Say what you think
Show what you do

Feedback interventions to support self-regulated quality improvement in physical therapy

Marjo Maas

Say what you think
Show what you do
Feedback interventions to support self-regulated quality improvement in physical therapy

Marjo Maas

The work presented in this thesis was carried out within the Radboud Institute for Health Sciences, at the department Scientific Center for Quality of Healthcare (IQ healthcare).
The studies in this thesis were financially supported by the HAN University of Applied Sciences, the Royal Dutch Society for Physical Therapy, and SURF foundation.

ISBN 978 94 6295 785 5

Cover and book illustrations
Luuk Huiskes
Design
Robbert Zweegman
Print
ProefschriftMaken || www.proefschriftmaken.nl

Say what you think
Show what you do
Feedback interventions to support self-regulated quality improvement in physical therapy

Marjo Maas

Proefschrift
ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college van decanen
in het openbaar te verdedigen op vrijdag 12 januari 2018
om 10.30 uur precies

door

Margaretha Johanna Maria Maas
geboren op 22 april 1958
te 's-Hertogenbosch

Promotoren	Prof. dr. M.W.G. Nijhuis-van der Sanden
	Prof. dr. C.P.M. van der Vleuten (Universiteit Maastricht)
Copromotoren	Dr. P.J. van der Wees
	Dr. Y.F. Heerkens (HAN)
Manuscriptcommissie	Prof. dr. G.A. Zielhuis (voorzitter)
	Prof. dr. A.W.M. Kramer (Leids Universitair Medisch Centrum)
	Dr. C.R.M.G. Fluit

Contents

Chapter 1
Introduction

A client who seeks the help of a physical therapist deserves the best possible care[1], provided by up-to-date educated professionals who can take responsibility for the quality of their services. These professionals should be capable to respond to rapidly changing client needs and to changing societal demands on the quality of healthcare services.[1–3] Quality aims for the best possible care that apply to the demands of 21st century healthcare, have been designed in 2001 by the Institute of Medicine focusing on how the healthcare delivery system can be designed to innovate and improve care. Six indicators for healthcare improvement were established:

1 Safe – avoiding injuries to patients from the care that is intended to help them.
2 Effective – providing services based on scientific knowledge to all who could benefit and refraining from providing services to those not likely to benefit (avoiding underuse and overuse, respectively).
3 Patient-centered – providing care that is respectful of and responsive to individual patient preferences, needs, and values and ensuring that patient values guide all clinical decisions.
4 Timely – reducing waits and sometimes harmful delays for both those who receive and those who give care.
5 Efficient – avoiding waste, including waste of equipment, supplies, ideas, and energy.
6 Equitable – providing care that does not vary in quality because of personal characteristics such as gender, ethnicity, geographic location, and socio-economic status.

In addition, the Institute of Medicine argued for the need for transparency in healthcare. When healthcare providers communicate effectively with their clients, and healthcare systems make information available, clients and their families are allowed to make informed decisions when selecting a health plan or choosing among alternative treatments.[2] Although these concepts are not new and adopted by healthcare policy makers world-wide, their implementation in clinical practice is still a challenge for individual healthcare providers and provider organizations, despite the variety of implementation strategies applied.[4–6] Suboptimal implementation also involves the healthcare domain of physical therapy. It should be an ethical obligation of physical therapists to self-direct their professional development process to adequately respond to current and future challenges and to publicly account for the quality of their services. The Royal Dutch Society for Physical Therapy (KNGF) developed a Masterplan Quality in Motion (MKIB)[7] that targets the

[1] The term 'care' includes all services described in the professional profile of the physical therapist.

development and implementation of an integrated quality system to meet the increasing societal and political demands on the quality of physical therapy care. It aims to self-regulate continuous improvement and quality assurance of both professional and organizational performance. The MKIB includes the development of quality indicators, the implementation of clinical practice guidelines and patient reported outcome measures (PROMS) for external accountability and internal quality improvement purposes. Self-regulation implies that professionals share the responsibility to account for the quality of their work. Quoting de Vijlder[8]: "Being professional is being accountable."

This thesis explores the concept of self-regulation by challenging physical therapists to take the assessor role in monitoring professional performance and providing performance feedback. The underlying rationale is that valid and reliable assessment of professional performance, needs professional judgment[9,10] and that bottom-up, intrinsically motivated quality improvement initiatives may yield better and more sustainable results than top-down, extrinsically motivated measures. If physical therapists would take the assessor role themselves in assessing the quality of their clinical and organizational performance, the quality and acceptability of feedback might improve.[11,12] In the context of this thesis, professional performance addresses the quality domains *effectiveness, client-centeredness*, and *transparency* of physical therapy care, representing the major challenges for professionals and organizations as explained in the following chapter.

Current and future challenges for performance improvement

To elaborate on the concepts of client-centered, effective, and transparent healthcare including its implications for professional development and behavior change, the following case will be used as an example. The case will be commented by a virtual physical therapist who takes the assessor role in performance assessment.

I met Eric in Greece while he was working on a research report and I was working on my PhD thesis. I saw him working outside on his terrace by the sea. He was a tall and slender man of 32 years old. He worked standing before a high placed (eye level) computer screen using a low placed (hip level) keyboard. This picture triggered my curiosity and Eric was willing to tell the story of his computer-related problems.

Eric is a scientist working daily at his computer. His complaints exist now

for five years. The onset was at the time when he worked on his PhD thesis – a stressful period because he was confronted with several obstacles to complete the thesis – and since then the symptoms have persisted unabated although he developed some pain reducing strategies that enable him to work. In the beginning, only his right forearm was involved, but later also the left arm, his shoulders and the upper spine. At the time he worked on his PhD thesis, he visited a physical therapist who informed him about the necessity of taking regular breaks to relax his forearm muscles and to increase muscle circulation, and to improve his working posture. Eric was treated with a massage of the forearm, a stretching program of the forearm muscles, and working posture improvement advice. Although the massage felt comfortable and the stretching program provided a temporarily relief, the complaints returned when working on his computer, despite improved working posture and taking regular breaks. After nine sessions, the treatment ended with disappointing results. He succeeded in completing his PhD thesis by carefully planning his activities and with pain medication.

Afterwards, he went for two months on a journey abroad, hoping that the symptoms would disappear with rest and they did. But when he started working again as a post-doc researcher, the complaints returned immediately. Eric felt frustrated, worrying about his future. He wondered why he never faced these problems while he was a boy, gaming each day for hours. He decided to visit another physical therapist. This therapist advised him to adjust his working place tailored to his length. He was provided with a training scheme addressing both his general physical fitness and his local muscle endurance while working on his computer according to a graded activity program. Progress was monitored by the Visual Analogue Pain Scale (VAS). Eric started to practice according to the provided graded activity scheme but his compliance to the program lowered by the disappointing results and the impact on his working scheme. Moreover, he didn't succeed in understanding the underlying rationale of this intervention despite his scientific background. In an attempt to meet his client's expectations, the physical therapist suggested to apply kinesio taping[2] *to reduce the load on his forearm muscles while working on the keyboard. Eric appreciated the serious involvement of his therapist but didn't agree with this proposal, because he had serious doubts about the effectiveness of this intervention and the sustainability of the results. On his request, the service unit of the university redesigned his working place according to pre-defined ergonomic guidelines. But these adjustments didn't work for Eric, in contrast, his complaints increased.*

From that moment on, Eric decided to design his own working place. He lifted his computer screen allowing him to work standing and continued

[2] Kinesio taping uses an elastic tape that is fixed onto the skin.

to do so until now. His complaints still exist, but with this self-invented solution he is able to cope with his complaints at work.

Client-centered healthcare

The concept of client-centered care will be explained by taking the assessor role in critical appraising the case of Eric.
Client-centered care implies that physical therapists consider the client perspective to understand what they need and what they don't need, which requires changing attitudes and advanced communication and collaboration skills.[2] It needs explicitly addressing the client's help-request, setting mutual accepted and achievable goals, choosing interventions tailored to client needs and preferences, and defining outcomes in terms of what is meaningful and valuable to the client.[1] This requires a paradigm shift from the professional perspective – traditionally focusing on treating signs and symptoms of limited physical functioning, to the client perspective focusing on enhancing daily activities and societal participation.
To describe health, the professional profile of physical therapists in the Netherlands refers to the concept of positive health (Huber, 2011): "Health as the ability to adapt and to self-manage, in the face of social, physical and emotional challenges".[13] To promote health according to this concept, physical therapists need new knowledge and skills enabling them to enhance healthy behaviors and to empower clients and their families (if relevant) to cope with limitations in daily functioning.[1,2,13] In addition, they need to effectively communicate with their colleagues and other healthcare professionals to align their services to clients' needs according to the local situation, as clients – in particular clients with chronic conditions – may move through many settings of care.[14,15] However, several barriers to the implementation of client-centered care in the physical therapy domain have been identified relating to both healthcare providers and receivers. Research showed that physical therapists often do not take patients' perspectives into account; they promote or recommend specific treatments rather than consider patients' ideas and preferences during the decision-making process.[16] In addition, they tend to immediately provide 'care', taking full responsibility for the outcomes instead of providing intervention alternatives and actively involve clients in the intended outcomes. A review of Shoeb *et al*.[17] showed that clients do not always feel the need to participate activively and physical therapists lack the communication skills to enhance active client involvement. The use

of measurement instruments – and more specific patient reported outcome measures (PROMs) – may support active client involvement. PROMs allow for identifying problems in daily activities that are meaningful to the client and may facilitate the dialogue on goal setting and treatment planning. Moreover, PROMs may trigger clients to monitor and ultimately self-direct their treatment process.[18–21] Although the use of measurement instruments has increased in recent years, routine use of PROMs in daily practice is not yet optimal.[22–24] In sum, client-centeredness in physical therapy shows room for improvement.

Taking the assessor role: was the care provided to Eric client-centered? Eric is a scientist. Although his expertise does not address health problems, he has the knowledge and skills to find relevant answers on the Internet. Surprisingly, Eric's views on his health problem were poorly addressed and in turn – for some reason – Eric didn't share these views with his therapist. His views on the nature of his health problem and his expectations regarding the treatment outcomes were not involved by the physical therapist in goal setting and treatment planning. Looking back, Eric's ideas should have been addressed to assess their validity and to prevent incongruence in outcome expectations. Involving Eric in goal setting, sharing decisions on treatment planning, and aligning outcome expectancies might have enhanced Eric's compliance to the training program and responsibility for the outcomes. The use of patient-specific outcome measures, such as the Patient Specific Complaints questionnaire (PSC)[19] might have supported this process.

Effective healthcare

To meet the increasing demands on the (cost-)effectiveness of physical therapy services, professionals are challenged to adopt new behaviors, and de-implement old behaviors based on the accumulating evidence on the effectiveness of interventions. Quoting the IOM: "When care does not match knowledge, it may fail to help – either by omission (failing to do what would help) or by waste (doing what cannot help).[2] For example, a client complaining of acute low back pain since a week without any limitations for spontaneous recovery, should be encouraged to trust on spontaneous recovery instead of providing unnecessary care.[25] The professional profile of physical therapists describes the concept of evidence-based practice as "the conscientious, explicit, and judicious use of

current best evidence in making decisions jointly with the client". It requires professional expertise and advanced clinical reasoning and communication skills to integrate the best available evidence, client needs and preferences and professional expertise in clinical decision-making.[26] Clinical practice guidelines provide the best available evidence on clinical problems to support the process of clinical reasoning and decision-making. However, guidelines are not available for each clinical problem, as in fact guideline availability is scarce, and if available, the context of the clinical problem might not be appropriate to apply the guideline, or patient preferences might conflict guideline recommendations.[6] Therefore, clinical decision-making often happens in the context of uncertainty, as no single best solution to a problem exists. Variation in clinical practice is therefore, to a certain extent, inevitable. However, when guidelines are both available and relevant regarding the client's problem, research on guideline implementation shows that their use in clinical practice is limited. The main bottlenecks for healthcare professionals are attributable to limited guideline knowledge, negative attitudes toward guidelines, and limited social and organizational support.[27–32] In addition, a study of Rutten *et al*.[33] on determinants of guideline adherence showed that physical therapists in the Netherlands do not hold realistic perceptions of their use of guidelines in clinical practice. In sum, evidence based practice in physical therapy shows room for improvement.

Taking the assessor perspective: was the service provided to Eric evidence-based?

According to the guideline on Complaints of the Neck, Arm and Shoulder (CANS) which was published in 2010 and applies to Eric's complaints, this health problem can be classified as nonspecific, meaning that the exact etiology is unknown.[34] According to the literature, multiple factors may have contributed to the onset and the development of his complaints: work related factors such as task content, working conditions, terms of employment, social relationships, and personal factors such as an ineffective approach to stress management. Both physical therapists who treated Eric could not rely on sound evidence on the diagnosis or treatment of Eric's health problem; they needed to deal with reasoning in the context of uncertainty.[26,35,36] The guideline on CANS argues for training programs that explicitly address physical, mental and contextual barriers for recovery, but does not recommend a specific intervention.

Although the approach of both physical therapists differs substantially, they both share their focus on physical functioning (muscle tension, muscle strength, muscle endurance, working posture) and working conditions (working place) consistent with guideline recommendations. However, personal factors – such as cognitions (beliefs, expectations), emotions (frustration) that might relate to the onset and the development of the health problem, were not explicitly addressed even though the literature is conclusive on the importance of involving these factors.[34,37] Contextual factors were addressed by advising working place adjustments and by empowering self-regulation of the working load. However, multi-professional collaboration with relevant healthcare providers might have prevented inadequate workplace solutions.[38]

The second physical therapist addressed behavioral factors (limited coping with pain in computer related activities) by introducing a graded activity program (a behavioral oriented approach) which has shown to be effective and is in line with guideline recommendations for clients with persisting CANS complaints. However, Eric lacked intrinsic motivation to adhere to the program. Motivational interviewing techniques might have identified hindering factors to program compliance, allowing to tailor the program to the stages of behavioral change.[39–41]

Transparent healthcare

To date, clients do not only depend on their caregiver as information resource. Via the Internet, they have access to all kinds of information sources, reliable or not. Physical therapists need communicate effectively and share information to validate this information if necessary and to allow for shared decision-making.[42] Adequate condition-specific information flow may also facilitate client's self-management.[37,43]

According to the view of the IOM, transparency also involves information describing the healthcare system's performance (including healthcare professionals and organizations), for example its cost-effectiveness.[3] Providing transparency requires routine data sampling of the process and outcomes of physical therapy services such as electronic health records, clinical performance measurement and PROM results.

However, due to increasing demands of health insurers, physical therapists resist extensive record keeping and performance and outcome measurement. They view record keeping as a dispropor-

tionate burden compared to actual client care, challenging their professional identity. Attempting to meet the quality criteria of the external audits conducted by insurance companies, they focus on the completeness of the record rather than providing accountability of their clinical reasoning and decision-making process. This lack of balance between the interests of healthcare providers and health insurers has been acknowledged by the Ministry of Health.[44] Research showed that external regulations – such as by health insurers – can potentially be effective, but the evidence is not convincing regarding the sustainability of the results and the strategy might induce unwanted consequences such as under-treatment of clients with multi-morbidity or disparities in healthcare delivery.[2,45]
Technology supports the transparency of physical therapy care; electronic client records enable sampling process and outcome data which can be used to inform clients, and – on an aggregate level – provide data that account for the effectiveness of physical therapy services. However, routine data entry in electronic health records including PROM data is still in the early stages of development. A study of Meerhoff *et al*.[24] showed that the data sampled from electronic health records in a national registry were usable for internal quality improvement purposes, but not robust enough for accountability purposes. In sum, improvement of the transparency of physical therapy services is desired.

