge as well as higher-level cognitive reasoning (Downing & Yudkowsky, 2009; Schuwirth & Van Der Vleuten, 2004). There are best practice principles for writing effective MCQs, and violating these standards is termed item writing flaws (IWFs) (Case & Swanson, 1998; Haladyna, Downing, & Rodriguez, 2002). IWFs reduce assessment validity by introducing the systematic error of construct-irrelevant variance, and have been shown to occur frequently in in-house examinations (Downing, 2005; Jozefowicz et al., 2002).

Quality assurance procedures around test item development and administration is necessary for high quality assessment (Van der Vleuten et al., 2010). These include faculty development programmes on proper item writing and blueprinting, review of items through internal review committees and psychometric evaluation, as well as student feedback. Item writing workshops have been shown to improve quality of MCQs in terms of difficulty and item discrimination, and reduces the frequency of IWFs (Abdulghani et al., 2015). Several studies have documented the effect of in-house peer review of MCQs (Abozaid, Park, & Tekian, 2017; Malau-Aduli & Zimitat, 2012; Wallach et al., 2006).

In our medical programme, MCQs are subject to review similar to that of the Maastricht model (Verhoeven, Verwijnen, Scherpbier, & Schuwirth, 1999). The departments write items based on a blueprint for the end-of-year examinations, and are entered into a web-based item bank. A multidisciplinary review committee (examination committee) reviews items for content, clarity and IWFs. In addition, one or two senior students are asked to comment on the examination draft. However, despite rigorous review, we still discover problematic items through post-test item analyses and student comments, as is experienced by other institutions (Verhoeven et al., 1999).

In an attempt to reduce the number of problematic items discovered post-test, we developed an additional review process suitable for in-house examinations in professional higher education. We were interested in consulting front line practitioners in the field, inviting them to share their thoughts on examination items through external, double-blinded peer review of MCQs in an undergraduate medical programme. The aim was to explore the following research questions:

(1) To what extent are items considered approved, needing revision or rejected by clinicians?

(2) What comments are provided by clinicians on items considered needing revision or rejected?

(3) To what extent are items changed by the item writer or examination committee following external peer review?

2. Materials and methods

2.1. The medical curriculum and assessment programme

The six-year undergraduate medical programme at the Norwegian University of Science and Technology (NTNU) is integrated and problem-based, featuring one oral and one written summative examination at the end of each year (Ware & Vik, 2009). Written examinations consist of 100–120 single best answer MCQs and several modified essay questions (MEQs). The examinations are pass or fail with a cut-off score of 65%.

2.2. The intervention: external peer review of MCQs

Clinicians as reviewers were recruited with the following inclusion criteria: (a) at least two years of postgraduate training, (b) not completed postgraduate training, although this did not apply to specialists in general practice, (c) does not teach at or write items for the faculty. These criteria were chosen because we considered junior doctors and general practitioners to be qualified to judge whether the content followed recommended clinical guidelines and practice, and its relevance for medical students. All reviewers were required to sign a research consent form and asked to complete a questionnaire on personal background information (Table 1). The study was approved by the Norwegian Centre for Research Data (project number: 45229).

In all, 172 reviewers were recruited, of which two reviewers later withdrew. Recruitment started among colleagues perceived to be highly professionally competent, and continued as snowball sampling in which reviewers recommended their own colleagues. Clinicians were recruited for a period of three years, and the annual work-load was estimated to be two hours. They received no

Table 1. Characteristics of reviewers

Age	
Mean, *years (min, max)*	32 (27, 63)
Gender	
Female, *n (%)*	75 (49.7)
Male, *n (%)*	76 (50.3)
Position	
General practitioner, *n (%)*	8 (5.3)
Junior doctor, *n (%)*	135 (89.4)
Mean number of months approved in specialist training (SD)	29.2 (16.2)
PhD student or researcher, *n (%)*	3 (2.0)
Other, *n (%)*	5 (3.3)
Workplace	
GP surgery, *n (%)*	19 (12.6)
District hospital, *n (%)*	67 (44.4)
University hospital, *n (%)*	61 (40.4)
Other, *n (%)*	4 (2.6)

Response rate: 151 (88%) responded to the questionnaire.

financial compensation for reviews, but were registered as external employees and given access to the university's resources, including IT facilities. We aimed at recruiting clinicians from multiple hospitals and GP surgeries, from different counties in Norway (n = 18), with a background from various medical schools (n = 13) and in various specialities (n = 33).