Taking the assessor perspective: was the care provided to Eric transparent?
Considering the adequacy of client-information, Eric's knowledge on his health problem was not explicitly addressed and information on condition-specific interventions – such as the guiding principles of a behavior oriented approach – was poorly provided allowing for false process and outcome expectancies.
Regarding performance and outcome measurement, the guideline on CANS recommends the Patient Specific Complaints Questionnaire (PSC) which focuses on perceived limitations in daily activities. Instead the Visual Analogue Pain Rating Scale (VAS) was used which focuses on pain, suggesting that pain reduction is the primary aim which is not recommended for clients with persisting CANS complaints. The guideline recommends interventions aiming for self-management, and measurement instruments should be consistent with this approach. Timely in-between measurement on the level of activities and participation by the PSC or a comparable PROM, might have changed the focus on pain and prevented ineffective care or waste.

Assessment of professional performance

Assessment of professional performance is applied with different aims, by different authorities, representing different interests. In the next sections we explore the aims of assessment, the assessors and their interest in professional performance.

Aims of professional performance assessment

Assessment of professional performance, including both clinical and organizational performance, can be applied for summative or formative purposes. Summative assessments are used to decide on academic progress, certification or accreditation. Formative assessments are used to support continuous learning and quality improvement.[46] On an individual level, formative results – conceived as feedback – can be used to identify gaps in actual performance and to inform the process of developing new knowledge, skills and attitudes.[47] Teams and organizations can use the results to evaluate their goals and to benchmark their output.[48] Irrespective of the purpose of performance assessment, the outcomes should be relevant to advance healthcare quality, either by improving undergraduate health professions education, or post-graduate professional development.

Actors in the assessment of the quality of physical therapy in the Netherlands

The Dutch government holds a register (BIG registry) focusing on certification and relevant expertise of healthcare providers.[49] Professional organizations of physical therapists hold a quality register, establishing minimum standards of performance, based on professional development activities.[50] Health insurance companies conduct a more comprehensive performance assessment system. The effectiveness of physical therapy is assessed by sampling process and outcome data. Feedback is provided by comparing the data to a benchmark on treatment session averages (treatment index) and client satisfaction (Consumer Quality Index, CQI). In addition, health insurers conduct audits focusing on both clinical and organizational performance. They use checklists to review electronic client records. These checklists trigger professionals to take a superficial and reductionist approach to quality improvement by striving for the completeness of their electronic health records, rather than account for the quality of their clinical reasoning and decision-making process. Research showed that assessment has a powerful impact on learning and this impact may be positive or negative.[51–53] Physical therapists question the validity of the audit

results as they are perceived as poorly reflecting their authentic clinical practice. Although the literature shows that external audits can potentially be effective, the evidence is not convincing regarding the sustainability of the results and the strategy might induce unwanted consequences such as under-treatment of clients with multi-morbidity or disparities in healthcare delivery,[2,45] such as preferences for accepting clients with health conditions requiring short interventions. A study of Scholte *et al.*[54] on the impact of a Dutch performance feedback system for physical therapists, based on indicator scores extracted directly from electronic health records, showed that financial incentives by health insurers negatively affected the use of feedback reports for quality improvement. A lack of 'belief' in the quality improvement system and 'distrust' among physical therapists towards health insurers were the major barriers to implementation. Feedback provided by a feedback source perceived as trustworthy, might be more effective.[4,55–57] In short, there is a need for performance assessment formats providing feedback that is perceived as meaningful and supportive in guiding sustainable quality improvement.

Quality criteria for professional performance assessment

Professional performance can be assessed by a variety of assessment formats that can be globally distinguished in standardized and non-standardized assessment. In standardized assessment, all assessment conditions are as much standardized as possible for all test takers such as assessment with clinical vignettes, standardized patients or to a lesser extent role-play.[10,47] When assessing real practice, standardization is impossible.[9] Real practice can be assessed by artefacts of professional performance such as electronic health records, video-recordings of real-life behaviors, or real-time observation. Classical quality criteria for clinical performance assessment are validity and reliability.[58] Validity refers to whether an instrument actually measures what it is purported to. Reliability refers to the reproducibility of the results. Inferences made on an assessment lacking validity and reliability may cause substantial harm to the assessed, especially when the stakes are high.[9]

When performance assessment is used as a tool for learning (formative assessment), its reliability and validity depend on the quality of the feedback provided and its impact on learning and improvement.[52,59,60] High quality formative assessment should produce high quality feedback guiding the process of continuous learning and improvement towards its intended goals. That poses substantial demands on the skills and attitudes of the performance

assessor.[10,61] Assessment of clinical performance addresses complex competencies related to complex behaviours. For example, diagnosing a clinical problem requires adequate communication skills, clinical examination skills, and the integration of different knowledge resources related to personal experience, available evidence, and client-related information.[9,47,62] Clinical reasoning is the cognitive process that guides the decision-making process, and is critical to the quality and safety of physical therapy care. Therefore, valid assessment of clinical reasoning and decision-making needs professional judgment. However, given the notion that clinical decisions are often made in situations of uncertainty about the correct diagnosis or best intervention as protocols and clinical practice guidelines are scarce, it is obvious that professionals may have different views on the best solution to a clinical problem. Therefore, bias is a natural given in the assessment of clinical performance. To reduce bias and increase the reliability and validity of the feedback provided, multiple views are needed on the observed performance as many pairs of eyes see more than one.[9] Multiple professional views may be presented by experts, peers or the self (self-assessment).[10,63]

Looking at the quality of feedback, the literature shows that the effectiveness of feedback depends on factors related to the feedback provider, the type of feedback, the feedback receiver, and the context in which feedback is provided.[64,65]

A systematic review of Ivers *et al*.[4] on the effects of feedback on professional practice and patient outcomes showed that feedback may be more effective when the source is a supervisor or colleague, it is provided more than once, it is delivered in both verbal and written formats, and when it includes specific and measurable goals and an action plan. Regarding the feedback receiver, research showed that professionals struggle with accepting feedback when it is incongruent with their self-assessment or threatens their self-esteem.[66] Its acceptability improves when feedback is provided in an environment of trust and mutual respect, provided in a neutral non-judgmental style.[55,56,67] Feedback is likely to be rejected when the provider is not perceived to be a credible and trustworthy source of information[65,68] or when it conflicts with personal or group norms and values.[69,70] In sum, when unstandardized performance assessment is used, the process of providing, receiving, and using feedback for quality improvement needs scaffolding.[10,57,71]

Aim of the thesis

This introduction showed that the client-centeredness, effectiveness, and transparency of physical therapy care needs improvement to adequately respond to current and future client needs, societal and political demands. The feedback provided by external authorities to guide the quality improvement process is lacking perceived validity and acceptability resulting in limited feedback use. Both validity and acceptability might improve when physical therapists would take the assessor role themselves in evaluating the quality of their clinical and organizational performance and provide each other of useful and acceptable improvement feedback.[11] Meanwhile, they might critically reflect on their own performance.
This thesis aims to explore the utility of educational programs aiming to advance the quality of physical therapy care. Central to the distinct programs is performance assessment as a feedback tool to stimulate professional development and behavior change in clinical practice. Although the educational program designs vary according to the targeted quality improvement domain, performance feedback is built from multiple assessor perspectives. Performance assessment includes the assessment of clinical reasoning, clinical skills, and organizational performance (clinical audit). Our research questions address:

1 How do physical therapists perceive interventions, based on performance feedback, aiming to advance the quality of physical therapy care?
2 What is the impact of interventions, based on performance feedback, on learning and professional behavior change?

Outline of the thesis

To answer these research questions we conducted seven studies addressing quality improvement programs using standardized and non-standardized performance assessment in both undergraduate education as post-graduate professional development.
In *chapter 2*, we describe a mixed methods study evaluating the impact of a peer assessment design on the improvement of clinical performance in undergraduate physical therapy education. *Chapter 3* presents an cluster randomized controlled trial in professional physical therapy practice comparing the effectiveness of peer assessment as an educational strategy to enhance adherence to

[3] Complaints of the Arm, Neck and Shoulder.

a low back pain guideline compared to case-based discussion. The critical success features of this intervention design are evaluated in a mixed methods study which is described in *chapter 4*. We used the results of the latter study to improve the peer assessment design and tested its effectiveness on the implementation of the guideline on CANS[3] in a cluster-randomized controlled trial in professional practice which is presented in *chapter 5*. Based on our research on peer assessment we designed a quality improvement system including both peer assessment and clinical audit aiming at self-regulated quality improvement. The development of this quality improvement system and its feasibility to self-regulate the quality of physical therapy services are evaluated in a mixed methods study which is presented in *chapter 6*. We used the results of the feasibility study to improve the quality system and tested its effectiveness on the improvement of the client-centeredness, effectiveness, and transparency of physical therapy services in four pilots with networks of physical therapists. This study is described in *chapter 7*. In *chapter 8* the development of a new performance assessment design is introduced including the evaluation of its validity for quality improvement purposes. Finally, in *chapter 9* the findings of the studies in this thesis are discussed with a critical reflection on the program design and implementation features and the consequences for ongoing program development and quality improvement in clinical practice.

Since this thesis is based on published journal articles, some overlap will be inevitable.

[3] Complaints of the Arm, Neck and Shoulder.

References

1 de Vries C, Hagenaars L, Kiers H, Schmitt M. KNGF Beroepsprofiel Fysiotherapeut. https://www.kngf.nl/vakgebied/vakinhoud/beroepsprofielen.html. Published 2014. Accessed June 1, 2017.

2 Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press; 2001.

3 Porter ME. What is value in health care? *N Engl J Med.* 2010;363(26):2477-2481.

4 Ivers N, Jamtvedt G, Flottorp S, *et al.* Audit and feedback: effects on professional practice and health care outcomes (Review). *Cochrane Database Syst Rev.* 2012;(7):1-227.

5 Dwamena F, Holmes-Rovne M, Gaulden C, *et al*. Interventions for providers to promote a patient-centered approach in clinical consultations (Cochrane Review). *Cochrane Libr.* 2012;(12):1-177.

6 Grol RP, Wensing M, Eccles MP, Davis DA, (Eds). *Improving Patient Care: The Implementation of Change in Health Care.* 2nd ed. Chichester, West Sussex: John Wiley & Sons, Inc.; 2013.

7 Masterplan Kwaliteit in Beweging Koninklijk Genootschap voor Fysiotherapie https://www.kngf.nl/vereniging/Programmas+en+projecten/mkib.html. Published 2015. Accessed December 27, 2016.

8 de Vijlder F. Professionals aan het roer. 2015. http://www.kwaliteitvanonderwijs.nl/wp-content/uploads/2015/07/De-echte-dingen-essays-over-de-kwaliteit-van-onderwijs-e-book.pdf. Accessed June 1 2017.

9 van der Vleuten CP, Schuwirth LWT. Assessing professional competence: From methods to programmes. *Med Educ.* 2005;39(3):309-317.

10 van der Vleuten CP, Sluijsmans DM, Joosten-ten Brinke D. Competence assessment as learner support in education. In: Mulder M, ed. *Competence-Based Vocational and Professional Education*. 1st ed. Springer International Publishing AG; 2017:607-630.

11 Pronovost PJ, Hudson DW. Improving healthcare quality through organisational peer-to-peer assessment: lessons from the nuclear power industry. *BMJ Qual Saf.* 2012;21(10):872-875.

12 Butterfield R, McCormick B, Anderson R, Ball J, White J, Eleftheriades C. *Quality of NHS Care and External Pathway Peer Review.*; 2012. https://www.chseo.org.uk/downloads/report3-peerreview.pdf. Accessed June 1 2017.

13 Huber M. How should we define health? *Br Med J.* 2011;(343).

14 O'Keefe M, Henderson A, Chick R. Defining a set of common interprofessional learning competencies for health profession students. *Med Teach*. 2017;5:463-468.

15 Paans W, Wijkamp I, Wiltens E, Wolfensberger MV. What constitutes an excellent allied health care professional? A multidisciplinary focus group study. *J Multidiscip Healthc* 2013;6:347-356.

16 Dierckx K, Deveugele M, Roosen P, *et al*. Implementation of shared decision making in physical therapy: observed level of involvement and patient preference. *Phys Ther.* 2013;93(10):1321-1330.

17 Schoeb V, Bürge E. Perceptions of patients and physiotherapists on patient participation: a narrative synthesis of qualitative studies. *Physiother Res Int*. 2012;17(2):80-91.

18 van der Wees PJ, Nijhuis-van der Sanden MW, Ananian JZ, Black N, Westert GP, Schneider EC. Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries. Milbank Q. 2015;93(4):788-825.

19 Stevens A, Beurskens A, Köke A, van der Weijden T. The use of patient-specific measurement instruments in the process of goal-setting: a systematic review of available instruments and their feasibility. Clin Rehabil. 2013;27(11):1005-1019.

20 Van Dulmen SA, Van Der Wees PJ, Staal JB, Braspenning JB, Nijhuis-Van Der Sanden MWG. Patient reported outcome measures (PROMs) for goalsetting and outcome measurement in primary care physiotherapy, an explorative field study. *Physiotherapy*. 2017;103(1):66-72.

21 Reuben DB, Tinetti ME. Goal-oriented patient care – an alternative health outcomes paradigm. *N Engl J Med.* 2010;363(1):1-3.

22 Stevens A, Moser A, Köke A, van der Weijden T, Beurskens A. The patient's perspective of the feasibility of a patient-specific instrument in physiotherapy goal setting: A qualitative study. *Patient Prefer Adherence*. 2016;(10):425-434.
23 Swinkels RAHM, Meerhoff GM, Custers JWH, van Peppen RPS, Beurskens AJHM, Wittink H. Using outcome measures in daily practice: Development and evaluation of an implementation strategy for physiotherapists in the Netherlands. *Physiother Canada*. 2015;67(4):357-364.
24 Meerhoff GA, van Dulmen SA, Maas MJM, Heijblom K, Nijhuis-van der Sanden MWG, van der Wees PJ. Development and evaluation of an implementation strategy for collecting data in a national registry and the use of patient-reported outcome measures (PROMs) in physical therapis practice: quality improvement study. *Phys Ther*. 2017;97:1-15.
25 Staal BJ, Hendriks EJ, Heijmans M, *et al*. KNGF-richtlijn Lage rugpijn. Konkinklijk Genootschap voor Fysiotherapie. http://www.fysionet-evidencebased.nl/index.php/component/kngf/richtlijnen. Gepubliceerd 2013. Accessed June 1, 2017.
26 Higgs J, Jones MA, Loftus S, Christensen N. *Clinical Reasoning in the Health Professions*. Third edit. Philadelphia: Elsevier Health Sciences; 2008.
27 Rutten GMJ, Harting J, Rutten STJ, Bekkering GE, Kremers SPJ. Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *J Eval Clin Pract*. 2006;12(5):491-500.
28 Bekkering GE, Tulder MW Van, Hendriks EJ, *et al*. Implementation of clinical guidelines on physical therapy for patients with low back pain : randomized trial comparing patient outcomes after a standard and active implementation strategy. *Phys Ther*. 2005;85(6):544-555.
29 Grimshaw J, Thomas R, Maclennan G, *et al*. Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technol Assess*. 2004;8(6).
30 Grol R, Wensing M, Bosch M, Hulscher M, Eccles M. Theorieën over implementatie. In: *Implementatie: Effectieve Verbetering van de Patiëntenzorg*. 4e ed. Amsterdam: Reed Business; 2011:43-68.
31 Swinkels ICS, van den Ende CHM, van den Bosch W, Dekker J, Wimmers RH. Physiotherapy management of low back pain: Does practice match the Dutch guidelines? *Aust J Physiother*. 2005;51(1):35-41.
32 Dannapfel P, Peolsson A, Nilsen P. What supports physiotherapists' use of research in clinical practice? A qualitative study in Sweden. *Implement Sci*. 2013;8:31.
33 Rutten GM, Kremers S, Rutten ST, Harting J. A theory-based cross-sectional survey demonstrated the important role of awareness in guideline implementation. *J Clin Epidemiol*. 2009;62(2):167-176.
34 Heemskerk MA, Staal JB, Bierma-Zeinstra SM, *et al*. KNGF-richtlijn Klachten aan de arm, nek en/of schouder (KANS).Published 2010. https://www.fysionet-evidencebased.nl/index.php/richtlijnen/richtlijnen/klachten-aan-de-arm-nek-enof-schouder-kans. Accessed June 1 2017.
35 Ajjawi R, Higgs J. Learning to reason: a journey of professional socialisation. *Adv Health Sci Educ Theory Pract*. 2008;13(2):133-150.
36 Ramaekers S, Kremer W, Pilot A, van Beukelen P, van Keulen H. Assessment of competence in clinical reasoning and decision-making under uncertainty: the script concordance test method. *Assess Eval High Educ*. 2010;35(6):661-673.
37 Hutting N, Staal JB, Heerkens YF, Engels JA, Nijhuis-van der Sanden MW. A self-management program for employees with complaints of the arm, neck, or shoulder (CANS): study protocol for a randomized controlled trial. *Trials*. 2013;14:1.
38 Heemskerk M, Staal JB, Bierma-Zeinstra S, *et al*. KNGF-richtlijn Klachten aan de arm, nek en/of schouder (KANS) [KNGF-guideline complaints of arm, neck and/or shoulder (CANS). Published 2010. http://www.fysionet-evidencebased.nl/index.php/component/kngf/richtlijnen. Accessed June 1 2017.
39 DiClemente CC, Velasques MM. Motivational interviewing and the stages of change. In: *Motivational Interviewing: Preparing People for Change*. 2nd ed. ; 2002:201-216.
40 de Vries N. *Managing the decline: Physical therapy in frail elderly*. Radboud Repos. 2015.

41 Ryan R, Deci E. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol.* 2000;55(1):68-78.
42 Institute of Medicin. *Improving Diagnosis in Health Care.* (Balogh EP, Miller BT, Ball JR, eds.). National Academies Press; 2015.
43 Hutting N, Heerkens YF, Engels JA, Staal JB, Nijhuis-van der Sanden MW. Experiences of employees with arm, neck and shoulder complaints: a focus group study. *BMC Musculoskelet Disord.* 2014;141(15):1-13.
44 Schippers E. Speech van de minister van VWS bij het Jaarcongres van het Koninklijk Nederlands Genootschap Fysiotherapie op de Dag van de Fysiotherapeut in Utrecht. 2015. https://www.rijksoverheid.nl/regering/inhoud/bewindspersonen/edith-schippers/documenten/toespraken/2015/11/06/speech-van-de-minister-van-vws-edith-schippers-bij-het-jaarcongres-van-het-koninklijk-nederlands-genootschap-fysiotherapie-op-de-dag-van-de-fysio. Accessed June 1 2017.
45 Eijkenaar F. Pay-for-performance for healthcare providers. Design, performance measurement, and (unintended) effects. *Health Policy* (New York). 2013;110(2-3):115-130.
46 Black P, Wiliam D. Developing the theory of formative assessment. *Educ Assessment, Eval Account.* 2009;21(1):5-31.
47 Epstein RM. Assessment in medical education. *N Engl J Med.* 2007;356(4):387-396.
48 Sessa V, Valerie I LM. *Continuous Learning in Organizations. Individual, Group, and Oranan izational Perspectives.* 1st ed. Mahwah, New Jersey: Lawrence Erlbaum; 2006.
49 Ministerie van Volksgezondheid en Welzijn. BIG register https://www.bigregister.nl/. Accessed February 27, 2016.
50 Koninklijk Genootschap van Fysiotherapeuten. Centraal Kwaliteitsregister Fysiotherapie. https://www.kngf.nl/vakgebied/kwaliteit/ckr.html. Published 2016. Accessed February 27, 2016.
51 Struyven K, Dochy F, Janssens S, Schelfhou W, Gielen S. The overall effects of end-of-course assessment on student performance: A comparison between multiple choice testing, peer assessment, case-based assessment and portfolio assessment. *Stud Educ Eval.* 2006;32(3):202-222.
52 Dochy F, Segers M, Gijbels D, Struyven K. Assessment engineering: Breaking down barriers between teaching and learning, and assessment. In: Boud D, Falchikov N, eds. *Rethinking Assessment in Higher Education: Learning for the Longer Term.* 2007:87-100.
53 Cilliers F, Schuwirth L, Herman N, Adendorff H, van der Vleuten C. A model of the pre-assessment learning effects of summative assessment in medical education. *Adv Heal Sci Educ Theory Pract.* 2012;17(1):39-53.
54 Scholte M, Neeleman-Van Der Steen CW, Van Der Wees PJ, Nijhuis-Van Der Sanden MWG, Braspenning J. The reasons behind the (non)use of feedback reports for quality improvement in physical therapy: A mixed-method study. *PLoS One.* 2016;11(8):1-16.
55 Hysong SJ, Kell HJ, Petersen LA, Campbell BA, Trautner BW. Theory-based and evidence-based design of audit and feedback programmes: examples from two clinical intervention studies. *BMJ Qual Saf.* 2016;0(6):1-12.
56 Payne VL, Hysong SJ. Model depicting aspects of audit and feedback that impact physicians' acceptance of clinical performance feedback. *BMC Health Serv Res.* 2016;16(1).
57 Sargeant JM, Lockyer J, Mann K, *et al*. Facilitated reflective performance feedback: developing an evidence- and theory-based model that builds relationship, explores reactions and content, and coaches for performance change (R2C2). *Acad Med.* 2015;90(12):1698-1706.
58 van Berkel H, Bax A, Joosten-ten Brinke. (Eds). *Toetsen in Het Hoger Onderwijs*. Springer; 2017.
59 Boud D. Assessment and learning : contradictory or complementary? In: Page K, ed. *Assessment for Learning in Higher Education*. London; 1995:35-48.
60 Falchikov N. *Improving assessment through student involvement: practical solutions for aiding learning in higher and further education*. 2nd ed. New York: Routledge Falmer; 2013.
61 Govaerts MJ, Schuwirth LW, Van der Vleuten CP, Muijtjens AM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract.* 2011;16(2):151-165.

62 Wass V, Van der Vleuten CP, Shatzer J, Jones R. Assessment of clinical competence. *Lancet*. 2001;357:945-949.
63 Charlin B, Gagnon R, Pelletier J, *et al.* Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ*. 2006;40(9):848-854.
64 Mann K, van der Vleuten CP, Eva KW, *et al*. Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. *Acad Med*. 2011;86(9):1120-1127.
65 Eva KW, Armson H, Holmboe E, *et al*. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ Theory Pract.* 2012;17:15-26.
66 Regehr G, Eva KW. Self-assessment, self-direction, and the self-regulating professional. *Clin Orthop Relat Res*. 2006;449:34-38.
67 Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. *Med Care.* 2009;47(3):356-363.
68 Sargeant J, Eva KW, Armson H, *et al*. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ.* 2011;45(6):636-647.
69 Ajzen I. Nature and operation of attitudes. *Annu Rev Psychol.* 2001;52:27-58.
70 Prochaska JO, Redding CA, Evers KE. Health behavior and health education. In: Glanz K, Rimer BK, Viswanath K, eds. *Health behavior and health education: theory, research, and practice*. 4th ed. Wiley & Sons; 2008:97-121.
71 Finn GM, Garner J. Twelve tips for implementing a successful peer assessment. *Med Teach.* 2011;33(6):443-446.

Chapter 2
Why peer assessment helps to improve clinical performance in undergraduate physical therapy education: a mixed methods design

Marjo Maas
Dominique Sluijsmans
Philip van der Wees
Yvonne Heerkens
Cees van der Vleuten
Ria Nijhuis-van der Sanden

BMC Medical Education, 2014;14(1):117

Abstract

Background

Peer Assessment (PA) in health professions education encourages students to develop a critical attitude towards their own and their peers' performance. We designed a PA task to assess students' clinical skills (including reasoning, communication, physical examination and treatment skills) in a role-play that simulated physical therapy (PT) practice.Students alternately performed in the role of PT, assessor, and patient. Oral face-to-face feedback was provided as well as written feedback and scores.
This study aims to explore the impact of PA on the improvement of clinical performance of undergraduate PT students.

Methods

The PA task was analyzed and decomposed into task elements.
A qualitative approach was used to explore students' perceptions of the task and the task elements. Semi-structured interviews with second year students were conducted to explore the perceived impact of these task elements on performance improvement. Students were asked to select the elements perceived valuable, to rank them from highest to lowest learning value, and to motivate their choices. Interviews were transcribed verbatim and analyzed, using a phenomenographical approach and following template analysis guidelines. A quantitative approach was used to describe the ranking results.

Results

Quantitative analyses showed that the perceived impact on learning varied widely. Performing the clinical task in the PT role, was assigned to the first place (1), followed by receiving expert feedback (2), and observing peer performance (3). Receiving peer feedback was not perceived the most powerful task element.
Qualitative analyses resulted in three emerging themes: pre-performance, true-performance, and post-performance triggers for improvement. Each theme contained three categories: learning activities, outcomes, and conditions for learning.
Intended learning activities were reported, such as transferring prior learning to a new application context and unintended learning activities, such as modelling a peer's performance. Outcomes related to increased self-confidence, insight in performance standards and awareness of improvement areas. Conditions for learning referred to the quality of peer feedback.

Conclusions

PA may be a powerful tool to improve clinical performance, although peer feedback is not perceived the most powerful element. Peer assessors in undergraduate PT education use idiosyncratic strategies to assess their peers' performance.

Background

Modern education in health professions aims at the development of reflective practitioners, capable of self-directing their professional development before and after graduation. Healthcare practitioners need to keep up with demands for improved quality of care and patient outcomes.[1] Peer Review is one of the strategies that healthcare practitioners apply for professional development, for upholding professional standards and to be accountable to stakeholders in healthcare.[2] Peer Assessment (PA) is a structured variant of Peer Review that can be described as the process whereby participants of similar status evaluate the performance of their peers and give quantitative and/or qualitative feedback. The strategy targets the development of a mutual accepted quality standard of performance by introducing peers with the 'assessor' or 'auditor' perspective. The PA approach implies that professional development is a shared responsibility and that individuals, teams and organizations may profit from the learning outcomes.[3] PA has become increasingly popular in health professions educational programs to encourage students to develop a critical attitude towards their own and their peers performance, anticipating on lifelong quality improvement demands in clinical practice. A study of Sluijsmans[4] showed that students in higher education, who are trained to critically reflect on the performances of their peers, simultaneously develop self-assessment skills that might help them to self-direct their learning process. Research showed that healthcare professionals have a limited ability to accurately self-assess their level of professional competence.[5] Self-assessment alone appears not to be a reliable source of information to identify shortcomings in clinical performance because practitioners tend to systematically over- or underestimate their level of competency.[6,7] The development of adequate self-perceptions requires additional information from external sources and comparing information with a performance standard.[8] Peers who are adequately trained in their peer assessor role, might provide the missing information to inform self-assessment and might uncover improvement areas that would remain undiscovered by self-assessment alone.[9]

PA in health professions education is applied with different educational goals and implemented in various educational formats.[10–12] Gielen[13] distinguishes two main goals of PA: PA as an 'assessment tool' and PA as a 'learning tool'. PA as an assessment tool refers to the ability of students to reliably and validly assess their peers. Most research on PA has conceived PA as an assessment tool. Peer judgment is either compared to faculty judgment or self-reports and the quality of PA is determined by a criterion validity approach.[12,14–17] This concept of PA is not applicable when PA is intended to inform self-assessment and improve performance. When PA is viewed as a 'learning tool' it aims to provide students with relevant improvement feedback.[18] In contrast with staff assessment, peer feedback is built up from multiple sources of information.[19] The quality criterion for PA as a learning tool can best be described by the concept of 'consequential validity', referring to the impact on student learning outcomes.[13,20–22] The majority of studies on the impact of PA on learning in health professions education report positive effects.[12] These studies however mainly focus on professional behavior such as rule-based adherence to behavioral norms, rather than (hands-on) clinical examination and treatment skills.[15,23–26] When it comes down to PA of clinical performance, validity evidence is scarce and limited to the medical domain.[12,16,27–29] However, diagnosis and treatment belong to the core business of healthcare practitioners and performance gaps might affect patient safety and intervention outcomes.[1] The implementation of PA of clinical performance in undergraduate health professions education is therefore desired. Research showed that one of the determinants of effective PA processes is training in PA skills.[10,11] When students are trained to adequately assess their peers and to provide meaningful improvement feedback, they might be well prepared to 'audit' their colleagues after graduation. Yet, we do not know how PA impacts on the improvement of clinical performance and validity evidence is needed.

We designed a complex PA task that aims to facilitate students to improve their clinical performance prior to work placement. Clinical performance included reasoning skills, communication skills and practical physical examination – and treatment skills. A mixed methods approach was taken to analyze the following research questions.

How does the PA task impact on the improvement of clinical performance in the perception of PT students?

1 Which elements of the PA task have a powerful impact on learning and what are factors conditional for learning?
2 Why do students perceive these task elements as powerful?

Methods

Study design

A qualitative approach was used to explore students' perceptions of the PA task and the distinct PA task elements. A quantitative approach was used to identify the elements that have the strongest impact on learning to strengthen the qualitative data.

Context and participants

This study was conducted within the Department of Physical Therapy at the HAN university of Applied Sciences in the Netherlands in 2008. The PA task was part of the course 'Physical therapy in primary care-2' that was offered in the second year of the bachelor program, prior to work placement. The course consisted of two blocks of seven weeks. Participation in the PA task was compulsory, but the use of the PA results was formative. Ten groups of twelve students completed the task (n=120). A purposive sample of 12 students was invited for interviews. Sampling was based on maximal variation in groups, gender and nationality.