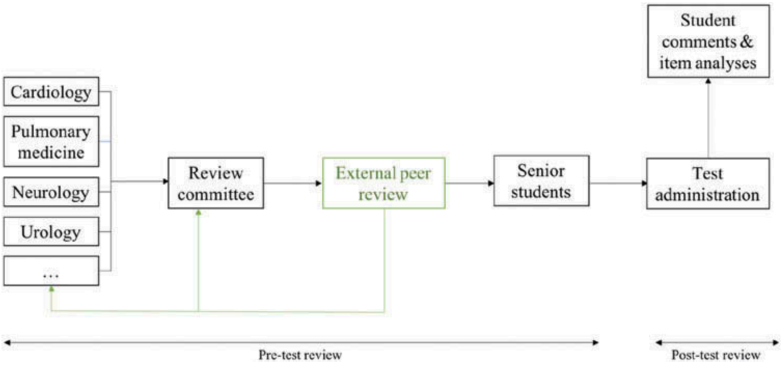
The external peer review was implemented as an additional step in the quality assurance procedure already in place (Figure 1). The items had been reviewed and approved by the multidisciplinary review committee prior to the external peer review. Items were sorted by subject and distributed to reviewers specialising in the topic covered by that item. For subjects that did not have a clear link to a medical specialty, items were pooled and divided between all reviewers. The reviewer and item writer did not know each other's identity. Each reviewer received one to ten items, and each item was assessed by only one reviewer. The whole review process, including distribution of items and completing the review, was carried out by way of a web-based item bank which could be accessed from home. Reviewers had two weeks to complete their review.

The reviewers received limited training with regards to item writing and reviewing. They were sent written information on the MCQ format and the review process, along with the item writing guidelines. Before the correct option was revealed to reviewers, they had to answer each item. Reviewers were asked to consider the questions' relevance, whether the correct option undoubtedly is the best option and the suitability of the explanation of the correct option. They had to indicate whether an item should be approved, revised prior to use, or rejected. If an item was deemed as needing revision or rejected, reviewers were asked to provide a comment. The reviews were disclosed to the item writers, who decided whether to revise or delete the item, or leave it unchanged. If an item was left unchanged, a comment explaining their decision had to be provided to the examination committee. The examination committee made the final decision on whether items should be included in the examination, and could also make changes to or delete items.

2.3. Data collection and analysis

Summative MCQs administered to students in Year 1, 2, 3 and 6 for the academic year of 2015/2016, in addition to previously used MCQs from Year 4, were externally reviewed. This study uses a mixed method approach, with both qualitative and quantitative data to answer the research questions. The following data were registered: Review decision, reviewer comments and whether the item was changed or deleted by the item writer/examination committee. The main part of this study focuses on a qualitative analysis of reviewers' comments to answer the second research question. Reviewers' comments were analysed using Systematic Text Condensation (STC) according to Malterud's description (Malterud, 2012). STC is a descriptive cross-case analysis used to capture significant themes in the empirical material. The analysis started by reading through all

Figure 1. External, double-blinded peer review. The green box indicates where the external peer review was incorporated in the item review process previously in use.



reviewers' comments to get an overall impression of the material and preliminary themes. Meaningful text representing reviewers' reasons for disapproving items was coded into the main themes, adding new themes as they became apparent. In the third step, subthemes within main themes were identified and the contents of each group were condensed into an artificial quote. Lastly, the content of each group was summarised in generalised descriptions and illustrated by selected quotes. All comments were read and themes discussed by three authors (SSS, VG, BL) to widen the analytic space. Quotes were translated, and edited only to improve readability.