The design of the PA task

The PA task was designed as an authentic, complex learning task. Performance of clinical skills was observed and evaluated by peers in a role play that simulated physical therapy practice. The task was pre-tested after the first block of seven weeks and evaluated in a pilot study including student interviews.
In the PA sessions, students alternately performed in three roles: physical therapist (PT) role, assessor role, and patient role. At the beginning of the role play, each group member received a short written clinical case. The simulated patient received an additional role description. In the PT role, students demonstrated relevant examination or intervention skills. In the assessor role, students provided immediate face-to-face oral feedback, written feedback and scores. In the patient role, students simulated the written clinical cases according to the role description and provided feedback afterwards. Table 1 shows the task procedure. Expert assessors (teachers) took part in the PA session in the role of end assessor, providing additional feedback if necessary and only when all peer feedback had been collected. Students were provided with a manual

Table 1 — Peer Assessment Task Procedure

Time	Task	Therapist role	Patient role	Assessor role
5 min.	Study written clinical case and clinical assignment	x	x	x
	Study simulation role information		x	
3-5 min.	Explain choice for intended examination or treatment	x		
8-10 min.	Perform examination – or treatment task	x		
3-5 min.	Fill out assessment form			x
4-5 min.	Provide oral improvement feedback		x	x
	Comment on feedback	x		
25-30 min				

that allowed them to prepare the task in advance. It contained the learning goals of PA, a structured task procedure and a set of short clinical cases, according to the *key-feature* concept.[30] These cases served as PA material to enhance the transfer of knowledge and skills to new problems[31]; students could choose to study them in advance or not. The manual also provided an assessment form, consisting of four global performance indicators that could be scored on a 7-point Likert scale and an open field for written comments. The form was validated in a previous study[32] and adapted to PA. Students were instructed in giving high-quality improvement feedback during the pilot and feedback guidelines were included in the manual.

Each peer group consisted of 6-7 students, which has showed to be an effective size for this purpose.[15,33] The task was presented prior to the final summative assessment of clinical performance that was decisive for the entrance of work placement. When the task was completed, students wrote a reflection report, using the PA feedback and scores. The reflection report served as participation evidence in their portfolio.

Data collection

The PA task was analyzed and decomposed by the method of Janssen-Noordman[34] to identify constituent task elements that might trigger improvement. The analysis of the PA task in task elements was discussed by a team of five experts until consensus was reached, and was validated by 12 participating students in the pilot study. Task analysis resulted in 13 task elements (table 2).

A semi-structured interview guide was designed on the basis of

pilot study results. Interviews were conducted by the principal researcher (MM) while notes were taken by a research assistant (EL). Students were invited for interviews by purposive sampling, aiming at maximal variation in groups, gender and nationality. The distinct task elements of the PA-task were presented on separate cards at the beginning of the interview. Students were asked to select the elements perceived to have a powerful impact on performance improvement and to rank the selected elements from highest learning value (rank 1) to the lowest. Task elements that were not selected were left out of the ranking procedure. Subsequently, students motivated their choices. Interviews were audio-taped after informed consent was obtained of each participant, and were transcribed verbatim. Interviews lasted between 45 and 60 minutes. Data collection was ended when saturation was reached, meaning that additional sampling would not contribute to new findings.

Data analysis

Ranking results of each selected task element were entered in IBM SPSS Statistics 20.0. Ranking numbers were re-coded into scores, awarding the first rank with the highest score and, the last rank with the lowest score. Frequencies were described and sum scores were calculated for each task element.

Five transcripts were studied and relevant quotes were coded independently by MM and EL. PA task elements were used as a-priori categories (defined in advance) to structure the data in a way that research question 2 was directly addressed. We followed the method of Nigel Kings' template analysis[35] that showed to be an adequate method for this purpose. Codes were discussed until consensus was reached and a coding scheme was created. Subsequently all transcripts were analyzed by MM and EL. New themes emerged from the data by constant comparison of codes and categories.

A data matrix was constructed that crossed task elements (a-priori categories) with themes and categories that emerged from the data.[36] Finally, a conceptual model of how PA affects learning was designed that fully fitted the data. To enhance credibility, the analysis process was checked by a project consultant (JB) and member checking was carried out among all interviewed students.

Ethical aspects

This project received approval from the Faculty board of Han University of Applied Sciences. All students volunteered to participate

Table 2 — Ranking of task elements according to perceived impact on performance improvement

Task[1]	Constituent[1] Task	Student (S) S1	S2	S3	S4	S5	S6	S7	S8	S9	S10	S11	S12	S13	S14	n[2]	Sum	R[3]
Prepare Task	**Study manual**		9											4		2	13	**9**
	Study Cases	8	8	9	4		9	8	4		4	6		7		10	67	**5**
Perform in PT role	**Give Performance**	9	6	8	9	9	8	9	8	7	9	8	9	8		13	107	**1**
	Receive peer FB	5		6	6	7	3	3	5	5	7	7	7		8	12	69	**4**
	Receive expert FB	6		7	7	8	2	4	6	6	8	9	8	5	9	13	85	**2**
	Receive Patient FB									4						1	4	**12**
	Receive Score							2		3					4	3	9	**10**
Perform in assessor role	**Observe performance**	4	4	4	5	6	7	7	9	9	5	4	5		6	13	75	**3**
	Give oral FB		7	3		5	6	6	7	8		3	4		5	10	54	**6**
	Give written FB					4										1	4	**12**
	Give Score		3				5									2	8	**11**
Perform in patient role		7	5		8		4	5		2	3	5		6		9	45	**7**
Write reflection report				5		3	1	1		1	6		6	9	7	9	39	**8**

[1] The blue entries were presented as cards for ranking

[2] N of students that selected the task for ranking

[3] Final rank order

and indicated their willingness to participate by signing an approved consent form.

Results

Quantitative analyses showed that all 13 presented task elements were selected, assigned to 12 ranks (2 tasks on rank 12). Table 2 shows that the perceived learning value of distinct task elements varied widely among students. The majority of students perceived 'performance in the physical therapist role' as the most valuable task element (1), followed by receiving expert feedback (2), and observing peer performance (3). Receiving peer feedback was not perceived the most powerful task element. Twelve interviews were conducted representing all groups and two additional interviews were needed to reach data saturation.
Qualitative analyses resulted in three major themes that explained how the PA task impacts the improvement of clinical performance: 1) pre-performance triggers, referring to the anticipatory cognitive motivators related to students' perceptions of the learning environment that were conceived as *feed forward*, 2) true-performance triggers, referring to the vast array of inputs elicited by performing the task that can be conceived as *internal feedback*, and 3) post-performance triggers, referring to knowledge of performance and knowledge of results that was conceived as *external feedback*. Each theme contained three categories: a) learning activities, b) learning outcomes, and c) conditions for learning. The results are summarized in table 3.

Pre-performance triggers

Expectations and personal goals

Students had different expectations of PA that colored their views. The majority of students viewed the assessment as an appropriate training prior to their summative assessment or their future professional practice. However, some students had little confidence in the assessor qualities of their peers.

"My expectations were not that high, because the group in which I worked was not so good. Yes, my expectations of the first PA were confirmed and so were my expectations of the second assessment. I needed to show that I participated to meet portfolio demands, and I did, but I was not satisfied ... the feedback was superficial and I was glad that there was an expert."

Table 3 — Summary of learning activities, learning results and conditions for learning related to distinct task elements

Triggers	Task elements	Learning activities	Learning results	Learning conditions
Pre-performance triggers *Feedforward*	Study manual Study cases	Self-study Practice	Knowledge of performance standards Reduction of performance anxiety	
True-performance triggers *Internal feedback*	Perform PT role	Cope with anxiety triggers Apply learning in new context Reason aloud Act methodically	Increased self-confidence Awareness of improvement areas	
	Perform in patient role	Empathise with patient problem		
	Observe performance	Matching intended performance with observed performance Modelling	Re-design of intended performance Increased self-confidence Knowledge of alternative performance	
	Give oral feedback Give written feedback Give score	Study criteria Structure information Empathize with peer Explicit views	Insight in performance standards	
Post-performance triggers *External feedback*	Receive peer feedback	Ask for clarification Analyse information	Knowledge of performance from different perspectives Knowledge of alternative performance Awareness of improvement areas	Peer is well prepared and has sufficient case-specific knowledge. Feedback is critical, specific, concrete, reveals strength and weakness and contains improvement suggestions. Feedback meets learning needs. Peer is involved in learning process.
	Receive expert feedback		Knowledge of expert standards Validation of peer feedback	Expert allows for discussion over criteria.
	Receive patient feedback		Knowledge of patient perceived aspects	Sufficient case-specific knowledge role-player.
	Receive score	Compare sum scores and domain scores	Knowledge of results compared to the group	Peer has enough courage to give low scores when necessary.
Reflection	Write reflection report	Select feedback Relate information to prior feedback Create new learning goals		

Study manual and clinical cases

Students felt triggered to study cases prior to assessment. The cases were new and represented a sample of the context in which prior learning needed to be applied. Students were motivated either out of fear of encountering unexpected demands or simply out of curiosity.

"In the previous PA I was not prepared and I stood there a bit nervous, waiting for what would happen. Now I have prepared the cases, I know what I can do and that feels much better."

True-performance triggers

Perform in the physical therapist role

The transfer of learning into a new context was perceived as challenging. Students embraced the opportunities for new clinical encounters. They were triggered by both curiosity and eagerness. Firstly, the PA context differed from the learning context because their actions were being watched. They needed to cope with anxiety triggers common in this type of performance.

"... You have an assignment and 12 eyes are watching you. You must also be able to perform under that pressure. You do not want to blunder before your classmates."

Successful coping resulted in increased self-confidence.

"... When you're insecure, but you have to perform the task, and then it turns out that you performed well, then you feel strengthened. Like I'm not someone that knows nothing and that feels good."

Secondly, the presence of a (simulated) patient required transfer of knowledge and skills to the specific content of the patient problem and the specific patient needs. Organized domain specific knowledge needed to be combined with new, unexpected information.

"... yes, in class you practice without a patient. When your skills are good, you do not bother. In PA you have to deal with a patient."

Students were confronted with having little professional language available to explain to their peers what they were planning to do and why. They were triggered to reason aloud.

"… usually it is in your head, but now you have to argue aloud. Normally you don't explain why you choose for a certain clinical test, but when you're asked, yes, you need to answer."

Similar to clinical reasoning aloud, the majority of students felt triggered to transform declarative knowledge (knows) into procedural knowledge (knows how) and performance (shows how).

"… and you do it sequentially, not just in pieces as in class, but the whole thing, in which all those pieces have to be glued together…"

Although students perceived performance in the physical therapist role as the most valuable task element, data analyses showed that this learning experience cannot be separated from learning in other roles.

"Critical appraisal of a peer's performance, is easier said than done. When you perform in the assessor role, you actually act as a physical therapist."

Students either reflected or anticipated on their performance in the physical therapist role by continuous comparison of personal performance with peer performance and personal views with peer views as shown in the following paragraphs.

Perform in the assessor role

In observing their peers, students reported learning activities taking place on a more or less unconscious level as well as to learning activities which can be clearly described. The unconscious level refers to mirroring and matching the observed performance to the virtual image of one's own performance. The more conscious level refers to using the peer as a 'model' to improve their own clinical performance. Although student reports come across both levels, some efforts are made to make a distinction in learning activities:

Matching

"… you are very focused on looking at what someone is doing. You learn a lot by just watching. Actually, when you observe someone else, you imagine how you would act yourself. Like, how could I stand there, how would I do that?"

Modelling

"For example ... you see certain actions, you make notes, and you might try them out later."

When giving peer feedback, their personal picture was compared to peer views. Students were challenged to structure, summarize and communicate their (implicit) observations. Giving feedback prompted discussion, providing them a deeper understanding of performance criteria.

"I express how I see it, how I think it should be done and there again you get a reaction. Also a kind of clinical reasoning actually."

The peer assessor view appeared to develop during the assessment process and students became aware of their views by reasoning aloud.

"When I see what others are doing wrong, then I ask myself: 'how am I doing that? And what is good?' Then I'm going to ask the rest of the group: 'how do you do that? And do you agree with the way this person did it?'"

Post-performance triggers

Receive expert feedback

Being in the middle of the course, with the end-course assessment ahead, students wished to know what improvements they should make to meet the expert standards. Expert feedback contributed to the credibility of peer feedback that advanced acceptance of a peers' judgment or advice.

"... hard to say ... a piece of approval so to speak. I need to be sure what I have to improve. Expert feedback is a kind of confirmation of peer feedback."

Receive peer – and simulated patient feedback

Most students valued peer feedback because of its variety and completeness. Students who were reluctant in asking for feedback during the course, mentioned the advantage of this task to obtain exclusive feedback.

"Yeah, I'm pretty insecure. I like it when someone specifically looks at me, that I receive personal attention. I easily push myself on the side."

The involvement of students in the learning process of their peers was generally considered to be an advantage. Peers were able to keep a record of errors what expert assessors usually do not and that enhanced peer feedback credibility and acceptance.

"... yes, maybe it is like you usually practice together and they know you better; they know when you make mistakes by nervousness, they know your positive and negative sides. And if you're using the same group again in PA, they know what you had to learn."

Yet this was not an argument in favor of feedback being just nice; students agreed on conditions for learning from peer feedback. They reported that the acceptability and the usefulness of peer feedback heavily relied on appropriate task-specific knowledge, sufficient task preparation and enough peer assessor skills. In addition, feedback should be critical, revealing strengths and weaknesses and should contain improvement suggestions. Even judgmental feedback was mentioned. So-called 'soft feedback' consisting of global comments on communicational aspects, missing any connection to clinical performance, was widely rejected.

"Critical feedback. May also be judging. Empathy is important, but I do not like someone to just repeat what has been said with a very sweet voice."

Students however did not ask for the feedback they wanted in advance. Instead they complained afterwards of receiving feedback that did not meet their expectations.

"How I communicate with patients, that I know by now. Clinical reasoning, that is currently important to me. I want to know for myself why I do the things I do and I want to be able to explain that when anyone else asks me to."

Give and receive written feedback and scores

Elaborative oral feedback was preferred over written feedback and scores. Received scores were not perceived reliable, because peers lacked objectivity from an interpersonal perspective. For the same reason some students felt reluctant in giving scores. They, however, reflected on scores in a meaningful way by trying to find a certain convergence in the domains that needed improvement.

"Suppose I have e.g. 20 points and someone else has 30, so I'd always look for a category where the difference is. The final score does not tell me so much. Although it is good to have a score for each category to see where you still need to work on something."

Write reflection report

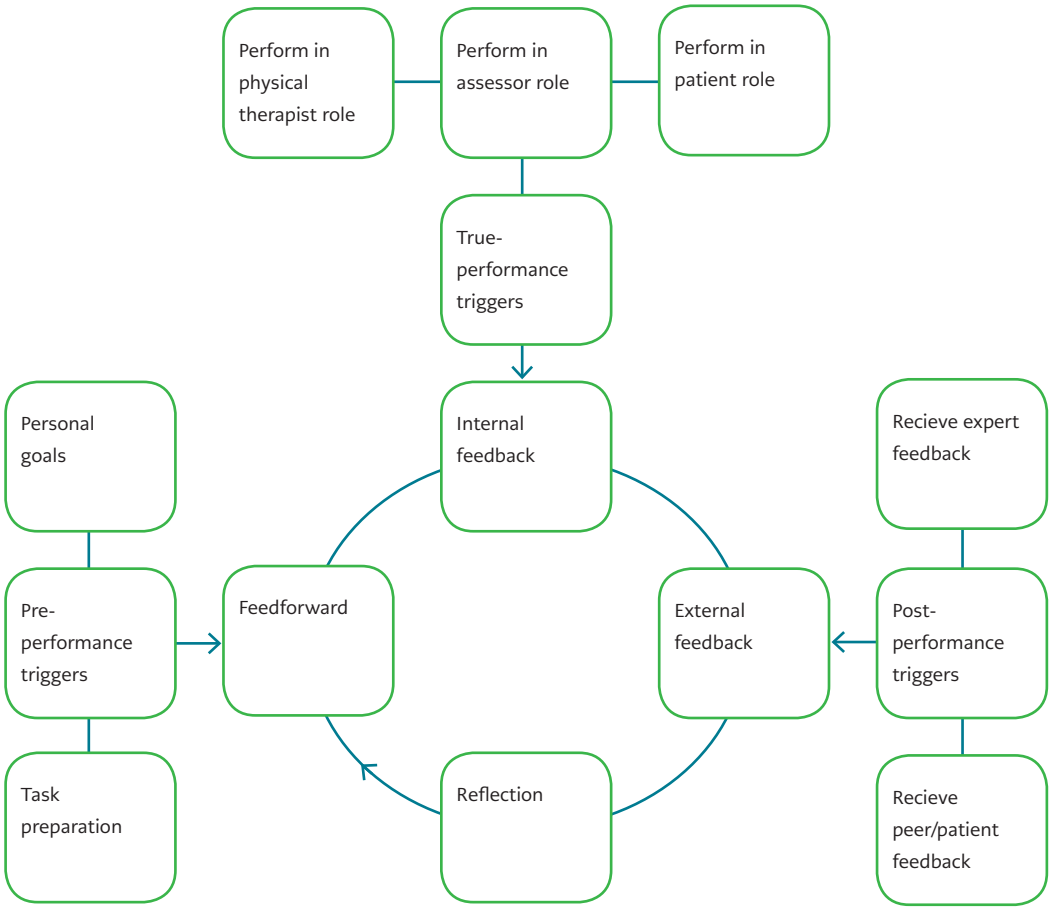
Writing reflection reports is perceived with mixed feelings. For some students it was helpful to self-direct their learning process, others perceived the (compulsory) task as unnecessary work load, especially for immigrants.

"… so I think … well I finished my assessment, I have received my feedback and now I also have to write it down. Actually, I do know enough. Why do that once again?"

Discussion

Our results show that the PA task contains a variety of elements that have a positive impact on the improvement of clinical performance. We developed a conceptual model, based on our results, that fully fits the data and that reflects how information is processed in PA to inform self-assessment (figure 1). The model shows that learning begins by anticipating the PA context as well as the PA content (pre-performance triggers), described as the 'backwash effect' of assessment[37] or the 'feed forward function' of assessment.[38]
Following the model, performance in different roles is the next phase. Performance in the physical therapist role was perceived to have the strongest impact on learning. This finding is in contrast with the general assumption that 'peer feedback' determines the impact of PA on learning.[39] However, social learning theory might provide an explanation. It emphasizes the importance of mastery experiences for performance improvement. Students needed to cope with anxiety triggers related to the context and content of PA. Successful coping resulted in increased self-confidence and awareness of strength and weaknesses. Studies of Bandura[40,41] show that mastery experiences are the strongest source of information for the development of self-efficacy beliefs and self-efficacy beliefs contribute significantly to the level of motivation for performance improvement.
Rush,[42] who studied the impact of PA on the performance of clinical skills in undergraduate nursing education, also reported high perceived learning value of performance in the nursing role, and found increased self-confidence as a dominant finding. Apparently, per-

Figure 1 — Conceptual Model of information processing



sonally perceived mastery evidence is more powerful than mastery evidence provided by peers in undergraduate education. The PA task challenged, or even forced, students to transfer knowledge and skills to a new application context which they apparently did not do spontaneously and that might be the key-feature of PA for improvement of clinical performance. Simons[43] argues that learners oftentimes do not and cannot know 'what' knowledge and skills need to be transferred and to 'what' new context, so they need help. Successful transfer of learning depends on the distance between the learning context and application context. A short distance (near transfer) refers to solving new problems in the same context. A long distance (far transfer) refers to solving new problems

in a new context. Apparently, the PA task construed a transfer gap that was 'near' enough to successfully bridge, but 'far' enough to be challenging. In effect, the task was in the 'zone of proximal development' as described by Vygotskiĭ.[44] Apart from these considerations that aimed to explain the superior perceived learning value of performance in the PT role, it should be noted that data analysis showed an interaction effect between performance in different roles. Learning experiences in the PT role may have been strengthened by performance in other roles and that may have influenced students' choice for the most valuable task element.

Concerning performance in the assessor role, we found results that were not reported by prior research. Peer assessors apply strategies to assess their peers that differ considerably from experts. Firstly, from a stakeholder perspective, students have different interests in observing the performance of their peers than expert assessors. Students have a need to improve their own performance whereby experts presumably do not. Thus peers may focus on different aspects than experts. Secondly, students obviously do not focus beforehand on critical features expressed in pre-determined criteria like expert assessors do. They use their peers, although not consciously. They 'match' and 'model' the observed performance to the image of their own performance. Research has revealed that the human motor system has mirroring capacity and is activated by observing motor actions made by others.[45] By mirroring the observed action, the brain is prepared to execute the same action. Calvo-Merino *et al*.[46] studied the differences in mirroring activity between watching an action that one has learned to do and an action that one has not. They compared experts in classical ballet with experts in capoeira (a traditional dance) observing both dancing styles and showed that mirroring activity is more powerful when expert dancers viewed movements that they had been trained to perform compared to movements they had not. The foregoing might explain students' engagement with observing their peers' performance. Although it is unknown how expert assessors actually view the performance of their students, it may be assumed that the 'virtual image' of the expert is different, as it is built and shaped by experience.[47] Thirdly, students observe more than experts. They are involved with the learning process of their peers and have more detailed knowledge of their learning needs than expert assessors have.

Receiving feedback represents the next phase in the model. Students preferred expert feedback over peer feedback because experts represent the performance standard for summative

decisions, which is obvious, because students depend on their judgment. Peer feedback, however, was valued because of its variety and its completeness and the involvement of peers in each other's learning was perceived a positive condition for identifying improvement areas. This finding is supported by several studies on PA in the health professions domain,[12,23–26,42] although some studies report reluctance of peers in giving face-to-face feedback.[33,48] Reflection represents the final phase in the model, referring to explicit conscious reflection on the PA-task that resulted in insight in strength and weaknesses and new learning goals. However, the perceived value of writing reflection reports was limited, which is understandable. Data show that reflection also occurred, although less explicitly, as a response to pre- and true-performance triggers as conceptualized by Schön's model of reflective practice.[49]

What this study adds to prior research is that peer judgment cannot be compared nor replaced by expert assessor judgments. However, the peer assessor view that develops during the PA process and the expert assessor view that represents the 'golden standard' may both provide students with rich just-in-time improvement feedback, built on multiple perspectives and connecting to their learning needs.[48] Research on PA revealed that effective PA processes depend on training and experience.[4,11,50] When peers continue to compare their personal views to peer views, they might gradually develop an internalized and mutually shared quality standard of performance that enhances professional development for now and after graduation. Future research should determine whether experienced peer assessors converge in their performance judgments.

Limitations

The generalizability of the qualitative data is limited. There is evidence that attitudes towards PA are gender- and cultural dependent[48] and that learning from PA depends on the PA-context, PA-content, and peer feedback preferences.[10,11,19,22,28,39,51] In addition, the generalizability of the quantitative data is limited because of the small sample size. It should be noted however, that the quantitative data were intended to strengthen the qualitative data and not vice-a-versa.

Conclusions

The PA task contains a variety of elements that have a positive impact on the improvement of clinical performance. It triggers intended learning through peer feedback and reflection as well as unintended

learning through matching and modeling a peer's performance. PA might be a powerful tool to help students in bridging the gap between the learning context and the application context. Peer feedback however is not perceived the most powerful task element in undergraduate physical therapy education and peer assessors use idiosyncratic strategies to assess their peers' performance.

Abbreviations

MM = Marjo Maas

EL = Els Lamers

JB = Joost de Beer

References

1 Wensing M, Grol R, Fluit C. Educatieve strategieën. In: *Implementatie: Effectieve Verbetering van de Patiëntenzorg*. 4th ed. Amsterdam: Reed Business; 2011:326-340.

2 Grol R. Quality improvement by peer review in primary care: a practical guide. *Qual Health Care*. 1994;3(3):147-152.

3 Pronovost PJ, Hudson DW. Improving healthcare quality through organisational peer-to-peer assessment: lessons from the nuclear power industry. *BMJ Qual Saf*. 2012;21(10):872-875.

4 Sluijsmans DMA, Van Merriënboer JJG, Brand-gruwel S, Bastiaens TJ. The training of peer assessment skills to promote the development of reflection skills in teacher education. *Stud Educ Eval*. 2003;29(1):23-42.

5 Davis DA, Mazmanian PE, Fordis M, Harrison R Van, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence. A systematic review. *JAMA*. 2006;296(9):1094-1102.

6 Eva KW, Regehr G. "I 'll never play professional football" and other fallacies of self-assessment. *J Contin Educ Health Prof*. 2008;28(1):14-19.

7 Rutten GM, Kremers S, Rutten ST, Harting J. A theory-based cross-sectional survey demonstrated the important role of awareness in guideline implementation. *J Clin Epidemiol*. 2009;62(2):167-176.

8 Epstein RM, Siegel DJ, Silberman J. Self-monitoring in clinical practice: a challenge for medical educators. *J Contin Educ Health Prof*. 2008;28(1):5-13.

9 Sargeant J, Eva KW, Armson H, *et al*. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ*. 2011;45(6):636-647.

10 Topping KJ. Methodological quandaries in studying process and outcomes in peer assessment. *Learn Instr*. 2010;20(4):339-343.

11 Van Zundert M, Sluijsmans D, van Merriënboer J. Effective peer assessment processes: research findings and future directions. *Learn Instr*. 2010;20(4):270-279.

12 Speyer R, Pilz W, Van Der Kruis J, Brunings JW. Reliability and validity of student peer assessment in medical education: a systematic review. *Med Teach*. 2011;33(11):572-585.

13 Gielen S, Peeters E, Dochy F, Onghena P, Struyven K. Improving the effectiveness of peer feedback for learning. *Learn Instr*. 2010;20(4):304-315.

14 Eva KW. Assessing tutorial-based assessment. *Adv Health Sci Educ Theory Pract*. 2001;6(3):243-257.

15 Dannefer EF, Henson LC, Bierer SB, *et al*. Peer assessment of professional competence. *Med Educ*. 2005;39:713-722.

16 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA*. 1993;269(13):1655-1660. http://www.ncbi.nlm.nih.gov/pubmed/8240483.

17 Violato C, Lockyer J. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: implications for self-directed learning. *Adv Health Sci Educ Theory Pract*. 2006;11(3):235-244.

18 Birenbaum M. Evaluating the assessment: sources of evidence for quality assurance. *Stud Educ Eval*. 2007;33(1):29-49.

19 Strijbos J-W, Sluijsmans D. Unravelling peer assessment: methodological, functional, and conceptual developments. *Learn Instr*. 2010;20(4):265-269.

20 Boud D, Falchikov N. Quantitative studies of student self-assessment in higher education: a critical analysis of findings. *High Educ*. 1989;18(5):529-549.

21 Tillema H, Leenknecht M, Segers M. Assessing assessment quality: criteria for quality assurance in design of (peer) assessment for learning – A review of research studies. *Stud Educ Eval*. 2011;37(1):25-34.

22 Van Gennip NAE, Segers MSR, Tillema HH. Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educ Res Rev*. 2009;4(1):41-54.

23 Cottrell S, Diaz S, Cather A, Shumway J. Assessing Medical Student Professionalism: An Analysis of a Peer Assessment. *Med Educ Online.* 2006;11(8):1-8.
24 Epstein RM. Assessment in medical education. *N Engl J Med.* 2007;356(4):387-396.
25 Schaub-de Jong MA, Cohen-Schotanus J, Dekker H, Verkerk M. The role of peer meetings for professional development in health science education: a qualitative analysis of reflective essays. *Adv Health Sci Educ Theory Pract.* 2009;14(4):503-513.
26 Nofziger AC, Naumburg EH, Davis BJ, Mooney CJ, Epstein RM. Impact of peer assessment on the professional development of medical students: a qualitative study. *Acad Med.* 2010;85(1):140-147.
27 Abdulla A. A critical analysis of mini peer assessment tool (mini-PAT). *J R Soc Med.* 2008;101(1):22-26.
28 Norcini JJ. Peer assessment of competence. Med Educ. 2003;37(6):539-543.
29 Archer JC, Norcini J, Davies HA. Use of SPRAT for peer review of paediatricians in training. *BMJ.* 2005;330(7502):1251-1253. doi:10.1136/bmj.38447.610451.8F.
30 Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ.* 2005;39(12):1188-1194.
31 Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Med Educ.* 2006;40(7):618-623.
32 Opheij M, Maas M, de Beer J. Onderzoek naar de inhoudsvaliditeit van het performance-assessment in de hoofdfase van de bacheloropleiding fysiotherapie. *Tijdschrift voor Med onderwijs.* 2006;25(2):88-95.
33 Falchikov N. Improving *Assessment through Student Involvement: Practical Solutions for Aiding Learning in Higher and Further Education*. 2nd ed. New York: Routledge Falmer; 2013.
34 Janssen-Noordman AMB, Merriënboer JJG, van der Vleuten CPM, Scherpbier AJJA. Design of integrated practice for learning professional competences. *Med Teach.* 2006;28(5):447-452.
35 King N, Cassel CM, Symon G. Using templates in the thematic analysis of texts. In: Cassell C, Symon G, eds. *Essential Guide to Qualitative Methods in Organizational Research*. 1st ed. London: Sage Publications; 2004:256-270.
36 Miles MB, Huberman MA. *Qualitative Data Analysis. An Expanded Sourcebook*. Thousand Oaks: Sage; 1994.
37 Biggs J. What the student does : teaching for enhanced learning. *High Educ Res Dev.* 2006;18(1):57-75.
38 Kluger AN, van Dijk D. Feedback, the various tasks of the doctor, and the feedforward alternative. *Med Educ.* 2010;44(12):1166-1174.
39 Liu NF, Carless D. Peer feedback: the learning element of peer assessment. *Teach High Educ.* 2006;11(3):279-290.
40 Bandura A, Locke E a. Negative self-efficacy and goal effects revisited. *J Appl Psychol.* 2003;88(1):87-99.
41 Bandura A. Self-Efficacy: *The Exercise of Control*. John Wiley & Sons, Inc.; 1997.
42 Rush S, Firth T, Burke L, Marks-Maran D. Implementation and evaluation of peer assessment of clinical skills for first year student nurses. *Nurse Educ Pract.* 2012;12(4):219-226.
43 Simons P. Transfer of learning: paradoxes for learners. *Int J Educ Res.* 1999;31(7):577-589.
44 Ormrod JE. *Human Learning*. 4th ed. Upper Saddle River: Pearson Prentice Hall; 2004.
45 Iacoboni M. *Mirroring people: The new science of how we connect with Others*. New York: Farrar, Straus and Giroux, 2009.
46 Calvo-Merino B, Glaser DE, Grèzes J, Passingham AE, Haggard P. Action observation and acquired motor skills: an FMRI study with expert dancers. *Cereb Cortex*. 2005 (15):1243-1249.
47 Govaerts MJ, Schuwirth LWT, van der Vleuten CPM, Muijtjens AM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract*. 2011(16):151-165.
48 Hanrahan S, Isaacs G. Assessing self and peer-assessment: the students'views. *High Educ Res Dev.* 2001;20(1):53-70.
49 Schön D. *The Reflective Practitioner: How Professionals Think in Action*. San Francisco: Jossey-Bass Inc; 1983.
50 Van Gennip NA, Segers MS, Tillema HH. Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learn Instr*. 2010;20(4):280-290.
51 Lin SSJ, Liu EZF, Yuan SM. Web-based peer assessment : feedback for students with various thinking-styles. *J Comput Assist Learn.* 2001;17:420-432.