3. Results

3.1. Review decisions

Of the 1353 items that were externally reviewed, 282 (20.8%) were not considered approved. Of these, 229 (16.9%) were judged as needing revision and 53 (3.9%) were rejected by reviewers (Table 2). Item writers and examination committees made changes to 115 (40.8%) of disapproved items. Of these, 96 (34.0%) were revised and 19 (6.7%) were deleted. In total, 8.5% of all the items reviewed were changed following external review.

3.2. Reviewer comments

Reviewers' comments fell within three main themes, each with three subthemes: *content relevance* (level of difficulty, importance of content, and cognitive level), *content accuracy* (missing information, content errors, and uncertainty about professional content) and *technical flaws* (spelling and language, structure, and lack of explanation of correct option).

3.2.1. Content relevance

The relevance of item content for medical students was a frequent reason for disapproving items. This included the level of difficulty, importance of content and cognitive level tested by the item.

Level of difficulty: Many reviewers commented that the item content was too difficult for undergraduates. Some remarked that the knowledge asked for was too in-depth, whereas others wrote that the content would be better suited in graduate medical education. One reviewer wrote: "This topic has far more relevance in specialist training than in final undergraduate examinations." Only three items were disapproved by reviewers on the basis of being too easy. One reviewer stated that the content should be presumed knowledge, and therefore unnecessary to ask about in an examination.

Importance of content: Some reviewers commented that the topic covered by items was peripheral as opposed to core areas of the curriculum, and others remarked that the item covered rare symptoms and diseases, and therefore unlikely to be encountered by junior doctors. One such comment was: "The item is irrelevant, and this type of detailed knowledge cannot be deemed essential for clinical practice." Irrelevance for later clinical practice was a frequent reason for not approving items.

Table 2. Review decision. Review decision and subsequent changes made to items by item writer/examination committee

	Review decision			Total
	Approved	Revision needed	Rejected	
Unchanged, *n*	n/a	125	42	1238
Revised, *n*	n/a	90	6	96
Deleted, *n*	n/a	14	5	19
Total, *n*	1071	229	53	1353

Cognitive level: Many items were disapproved because they only tested recall of knowledge. Reviewers commented that such facts would either not be relevant in clinical practice or when needed can be looked up in the literature, as illustrated by the following two quotes: “A very narrow question that only tests students’ ability to recall knowledge,” and “the question should be more comprehensive, enabling students to use their reasoning skills to a greater extent.” This applied especially to items asking for numbers or percentages, for example prevalence of disease.

3.2.2. Content accuracy

The accuracy of the item content was also commonly remarked by reviewers. These comments related to items that were missing key bits of information, had errors in the content or items where reviewers were unsure about the accuracy of the content.

Missing information: Most comments on content accuracy related to missing information in the stem or options, as exemplified by the quote: “There is not enough information in the stem to provide a good and unambiguous answer.” In a few cases, reviewers specified that certain details were missing, thereby making more than one option correct. In one such item, the reviewer wrote: “It should be specified in the stem that this applies for children with a birth weight above 2.5 kg. For children with low birth weight, option C is the correct answer.” Many reviewers commented that the stem or options did not provide information that would normally be present in a real clinical situation, making the item hypothetical rather than realistic: “The item should include more information such as temperature, heart rate and blood pressure, which you would have access to in a real clinical situation.”