Chapter 3
Effectiveness of peer assessment for implementing a Dutch physical therapy low back pain guideline: a cluster randomized, controlled trial

Simone van Dulmen
Marjo Maas
Bart Staal
Geert Rutten
Henri Kiers
Ria Nijhuis-van der Sanden
Philip van der Wees

Physical Therapy, 2014;94(10):1396–1409

Abstract

Background

Clinical practice guidelines are considered important instruments to improve quality of care. However, success is dependent on adherence, which may be improved using peer assessment, a strategy in which professionals assess performance of their peers in a simulated setting.

Objective

To determine whether peer assessment is more effective than case-based discussions to improve knowledge and guideline consistent clinical reasoning in the Dutch physical therapy guideline for low back pain (LBP).

Design

Cluster-randomized controlled trial was conducted.

Setting and participants. Ten Communities of Practice (CoPs) of physical therapists were cluster-randomized (N=90): six CoPs in the peer assessment group (n=49) and four CoPs in the control group (n=41).

Intervention. Both groups participated in four educational sessions and used clinical cases. Peer assessment group reflected on performed LBP management in different roles. The control group used structured discussions.

Measurements. Outcomes were assessed at baseline and at six months. The primary outcome measure was knowledge and guideline-consistent reasoning, measured with 12 performance indicators using four vignettes with specific guideline related patient profiles. For each participant the total score was calculated by adding up the percentage scores (0-100) per vignette, divided by four. Secondary outcome measure was reflective practice as measured by the Self-Reflection and Insight Scale.

Results

Vignettes were completed by 78 participants (87%). Multilevel analysis showed an increase in guideline-consistent clinical reasoning of 8.4% in the peer assessment groups whereas the control groups showed a decline of 0.1% (estimated group difference = 8.7%, 95%CI: 3.9 to 13.4). No group differences were found on self-reflection.

Limitations. The small sample size, a short-term follow-up, and the use of vignettes as a proxy for behavior were limitations of the study.

Conclusions

Peer assessment leads to an increase in knowledge and guideline-consistent clinical reasoning.

Introduction

Clinical practice guidelines are considered as useful tools for quality improvement.[1] However, successful implementation is necessary to decrease the gap between research and current practice, and to reduce costs and unwanted variability in practice.[2,3] Adherence to guideline recommendations for patients with low back pain (LBP) is associated with improved quality of care, increased activities, fewer visits and better outcomes.[4,5] Especially for patients with LBP new research results have accumulated over the past years, requiring an update of the guideline for physical therapy management of LBP.[6] It is a challenge to implement a revised guideline when physical therapists already have a lot of experience in treating these patients. Physical therapists may have to change their behavior based on new research findings, so they need to be aware of the sometimes small but determining differences.[7]

To be successful in implementation several barriers should be addressed, including barriers on individual, social, organizational, economic or political levels.[8-11] Comprehensive implementation strategies are essential to increase adherence to guideline recommendations. Research shows that guideline consistent behavior in physical therapy shows room for improvement.[12-15] The most important discrepancy between current practice and guideline recommendations in physical therapy is related to knowledge and skills, awareness of or familiarity with guidelines, and external factors.[9,13,16,17] As regards the Dutch LBP guideline, physical therapists in the Netherlands are no exception in this respect.[18-20] A qualitative study of Harting *et al.* identified barriers to the adoption process of guidelines, lack of practical skills and unfavorable attitude for using guidelines.[16] The use of measurement instruments is limited as result of a lack of knowledge for applying, scoring and interpreting measurement instruments.[13,21]

To improve the uptake of guidelines in physical therapy, implementation strategies should be focused on improvement of knowledge, skills, attitude and awareness of guideline adherence.[7,16,18] Small-group education and peer review are widely used methods for guideline implementation and changing professionals' performance, to support critical appraisal of personal quality of care.[22-25] *Small-group*

education can be defined as continuing medical education or skills training on specific subjects in a small group of peers. *Peer review* is defined as a "continuous, systematic, and critical reflection by a number of care providers, on their own and colleagues' performance, with the aim of achieving continuous improvement of the quality of care".[24] Peer review may include different methods, such as consensus development, evaluation of performance, practice visit or peer assessment. Peer assessment is a specific form of peer review in which professionals assess (*judge*) the performance of their peers using relevant criteria and providing feedback.[26]

For implementing the revised LBP guideline we hypothesized that peer assessment as a specific peer review strategy could be appropriate to change professional behavior. Peer assessment aims at increasing self-reflection and improving awareness of actual performance. Triggers for learning and change concern both providing and receiving feedback.[25,27-30] Several studies have been conducted to study the impact of peer assessment. Ramsey *et al*.[31] demonstrated that peer ratings provide a practical method to assess the performance of practicing physicians on clinical skills, humanistic qualities, and communication skills.[31] A review of Overeem *et al*.[32] showed that 61-72% of participating physicians reported a change in their behavior using peer assessment.[32] Similar results were found by Sargeant *et al*.[33] using a multisource feedback tool including peer assessment by family physicians.[33] Case-based discussions are commonly used in postgraduate education as a strategy for implementing guidelines, stimulating reflection, and integrating scientific knowledge in clinical reasoning and decision-making.[34-36] The main difference between peer assessment and case-based discussions is that peer assessment focuses on assessment of performance rather than discussions.

Knowledge of – and adherence to the guideline can be assessed in different ways. Self-reports are practical and inexpensive to measure clinical performance, although they may overestimate guideline adherence.[37] Using medical record review might be problematic in achieving a sufficient case mix.[38] Measurement by direct observations or using standardized patients is expensive and time consuming.[32,38,39] Clinical vignettes are written patient cases that approach as much as possible the authentic context of practice. They require factual guideline knowledge as well as guideline-consistent clinical reasoning in the context of a clinical problem. Therefore, vignettes are a suitable means of assessing knowledge and evaluating guideline consistent clinical reasoning. In assessing intentional behavior, clinical vignettes are a proxy for guideline adherence and clinical behavior.[19,38,40-44]

In the present study we compared the tailored peer assessment strategy with the case-based discussion strategy in post graduate education. Both groups used the same clinical written patient cases. The intervention group used peer assessment in which they reflected on performed LBP management in three roles: patient, physical therapist and assessor. Additionally, they developed and evaluated a personalized improvement plan. The control group used structured case-based group discussions with written clinical cases. The effect on knowledge and guideline consistent reasoning was measured using clinical vignettes: descriptions of four patient cases with specific guideline related patient profiles.
We hypothesized that peer assessment is more effective to improve guideline knowledge, guideline consistent clinical reasoning, and reflective practice than case-based discussions as regular activities in postgraduate education. The objective of our study was therefore to compare the peer assessment strategy with the case-based discussion strategy in post-graduate education. We used the updated Dutch LBP guideline for physical therapists[6] because of the high prevalence of this condition in clinical practice.

Method

Design overview

We conducted a cluster-randomized controlled trial among Communities of Practice (CoPs) of Dutch physical therapists from January to September 2010 to evaluate the effect of an implementation strategy on guideline knowledge and guideline consistent clinical reasoning. Both educational programs (peer assessment and the case-based discussions) included multifaceted strategies to improve knowledge and clinical reasoning skills according to the Dutch LBP guideline for physical therapy.[6] Both educational approaches consisted of a series of four 2-hours meetings during a 6-month period. Changes in knowledge and guideline consistent clinical reasoning were assessed with vignettes at baseline and at six months.

Setting and Participants

In September 2009, all contact persons of the approximately 800 existing CoPs within the professional body of physical therapists in the Netherlands (Royal Dutch Society for Physical Therapy) received an electronic newsletter from the secretary of the Royal Dutch Society for Physical Therapy with an invitation to choose a topic out of the approximately 30 post-graduate educational programs of the coming year. One of the programs was an educational trajectory for

implementing the updated LBP guideline. CoPs are small groups of 5–15 physical therapists that share the same setting, specialization, or interests and who work together on quality improvement by choosing each year an educational program. The CoPs are broadly oriented and may include many different specializations, e.g. specializations on pediatric physical therapy, and may also include physical therapists working in both primary and secondary care. CoPs of physical therapists treating patients with LBP on a regular basis were eligible for inclusion in this educational trajectory.
A meeting was organized for the interested contact persons in November 2009 to provide information about the aim of the project and study procedures. After this meeting the CoPs could decide to participate in the study.
We explored the required sample size based on an estimated important difference of at least 5% for the primary outcome measure, with an anticipated intra-class correlation (ICC) of 0.05, and 10% loss to follow up. Our estimation was based on the effectiveness of audit and feedback, which generally leads to small but potentially important improvements in professional practice with an overall improvement of adherence to desired practice of 5%.[45] This procedure resulted in a required inclusion of n=103 physical therapists in 12 clusters, which we used as target for our study.

Randomization and intervention allocation

All participants of committed CoPs visited 1 of 2 joint meetings organized in January 2010, where the updated LBP guideline was presented and modifications in the revised guideline were explained. The participants were informed that the study consisted of two educational strategies and that both strategies were comparable and required an identical time investment in 4 meetings. Randomization at CoP level was conducted after the meetings. CoPs were randomized using a computerized randomization system. An independent research assistant (A.S.) who was not blinded for the allocation drew up an allocation schedule using a computerized random number generator, listed them by the number of the CoP, and subsequently assigned them to the peer assessment group or the case-based discussion group according to the allocation schedule, and informed the contact persons of the CoPs of the allocation by e-mail. The research assistant safeguarded the allocation codes, which were only revealed after the data analysis. The principal investigators (P.W. and S.D.) did not attend the meetings with the CoPs and were blinded for the allocation of the CoPs throughout the study. After the allocation, the participating physical therapists

received an electronic questionnaire to gather demographic information. Meetings were each 4-6 weeks, depending on the available working schedule of the physical therapists.

Intervention: problem-based peer assessment

Peer assessment was aimed at improving guideline consistent knowledge, *clinical* reasoning skills and performance. In peer assessment clinical performance was directly observed and evaluated by peers in a simulated setting. Participants received a peer assessment manual in advance, which contained a description of the peer assessment procedure, a time schedule and instructions for providing constructive feedback. Performance was assessed with a scoring sheet containing performance criteria that could be scored on a 7-point scale (1= much improvement needed, to 7= no improvement needed) and some space for qualitative feedback. Performance categories addressed the diagnostic process (choice for diagnostic tests and measurement instruments, performance of clinical tests, and evaluation of outcomes) and the intervention process (choice of interventions, performance of interventions, and evaluation of outcomes).

The scoring sheet was developed and validated in another study and slightly modified to the new guideline criteria.[46] The peer assessment CoPs were coached by an expert assessor (MM), a physical therapist with expertise on LBP and an experienced teacher. The expert assessor participated in the role of process moderator and end-assessor, providing additional feedback only if necessary and when all peers had given their feedback.

During the first 2 meetings written cases were presented, accompanied by assignments for patient role performance. Participants performed in 3 roles: the physical therapist, assessor and the patient role. In the physical therapist role they were blinded for the simulation role description of the patient so it was not known in advance what specific clinical problem was simulated. Performance in the physical therapist role included communicative skills, hands-on diagnostic and treatment skills. Choices for diagnosis and treatment were explicated by reasoning aloud. In the assessor role, participants observed the performance of their peers and provided them with oral and written feedback. In the patient role, participants simulated a clinical problem according to brief simulation guidelines. Each participant developed a personal plan for improvement, including an action plan, based on feedback and assessment of colleagues during the first 2 meetings, and a strengths, weaknesses, opportunities and threats (SWOT) analysis

of their own performance, which was evaluated and discussed with their peers during the third meeting. They clarified their plan and received feedback from the other participants. In the final meeting, participants evaluated their action plan and another session of peer assessment was scheduled. This session was identical to the first two meetings; however, patient cases were adapted by MM to meet the specific learning needs of the participants, such as screening of "red flags".

Control: case-based discussion

Routine case-based discussion was aimed at improving guideline consistent knowledge and reasoning skills. Participants received a program manual that contained a structured program schedule, including a description of the case discussion procedure, a time schedule, and cases for each meeting that were given in advance. For each meeting, assignments were given to guide and evaluate the case-discussion process: 1) supportive questions for unraveling the problem, 2) supportive questions for establishing a physical therapy diagnosis and an intervention plan, and 3) assignments to make a summary of the discussions of each meeting. After each meeting, learning results were evaluated by the group. Each participant had to explicate his/her lessons learned. During the fourth meeting 25 written statements about the anatomical and physiological structures, etiology, diagnosis, and treatment were discussed. After this meeting, participants individually answered the statements as being true or false via an online system and received feedback on each answer from the research assistant. There was no external coach to guide the discussion process, because CoPs were familiar with this educational format.

Outcomes and follow-up

Outcomes were assessed at baseline and after 6 months when both groups had finalized their meetings. Primary outcome measure was knowledge of the LBP guideline and guideline consistent reasoning which was measured by 4 clinical vignettes, developed by Rutten *et al*.[19] The vignettes were validated and showed to have an adequate validity as a proxy measure for physical therapists' adherence to the LBP guideline.[19] These vignettes were modified to the updated guideline.[6] The vignettes represented 4 patient profiles: 1) a patient with acute non-specific LBP and an expected normal recovery process; 2) a patient with sub-acute non-specific LBP and an imminent delay in the recovery process (indicating that the activities and participation showed no progress during the past three weeks); 3) a patient with

sub-acute non-specific LBP and a delayed recovery with intervening psychosocial factors; and 4) a patient with LBP due to an underlying, serious spinal pathology (e.g. inflammatory process, tumor etc.). Profiles 1 through 3 also are presented in the LBP guideline, profile 4 is described in the guideline, but it is not labeled as a profile.
Text in the vignettes was presented in separate blocks similar to the steps in the guideline. Each block was followed by questions[40] related to the assessment of patients' complaints, diagnostic activities, profile selection, the use of health outcome questionnaires, whether they would contact the referring physician, treatment objectives and strategies, expected number of treatment sessions, information and advice to be provided, planned evaluation, aftercare, and a report to the referring physician.
Participants were asked to complete the questionnaires online after the joint meetings but before the start of the first group sessions, and post-intervention within four 4 after finishing the final group sessions. The score for each vignette depended on the specific guideline recommendations for specific patient profiles. Per vignette and for each step in clinical decision-making a performance indicator was used to measure guideline knowledge and guideline consistent clinical reasoning, in total twelve indicators (Table 1). Performance indicators have been defined as measurable elements of practice performance that can be used to assess the quality of care.[47,48] Per vignette, for each indicator, one or more questions were formulated. Answers that matched the recommendation in the guideline were given a point, whereas answers that contravened the recommendation were given no points. For each indicator, a percentage score was calculated by dividing the actual number of correct answers by the maximum possible score and multiplying the result by 100. For each vignette the total percentage score was calculated based on the indicator scores divided by the number of indicators. In addition, a mean percentage score for overall guideline adherence was calculated by adding the 4 vignette scores and dividing the total by 4 with a score range from 0 (minimal knowledge/guideline consistent clinical reasoning) to 100 (maximal knowledge/guideline consistent clinical reasoning). This method is known as the patient average method.[49] Scores on the vignettes were calculated when at least 75% of the indicators were completed and the overall score was calculated when at least 3 vignettes were completed (in which case the total score was divided by 3.
The secondary outcome measure was self-reflection, measured by the Self-Reflection and Insight Scale (SRIS).[50] The SRIS is a validated instrument to measure the process of self-reflection and insight

Table 1 — Performance indicators to measure guideline adherence based on profiles of patients with low back pain in clinical vignettes

	Indicator	Description
1	Red flags assessed correctly	Identification of dangerous or potentially dangerous findings in the history or examination, e.g. pain at night or unexpected body loss
2	Assessment of the patients' complaints	To assess all relevant domains in relation to a patients' health: body function, activity, participation, environmental and personal factors (according to the International Classification of Functioning, Disability and Health)
3	Correct choice of the patient profile	Patient profile determined by the course of the symptoms and factors that prevent recovery (profile 1: non-specific acute LBP and a normal recovery process; profile 2: non-specific sub-acute LBP and an imminent delay in the recovery process; profile 3: non-specific sub-acute LBP and a delayed recovery with intervening psychosocial factors; and profile 4 (not a formal profile in the guideline): LBP due to an underlying, serious spinal pathology
4	Contacting the physician in case of red flags	Contacting the physician in case of LBP due to a suspected underlying, serious spinal pathology (profile 4)
5	Choice of examination objectives related to the patient profile	Examination objectives on domains of body function, activity, participation, environmental and external factors
6	Choice of treatment objectives related to the patient profile	Treatment objectives on domains of body functions, activities, participation, environmental and personal factors
7	Choice of treatment strategies related to patient profile	Recommendations are described on treatment strategies at the start and at a later stage of the treatment
8	Number of intervention sessions	Number of sessions is limited to a maximum of three in case of acute LBP with normal course
9	Adequate information is provided	Recommendations are described on treatment strategies at the start and at a later stage of the treatment
10	Health outcome questionnaires have been applied	Measurements for diagnostic and evaluation, Numeric Rating Scale (NRS), Quebec Back Pain Disability Scale (QBPDS) or Patient-Specific Function Scale (PSFS)
11	Written report to physician	Report to the physician with information about diagnosis, intervention, number of session
12	Aftercare has been arranged	Information about what to do in case of a recurrence

that is presumed as conditional to self-directed change. Reflection allows assimilation and reordering of concepts, skills, knowledge, and values into pre-existing knowledge structures and is therefore conditional for learning new knowledge, skills, and behavioral change.[27,51,52] The SRIS is a self-administered, 20-item closed questionnaire with a 5-point Likert scale measuring engagement and insight in self-reflection, and the need for self-reflection. The total score could range from 20-100, with higher scores indicating more self-reflection.The validated version of Roberts and Stark[53] was translated by two researchers (P.W. and M.M.) and expert validity of this version was obtained by three experts who judged the translation. Their comments were used to improve the Dutch version of the SRIS.

Data Analysis

The characteristics of the participants in the two groups were described and tested for differences between the two arms using Chi-square tests, unpaired t-tests and Mann-Whitney U tests. Post-intervention mean total scores on the 4 vignettes of each participant were included as outcome variable in a multilevel model, and baseline scores were included as covariates. Baseline characteristics were considered confounders if they were 1) significantly associated with the outcome variable, and 2) significantly different between the groups. If both conditions were met, they were added as covariate to the multilevel model to adjust for confounding. Identical analyses were performed with the follow up score on the SRIS questionnaire as outcome variable. Statistical significance was tested using two-sided tests at *P*-value of <.05. To determine the associations of the score at CoP level, we calculated the Intra-Class Correlations (ICC)[54] of the scores on the vignettes from the output of the multilevel analysis with covariance parameters included. For each indicator we calculated mean scores at baseline and at 6 months for both groups. All statistical analyses were performed using IBM SPSS Statistics for Windows, version 20.

Role of funding source

This study was a researcher-initiated study, primarily funded by the KNGF, with co-funding of the Radboud University Medical Center, the Scientific Institute for Quality of healthcare, and the HAN university of Applied Sciences. The KNGF had no role in the conduct of this study, analysis or interpretation of data, or manuscript preparation.

Results

The flow of participants is presented in Figure 1. After the invitation 35 contact persons of the CoPs visited the information meeting. Thirteen contact persons were potentially interested to participate in our study. Based on these expected CoPs a randomization scheme was determined. Three decided not to participate because of other priorities. Ten CoPs were initially cluster-randomized (N=90); 6 CoPs in the peer assessment group (n=49) and 4 CoPs in the control group (n=41).

Table 2 presents characteristics of the participants and their practices (n=78). The mean age of the participants was 42.7 years (SD = 11.6), with mean practice experience of 18.7 years (SD = 11.0), and 56% of the participants were female. The participants comprised a representative sample on age and gender when compared with national reference data.[55,56] Differences in gender and the amount of LBP patients per year were not statistically significant between the groups. The years of experience were significantly higher in the control group but no relationship was found between the scores on the vignettes and years of experience. A significant difference was found for the proportion of manual therapists between the groups. Moreover, manual therapists had significantly higher scores on the vignettes, so this was added as covariate in the analysis.

The primary outcome measure could be analyzed for 78 of allocated participants (87%). After randomization, 3 participants dropped out of the study: 1 in the peer assessment group and 2 in the case-based discussion group. Four participants of the peer assessment group and 5 participants of the case-based discussion group had incomplete scores on the vignettes.

Table 3 shows the mean scores for the indicators at baseline and at follow-up for the multilevel analysis. Mean increase was 8.4% in de peer assessment group, whereas the scores in the control group declined with 0.1%. Improvement scores on vignettes at 6 months post-intervention were significantly higher in the peer assessment group, with an estimated group difference of 8.7% ($P \leq .001$; 95% Confidence Interval [CI]: 3.9 to 13.4). Twenty participants did not complete the SRIS, so the secondary outcome measure could be analyzed for 70 participants (78%). Mean baseline score of the peer assessment group was 74.0 points and 79.9 points of the control group. The improvement on the SRIS questionnaire was 2.5 points in the peer assessment group and 0.5 points in the control group. The estimated group difference in improvements between the two groups was non-significant (-0.69 points, $P = 0.63$; 95%CI: -3.5

Figure 1 — The flow of participants

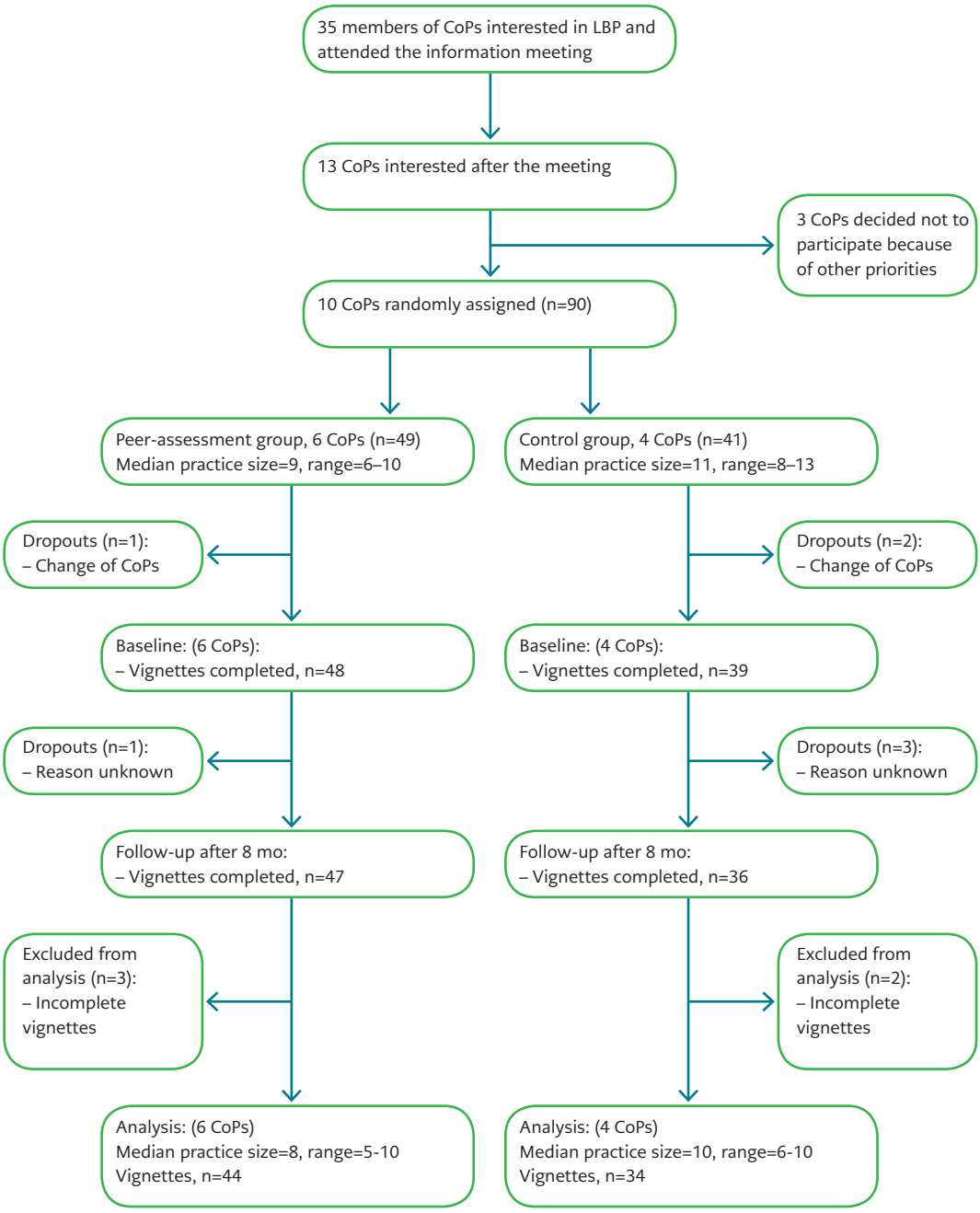


Table 2 – Physical Therapist and practice characteristics

Characteristic	Peer assessment group (n=44)	Control group (n=34)[a]
Age mean (SD)	40.4 (12.4)	45.8 (9.9)
Sex (male/female)	17/27	18/16
Hours worked per week mean (SD)	32.5 (9.6)	32.2 (10.5)
Treatment of patients with LBP per year		
<25	11	11
25-50	12	3
50-75	6	3
76-100	5	3
>100	10	13
Manual therapist, n	8	17
Years of experience mean (SD)	16.5 (11.9)	21.2 (9.2)

[a] Information on one participant in the case-based group was missing for the variables of age, hours worked per week, treatment of patients with low back pain per year, and years of experience

Table 3 — Effect of intervention on therapist knowledge, clinical reasoning, and self-reflection

Measure	Peer assessment group		Case-based discussion group		Intervention effect (95% CI)[a]	*P*
	Baseline	Follow up	Baseline	Follow up		
Vignettes				66.7 (13.1)	8.7 (3.9 to 13.4)	.001*
Mean	63.7	72.0	66.8			
SD	10.2	11.6	10.1			
Range (0-100)	44-89	41-98	47-84	47-87		
SRIS[b]						.63
Mean	74.0	76.5	79.9	80.4	- 0.69 (-3.5 to 2.2)	
SD	8.5	9.2	8.6	8.6		
Range (0-100)	54-96	59-95	56-100	63-99		

[a] 95% Confidence Interval, [b] Self-Reflection and Insight Scale, * P ≤.05

Table 4 — Mean scores and effect of intervention on therapist knowledge and clinical reasoning on the individual performance indicators

	Indicator	Peer assessment group		Case-based discussion group			
		Baseline Mean (SD)	Follow-up Mean (SD)	Baseline Mean (SD)	Follow up Mean (SD)	Intervention effect (95% CI)	*P*
1	Red flags assessed correctly	66.3 (14.3)	90.9 (15.3)	70.6 (14.9)	88.3 (15.5)	6.7 (0.3 to 13.1)	.04*
2	Assessment of the patient's complaints	84.8 (15.4)	81.3 (14.7)	86.0 (14.2)	82.3 (15.7)a	-1.5 (-8.6 to 5.7)	.69
3	Correct choice of the patient profile	44.3 (28.5)	52.3 (23.0)	51.5 (31.2)	41.7 (23.0)	13.1 (2.2 to 24.0)	.02*
4	Contacting the physician in case of red flags	89.2 (12.5)	92.6 (12.7)	92.0 (13.4)	89.8 (14.0)	6.8 (1.1 to 12.4)	.02*
5	Choice of examination objectives related to the patient profile	60.5 (10.9)	58.9 (12.9)	54.1 (15.3)	52.9 (15.9)	1.2 (-5.1 to 7.6)	.70
6	Choice of treatment objectives related to the patient profile	60.1 (20.4)	67.8 (18.6)	55.2 (19.5)	61.2 (17.8)	8.2 (-1.8 to 18.1)	.11
7	Choice of treatment strategies related to the patient profile	71.6 (14.9)	77.3 (15.4)	67.7 (14.0)	62.8 (19.0)	12.4 (1.5 to 23.3)	.03*
8	Number of intervention sessions	70.7 (24.9)	85.8 (16.7)	79.6 (22.2)	85.1 (16.9)	3.1 (-5.8 to 12.1)	.49
9	Adequate information is provided	43.4 (22.42	46.6 (21.7)	39.5 (23.8)	46.2 (29.3)	-0.6 (-13.6 to 12.3)	.93
10	Health outcome questionnaires have been applied	60.9 (23.5)	64.1 (22.5)	53.3 (21.0)	56.6 (22.6)	5.8 (-4.4 to 16.0)	.26
11	Written report to the physician	81.6 (30.0)	79.8 (31.8)	86.3 (27.4)	86.0 (30.8)	-3.9 (-18.2 to 10.3)	.58
12	Aftercare has been arranged	85.3 (32.0)	94.6 (16.2)	91.2 (26.3)	91.7 (20.7)	4.9 (-2.0 to 11.8)	.16

* $P \leq .05$

to 2.2). The ICC was <0.00, indicating that the clustering effect is negligible after adjusting for covariates. Table 4 presents the mean scores for the performance indicators at baseline and at 6 months and the results of the multilevel analysis for each indicator.