Content errors: Reviewers also came across content errors, some of which related to improbable symptoms or findings. Other items were based on outdated guidelines or classification systems. The following quotes illustrate typical examples of errors:

> “The question asks for a probable diagnosis in a patient with a broad complex tachycardia with a ventricular rate that fits well with an atrial flutter with 2:1 conduction. Most patients with a broad complex tachycardia and previous history of MI will have ventricular tachycardia, but not at this ventricular rate … If the ventricular rate is changed to a higher rate, the answer will be correct.”
>
> “The classification used for endometrial hyperplasia is outdated based on WHO guidelines … ”

Uncertainty about professional content: Several reviewers expressed uncertainty about the accuracy of the content. This included uncertainty about whether the content was in line with updated guidelines or current literature, or whether the stem and option were realistic or had missing information, such as: “[I am] unsure whether 15% is right. I have found 20–30% in the literature,” and “I am unsure whether the correct option complies with national guidelines … ”

3.2.3. Technical flaws

Item writing flaws that related to language or structure of the items, here termed technical flaws, were often commented by reviewers.

Spelling and language: In some items, spelling mistakes and typographical errors were pointed out. A few reviewers commented that abbreviations, eponyms and dialect words should be avoided for clarity, as in this case: “Eponyms such as Conn’s syndrome should be avoided.” Other comments related to imprecise wording in the stem and question, long and information dense options, response options with lists of shuffled words, and negative wording. These comments can be summarised in the following quote: “Some students might answer this incorrectly because they are confused by the question.”

Structure: Reviewers commented that some items had superfluous stems or did not have stems at all. Other comments pertained to clues as to which option was correct, for example grammatical clues or longest option being correct.

Explanation of correct option: Many items lacked an explanation of the correct option. Where an explanation was provided, reviewers often requested that explanations be more in-depth or to a larger degree explain certain concepts of the item. Some reviewers commented that distractors should also have an explanation of what did not make this the best option, thereby increasing the learning potential of the item.

4. Discussion

In this study, we implemented external double-blinded peer review of MCQs for in-house examinations. Results showed that of the 1353 items reviewed, 20% of items were either rejected or judged as needing revision by reviewers. Subsequently, changes were made to 40% of disapproved items, which constitutes almost 10% of the total number of MCQs that were reviewed. Relevance and accuracy of item content, as well as technical item writing flaws, were the three main reasons for disapproving items for use.

The double-blinded peer review system ensures that review is not biased by gender, affiliation or seniority, and that reviews can be honest and without fear of retaliation (Shaw, 2015). In higher education, a limitation of internal review can be a reluctance to criticise colleagues, especially when the individual writing that item is considered an expert on the topic (Jozefowicz et al., 2002). In this study, we chose to use junior doctors and general practitioners as reviewers. We asked that they review items on the basis of being clinicians, thereby providing a practitioner's perspective which draws on experience and tacit knowledge. There may be advantages of using junior doctors and general practitioners rather than content experts. Their generalist perspective may contrast that of experienced academic staff responsible for teaching and developing test items, who may overestimate the importance of learning the details in their field (Mcleod & Steinert, 2015). Indeed, standard setting studies have demonstrated that expert judges tend to set unrealistically high passing scores, which could indicate that they expect too much of novice learners (Kane, Crooks, & Cohen, 1999). However, by allowing item writers to decide whether to change the item following review, experts remained responsible for item content. In this way, reviews provided input on item content, rather than a final say. Assessment authenticity and validity may increase when content is informed both by experienced academic staff and front line clinicians' perspective on what is important to know (American Board of Internal Medicine, 2016).

Content relevance emerged as one of the main themes in reviewers' comments on disapproved items. Reviewers commented that many items were too difficult to be appropriate in an undergraduate setting, demanding knowledge that was too in-depth or that concerned rare conditions. Another aspect was the importance of the content tested, with reviewers commenting that items were irrelevant for clinical practice or that items only tested recall of knowledge, as opposed to application and reasoning. This finding is in line with Koens, Rademakers, and Ten Cate (2005) who found that although test items were designed by item writers to assess core medical knowledge, many were judged as testing non-core knowledge by clinicians. The occurrence of test items of low relevance may reflect differing views on what constitutes relevance (Janssen-Brandt, Muijtjens, & Sluijsmans, 2017; Koens, Custers, & Ten Cate, 2006; Koens et al., 2005). In order to reach a more consistent and accurate interpretation of the relevance, Janssen-Brandt et al. (2017) suggest using a rubric of five criteria: 1) medical knowledge (requires study and understanding of medicine), 2) ready knowledge (cannot be looked up quickly), 3) incidence in practice (how often knowledge is needed in practice) 4) prevalence or high-risk (needed for high-prevalence or high-risk situations), and 5) foundation in the medical curriculum. The link between test content and professional practice is especially important in professional higher education in order to make sound inferences about licensing and certification, and irrelevant content is therefore a major threat to test validity (Downing, 2002; Norcini & Grosso, 1998).

Another main theme from reviewers' comments was content accuracy. While most comments related to lack of sufficient information, leaving items too imprecise to identify one best option, others related to errors in the content. These ranged from uncertainties about the accuracy of the professional content to content errors, such as items that were based on outdated guidelines or classification systems. The rapid growth of medical knowledge poses a challenge to deciding and updating curriculum and assessment content. Additionally, if items contain information that is medically inaccurate, the testing effect may increase the likelihood of students remembering erroneous information (Rohrer & Pashler, 2010). This may be especially relevant when storing items in an item bank for reuse on later examinations, running the risk of items becoming outdated in a short period of time (Sadaf, Khan, & Ali, 2012).

The last main theme that emerged from the peer review encompass technical aspects of MCQs, such as errors relating to structure, clues, language and spelling, and items missing an explanation of the correct option. Poorly written MCQs may be falsely more difficult or easy, and be differentially confusing for different subgroups of students, thereby decreasing the fairness of the assessment (Downing, 2002; McCoubrie, 2004; Tarrant & Ware, 2008). Although important, technical aspects could probably be equally or better reviewed by strengthening in-house review. By reducing the frequency of IWFs, in-house peer review has been shown to improve psychometric properties of examinations (Abozaid et al., 2017; Malau-Aduli & Zimitat, 2012; Wallach et al., 2006).

Feasibility of the peer review process was important for implementation in an in-house setting, with fewer financial and staff resources available. The number of items reviewed in one year is approximately the number of MCQs needed yearly for examinations and reassessment in our programme. A small annual work load per reviewer and an IT solution that enabled reviewers to work from home were essential for recruiting reviewers as they received no financial compensation. Furthermore, the IT solution (our web-based item bank) supported the entire review process, including distribution of items to reviewers, review, distributing reviewer comments to item writers and editing items, thereby minimising administrative costs.

The novelty of this study is the implementation of quality assurance of MCQs that is new to an in-house setting. External review could be suitable for other professional higher education programmes, where front-line practitioners can provide useful input on assessment content. In this study, external reviewers received limited training in item writing guidelines and were asked to assess items on the basis of being clinicians. The qualitative data give insight into why junior doctors and general practitioners thought many items should be revised or not be used in examinations. In order to see whether external peer review can affect measures such as reliability and item discrimination, the authors suggest future studies should look into psychometric effects, in addition to its long-term effects on item quality.

5. Conclusions

This study showed that external, double-blinded peer review of MCQs can be implemented for in-house examinations. Approximately one in five items were rejected or judged as needing revision and of these, two in five items were later changed by the item writer. There were three main reasons for not approving items for use: (i) Relevance of item content, (ii) accuracy of item content, and (iii) technical item writing flaws. Using front-line practitioners to review examination content may lead to more relevant and accurate items, increasing the validity of summative assessments.

Acknowledgements
The authors would like to thank all peer reviewers in hospitals and GP surgeries across Norway.

Funding
This work was supported by NTNU Teaching Excellence.

Author details
Susanne Skjervold Smeby[1]
E-mail: susanne.s.smeby@ntnu.no
Børge Lillebo[2,3]
Vidar Gynnild[4]
ORCID ID: http://orcid.org/0000-0001-7589-6057
Eivind Samstad[1,5]
Rune Standal[6]
Heidi Knobel[1,7]
Anne Vik[8,9]
Tobias S. Slørdahl[1,10]
ORCID ID: http://orcid.org/0000-0001-7488-4863
[1] Department of Clinical and Molecular Medicine, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[2] Department of Circulation and Medical Imaging, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[3] Clinic of Medicine and Rehabilitation, Levanger Hospital, Nord-Trøndelag Hospital Trust, Levanger, Norway.
[4] Department of Education and Lifelong Learning, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.

[5] Clinic of Medicine and Rehabilitation, Ålesund Hospital, Møre og Romsdal Hospital Trust, Ålesund, Norway.
[6] Faculty of Medicine and Health Sciences, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[7] Department of Oncology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway..
[8] Department of Neuromedicine and Movement Science, Norwegian University of Science and Technology (NTNU), Trondheim, Norway.
[9] Department of Neurosurgery, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.
[10]Department of Haematology, St. Olavs Hospital, Trondheim University Hospital, Trondheim, Norway.

Declaration of interest statement

The authors declare that they have no competing interests.

Citation information

Cite this article as: Improving assessment quality in professional higher education: Could external peer review of items be the answer?, Susanne Skjervold Smeby, Børge Lillebo, Vidar Gynnild, Eivind Samstad, Rune Standal, Heidi Knobel, Anne Vik & Tobias S. Slørdahl, *Cogent Medicine* (2019), 6: 1659746.

References

Abdulghani, H. M., Ahmad, F., Irshad, M., Khalil, M. S., Al-Shaikh, G. K., Syed, S., … Haque, S. (2015). Faculty development programs improve the quality of multiple choice questions items' writing. *Scientific Reports*, *5*. doi:10.1038/srep09556

Abozaid, H., Park, Y. S., & Tekian, A. (2017). Peer review improves psychometric characteristics of multiple choice questions. *Medical Teacher*, *39*(sup1), S50–S54. doi:10.1080/0142159X.2016.1254743

American Board of Internal Medicine. (2016). More physicians invited to rate exam topics by relevance in practice and to help set exam standard. Retrieved from https://www.abim.org/news/abim-engages-physicians-on-updates-potential-changes-to-moc-assessments.aspx.

Baartman, L. K., Bastiaens, T. J., Kirschner, P. A., & Van der Vleuten, C. P. (2006). The wheel of competency assessment: Presenting quality criteria for competency assessment programs. *Studies in educational evaluation*, *32*(2), 153–170.

Case, S. M., & Swanson, D. B. (1998). *Constructing written test questions for the basic and clinical sciences*. Philadelphia: National Board of Medical Examiners.

Downing, S. M. (2002). Threats to the validity of locally developed multiple-choice tests in medical education: Construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education*, *7*(3), 235–241.

Downing, S. M. (2005). The effects of violating standard item writing principles on tests and students: The consequences of using flawed test items on achievement examinations in medical education. *Advances in Health Sciences Education*, *10*(2), 133–143. doi:10.1007/s10459-004-4019-5

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*. New York: Routledge.

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education*, *15*(3), 309–333. doi:10.1207/S15324818AME1503_5

Janssen-Brandt, X. M., Muijtjens, A. M., & Sluijsmans, D. M. (2017). Toward a better judgment of item relevance in progress testing. *BMC medical education*, *17*(1), 151.

Jozefowicz, R. F., Koeppen, B. M., Case, S., Galbraith, R., Swanson, D., & Glew, R. H. (2002). The quality of in-house médical school examinations. *Academic Medicine*, *77*(2), 156–161. doi:10.1097/00001888-200202000-00016

Kane, M., Crooks, T., & Cohen, A. (1999). Designing and evaluating standard-setting procedures for licensure and certification tests. *Advances in Health Sciences Education*, 4(3), 195–207. doi:10.1023/A:1009849528247

Koens, F., Custers, E. J., & Ten Cate, O. T. (2006). Clinical and basic science teachers' opinions about the required depth of biomedical knowledge for medical students. *28*(3), 234–238. doi:10.1080/01421590500271183

Koens, F., Rademakers, J. J., & Ten Cate, O. T. (2005). Validation of core medical knowledge by postgraduates and specialists. *Medical teacher*, *39*(9), 911–917. doi:10.1111/j.1365-2929.2005.02246.x

Malau-Aduli, B. S., & Zimitat, C. (2012). Peer review improves the quality of MCQ examinations. *Assessment & Evaluation in Higher Education*, *37*(8), 919–931. doi:10.1080/02602938.2011.586991

Malterud, K. (2012). Systematic text condensation: A strategy for qualitative analysis. *Scandinavian Journal of Public Health*, *40*(8), 795–805. doi:10.1177/1403494812465030

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher*, *26*(8), 709–712. doi:10.1080/01421590400013495

Mcleod, P., & Steinert, Y. (2015). Twelve tips for curriculum renewal. *Medical Teacher*, *37*(3), 232–238. doi:10.3109/0142159X.2014.932898

Melnick, D. E. (2009). Licensing examinations in North America: Is external audit valuable? *Medical Teacher*, *31*(3), 212–214.

Norcini, J., & Grosso, L. J. (1998). The generalizability of ratings of item relevance. *Applied Measurement in Education*, *11*(4), 301–309.

Rohrer, D., & Pashler, H. (2010). Recent research on human learning challenges conventional instructional strategies. *Educational Researcher*, *39*(5), 406–412. doi:10.3102/0013189X10374770

Sadaf, S., Khan, S., & Ali, S. K. (2012). Tips for developing a valid and reliable bank of multiple choice questions (MCQs). *Education for Health*, *25*(3), 195. doi:10.4103/1357-6283.109786

Schuwirth, L. W., & Van Der Vleuten, C. P. (2004). Different written assessment methods: What can be said about their strengths and weaknesses? *Medical Education*, *38*(9), 974–979. doi:10.1111/j.1365-2929.2004.01916.x

Shaw, D. M. (2015). Blinded by the light. *EMBO Reports*, *16* (8), 894–897. doi:10.15252/embr.201540943

Swanson, D. B., & Roberts, T. E. (2016). Trends in national licensing examinations in medicine. *Medical Education*, *50*(1), 101–114. doi:10.1111/medu.12810

Tarrant, M., & Ware, J. (2008). Impact of item-writing flaws in multiple-choice questions on student achievement in high-stakes nursing assessments. *Medical Education*, *42*(2), 198–206. doi:10.1111/j.1365-2923.2007.02957.x

Van der Vleuten, C., Schuwirth, L., Scheele, F., Driessen, E., & Hodges, B. (2010). The assessment of professional competence: Building blocks for theory development. *Best Practice & Research Clinical Obstetrics & Gynaecology*, *24*(6), 703–719. doi:10.1016/j.bpobgyn.2010.04.001

Verhoeven, B., Verwijnen, G., Scherpbier, A., & Schuwirth, L. (1999). Quality assurance in test construction: The approach of a multidisciplinary central test committee/ Commentary. *Education for Health*, *12*(1), 49.

Wallach, P. M., Crespo, L., Holtzman, K., Galbraith, R., & Swanson, D. (2006). Use of a committee review process to improve the quality of course examinations. *Advances in Health Sciences Education*, *11*(1), 61–68. doi:10.1007/s10459-004-7515-8

Ware, J., & Vik, T. (2009). Quality assurance of item writing: During the introduction of multiple choice questions in medicine for high stakes examinations. *Medical Teacher*, *31*(3), 238–243. doi:10.1080/01421590802155597

***Cogent Medicine* (ISSN: 2331-205X) is published by Cogent OA, part of Taylor & Francis Group.**

ISBN 978-82-326-7770-2 (printed ver.)
ISBN 978-82-326-7769-6 (electronic ver.)
ISSN 1503-8181 (printed ver.)
ISSN 2703-8084 (online ver.)

NTNU
Norwegian University of
Science and Technology