Discussion

Our results confirmed the hypothesis that the tailored peer assessment strategy was more successful to increase knowledge and clinical reasoning consistent with recommendations in the LBP guideline, compared to routine case-based discussions. This effect may be explained by the combination of different educational strategies: dissemination of the guideline, in-depth assessment of the guideline in a problem solving process, assessment of performance, individualized well-timed performance feedback, and an individually tailored improvement plan. Peer assessment did not result in improved reflective practice.

The strength of the peer assessment strategy is that participants performed different roles, which leads to a reflection on the guideline from various perspectives. In the assessor role, they had to reflect on professional qualities of colleagues using guideline recommendations as gold standard. This facilitates the ability to improve clinical skills while comparing the observed performance of colleagues with their own performance level and the guideline. In the physical therapist role, participants reflected on their own knowledge and performance using the feedback of their peers. In the patient role they were able to reflect on the communication and perception of diagnosis and treatment from the patients' perspective. This triangle of feedback might increase reflection and the awareness of individual shortcomings which are considered key factors in guideline implementation and improvement of professional practice.[7,57] In addition, the feedback was used to develop a tailored and individualized improvement plan. Finally, the peer assessment groups were coached by an expert assessor. By representing the 'gold standard', the expert assessor might have played an important role in stimulating and reinforcing the feedback, and avoiding long discussions without endpoint or consensus. It is not clear which aspects of the educational process are attributed most to the results of this study.

Intentional change of professional behavior and improved knowledge of guidelines does not necessarily lead to a concurrent change in patient outcomes. In various studies better guideline adherence and professional behavior was not associated with improved patient

outcomes.[12,58,59] Reviews focused on the effect of audit and feedback demonstrated that patient outcomes were less commonly measured and showed mixed results.[45,60] We found only one study in which similar implementation strategy was performed with outcomes on patient level. In this study peer assessment was used to improve care for patients with asthma/chronic obstructive pulmonary disease by general practitioners and showed no differences in provided care or in patients' health status.[61] Therefore, the results of our study must be interpreted with caution. Further evaluation of this strategy with appropriate designs to measure outcomes at patient level is needed.

Although peer assessment can be a process that fosters reflection on professional quality,[62] in our study self-reflection as measured with the SRIS did not improve. For self-reflection as secondary outcome measure no cut-off values for clinical importance were set. We hypothesized that an improvement of minimal 5% would be of clinically importance, based on studies that assessed the effectiveness of implementation studies,[25,63,64] and on audit and feedback.[45] There are very few studies comparing performance with self-reflection as measured with the SRIS. The SRIS was used in a course that aimed at improving reflective practice of social work students and resulted in a significant improvement on the SRIS of 14.6 points.[65] Another study that used reflection as an approach to learn showed improvement of 1.3 points.[66] These large variations in improvements made it difficult to reflect potential clinical important differences. Both intervention groups in our study showed fairly high baseline scores and comparable improvement scores, indicating both interventions affected conscious reflection. The process of reflection is influenced by individual aspects and practice context.[67] Further research is necessary to identify the role of reflection in this implementation strategy and to test the validity of the SRIS in postgraduate education.

In assessing differences between the peer assessment and case-based discussion group in knowledge and guideline consistent clinical reasoning per indicator, we found lower baseline scores and significant improvements in the assessment of red flags (indicator 1), choice of the patient profile (indicator 3), contacting the referring physician in case of red flags (indicator 4), and the choice of treatment strategies (indicator 7). All these indicators include recommendations that were modified in the revised guideline. This might explain the lower baseline scores for these indicators, allowing for more improvement potential. We adjusted for the proportion of manual therapists in the multilevel analysis, because there was a significant

difference in the proportion of manual therapists between the two groups and their baseline scores on the vignettes, with higher score of manual therapists. Manual therapists are presumed to be familiar with topical results from clinical research on LBP, which might explain the difference on the baseline score.

Limitations

Our study has several limitations and in the light of these, the results should be interpreted with caution. First, although peer assessment did improve knowledge and clinical reasoning consistent with recommendations in the LBP guideline, we have not demonstrated that the intervention has changed the actual behavior of physical therapists in clinical practice, or resulted in better patient health outcomes. Vignettes, by construct, do not capture all important elements of care that are critical to overall patient well-being.[40] Vignettes are assumed to rather measure attitudes and perceptions rather than actual behavior,[68] although recent studies have demonstrated the validity of vignettes as (proxy) measure for clinician performance.[19,38,40-42,44] While the validity of vignettes used in this study was deemed acceptable,[19] other measurement instruments may be desirable.[38,44] The results of our study can be used for further analysis of using vignettes to improve knowledge and guideline consistent clinical reasoning and to assess the relationship with clinical practice and patient outcomes.

Second, our study was conducted with small, although representative,[55,56] self-selected sample of CoPs and physical therapists, herewith threatening external validity of the study. However, this self-selection is common in postgraduate educational interventions. Therefore, the results may be generalizable to health professionals who are motivated to improve their quality of care and adherence to clinical practice guidelines, and if well supervised, this method can be integrated in regular teams in primary care practices and inpatient facilities such as hospitals. We anticipate on developing a course, including a training manual, for expert assessors in a national implementation program. Vignettes need to be developed for each new guideline. After allocation of 90 participants, the primary outcome was measured for 78 participants (87%), so this is a high percentage of the initial number of participants. The characteristics of dropouts and participants did not differ in mean age and working hours per week or baseline scores on the vignettes, which suggests that this small number of dropouts did not influence the results (data not shown). However, we did not evaluate the reasons for dropout so it is unknown if and in which way this group affected the results.

Third, the scores on the vignettes of both groups at baseline appear to be rather high compared with results from other adherence studies.[14,19,25,69,70] This might be explained by the interactive educational meeting both groups received before completing the baseline vignettes. It is known that interactive workshops can change professional practice,[71,72] which may have resulted in higher baseline scores. Furthermore, this was an updated guideline and participants may have received educational training in the LBP guideline previously. This is confirmed by the lower baseline scores on the indicators with items that were modified in the revised guideline. Moreover, registration for participation was voluntary, so selection bias of participants could have influenced the scores. They probably volunteered because they were interested in LBP or may have had a more positive attitude towards clinical guidelines, and they may have been familiar with the latest evidence in this field. Despite the high baseline score, the effect of the intervention in the peer assessment group was 8.7%. We estimated a minimal important difference in knowledge and guideline consistent reasoning of 5%, based on improvements in professional practice using audit and feedback.[45] An update of this review showed a median improvement of 4.3% for dichotomous outcomes and 1.3% for continuous outcomes.[60] Systematic reviews that assessed the effectiveness of guideline implementation showed improvements in process of care ranging from 5-10%.[73,64,74] The results of our study fall within this range of results.
Fourth, the peer assessment was primarily focused on knowledge and clinical reasoning of individual physical therapists and we did not specifically address organizational and contextual barriers in the peer assessment. However, the participating physical therapists developed a personal improvement plan that allowed for addressing organizational barriers. Fifth, we did not conduct an economic evaluation of the peer assessment program, which could be included in follow-up studies.

In conclusion, our study demonstrates that peer assessment is an effective method to improve guideline knowledge and guideline consistent clinical reasoning. Our findings are a first step toward further use of peer assessment to support the implementation of clinical guidelines and to identify areas where knowledge of guidelines should be improved. More work is needed to assess consistency of results at patient level in clinical practice, and with professionals who are not necessarily prepared to reflect critically on their own performance. Further research should address which aspects of the educational process can be attributed to the results

and to assess the impact on self-reflection. Peer assessment can be integrated in CoPs of other professions as well if well prepared and supervised. Large scale implementation can be explored by teaching expert assessor skills to group leaders within CoPs and development of vignettes for other guidelines.

Abbreviations

MM = Marjo Maas

HE = Henk van Enck

References

1 Grimshaw JM, Thomas RE, MacLennan G, *et al*. Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technol Assess*. 2004;8(6):1-72.

2 Grol R, Wensing M. *Improving patient care; the implementation of change in clinical practice.* London: Elsevier; 2005.

3 Gundersen L. The effect of clinical practice guidelines on variations in care. *Ann Intern Med.* Aug 15 2000;133(4):317-318.

4 Fritz JM, Cleland JA, Brennan GP. Does adherence to the guideline recommendation for active treatments improve the quality of care for patients with acute low back pain delivered by physical therapists? *Med Care.* Oct 2007;45(10):973-980.

5 Rutten GM, Degen S, Hendriks EJ, Braspenning JC, Harting J, Oostendorp RA. Adherence to Clinical Practice Guidelines for Low Back Pain in Physical Therapy: Do Patients Benefit? *Phys Ther.* 2010;90(8):1111-1121.

6 Staal JB, Hendriks EJM, Heijmans M, *et al*. *Richtlijn Lage-Rugpijn voor Fysiotherapie en Manuele therapie [Guideline low back pain for physical therapy and manual therapy]*. Amersfoort, The Netherlands: Royal Dutch Society for Physical Therapy;2010.

7 Rutten G, Kremers S, Rutten S, Harting J. A theory-based cross-sectional survey demonstrated the important role of awareness in guideline implementation. *J Clin Epidemiol.* Feb 2009;62(2):167-176 e161.

8 Cabana MD, Rand CS, Powe NR, *et al*. Why don't physicians follow clinical practice guidelines? A framework for improvement. *JAMA*. 1999;282(15):1458-1465.

9 Cote AM, Durand MJ, Tousignant M, Poitras S. Physiotherapists and use of low back pain guidelines: a qualitative study of the barriers and facilitators. *J Occup Rehabil.* 2009;19(1):94-105.

10 Grol R, Grimshaw J. From best evidence to best practice: effective implementation of change in patients' care. *Lancet*. 2003;362(9391):1225-1230.

11 Fleuren M, Wiefferink K, Paulussen T. Determinants of innovation within health care organizations: literature review and Delphi study. *Int J Qual Health Care*. Apr 2004;16(2):107-123.

12 van der Wees PJ, Jamtvedt G, Rebbeck T, de Bie RA, Dekker J, Hendriks EJ. Multifaceted strategies may increase implementation of physiotherapy clinical guidelines: a systematic review. *Aust J Physiother*. 2008;54(4):233-241.

13 Bekkering GE, Engers AJ, Wensing M, *et al*. Development of an implementation strategy for physiotherapy guidelines on low back pain. *Aust J Physiother.* 2003;49(3):208-214.

14 Li LC, Bombardier C. Physical therapy management of low back pain: an exploratory survey of therapist approaches. *Phys Ther.* 2001;81(4):1018-1028.

15 Fleuren M, Jans M. *Basisvoorwaarden voor de implementatie van de KNGF-richtlijnen*. Leiden: TNO Kwaliteit van Leven; 2008.

16 Harting J, Rutten GMJ, Rutten STJ, Kremers SP. A qualitative application of the diffusion of innovations theory to examine determinants of guideline adherence among physical therapists. *Phys Ther.* 2009;89(3):221-232.

17 Stevens JG, Beurskens AJ. Implementation of measurement instruments in physical therapist practice: development of a tailored strategy. *Phys Ther.* 2010;90(6):953-961.

18 Bekkering GE, Hendriks HJM, Van Tulder MW, *et al*. Effect on the process of care of an active strategy to implement clinical guidelines on physiotherapy for low back pain: a cluster randomised controlled trial. *Qual Saf Health Care*. 2005;14(2):107-112.

19 Rutten GMJ, Harting J, Rutten STJ, Bekkering GE, Kremers SPJ. Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *J Eval Clin Pract.* 2006;12(5):491-500.

20 Swinkels IC, van den Ende CH, van den Bosch W, Dekker J, Wimmers RH. Physiotherapy management of low back pain: does practice match the Dutch guidelines? *Aust J Physiother.* 2005;51(1):35-41.

21 Swinkels RA, van Peppen RP, Wittink H, Custers JW, Beurskens AJ. Current use and barriers and facilitators for implementation of standardised measures in physical therapy in the Netherlands. *BMC Musculoskelet Disord.* 2011;12:106.
22 Speyer R, Pilz W, Van Der Kruis J, Brunings JW. Reliability and validity of student peer assessment in medical education: a systematic review. *Med Teach*. 2011;33(11):e572-585.
23 Beyer M, Gerlach FM, Flies U, *et al*. The development of quality circles/peer review groups as a method of quality improvement in Europe. *Fam Pract.* 2003;20(4):443-451.
24 Grol R. Quality improvement by peer review in primary care: a practical guide. *Qual Health Care.* 1994;3(3):147-152.
25 Grol R. Successes and failures in the implementation of evidence-based guidelines for clinical practice. *Med Care.* 2001;39(8 Suppl 2):II46-54.
26 Falchikov N. *Learning Together. Peer Tutoring in Higher Education.* London: UK: RoutledgeFalmer; 2001.
27 Branch WT, Jr., Paranjape A. Feedback and reflection: teaching methods for clinical settings. *Acad Med.* 2002;77(12):1185-1188.
28 Orsmond P, Merry S, Reiling K. Biology students' utilization of tutors' formative feedback: a qualitative interview study. *Assess Eval High Educ.* 2005;30(4):369-386.
29 Searby M, Ewers T. An evaluation of the use of peer assessment in higher education: a case study in the School of Music, Kingston University. *Assess Eval High Educ.* 1997;22(4):371-383.
30 Hanrahan SJ, Isaacs G. Assessing Self- and Peerassessment: The students' views. *High Educ Res Dev*. 2001;21(1):53-70.
31 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA*. 1993;269(13):1655-1660.
32 Overeem K, Faber MJ, Arah OA, *et al*. Doctor performance assessment in daily practise: does it help doctors or not? A systematic review. *Med Educ.* Nov 2007;41(11):1039-1049.
33 Sargeant JM, Mann KV, Ferrier SN, *et al*. Responses of rural family physicians and their colleague and coworker raters to a multi-source feedback process: a pilot study. *Acad Med*. Oct 2003;78 (10 Suppl):S42-44.
34 Dolmans D, Schmidt H. The advantages of problem-based curricula. *Postgrad Med J.* Sep 1996;72(851):535-538.
35 Zaher E, Ratnapalan S. Practice-based small group learning programs: systematic review. *Can Fam Physician.* 2012;58(6):637-642.
36 Smits PBA, Verbeek J, De Buisonje CD. Problem based learning in continuing medical education: a review of controlled evaluation studies. *BMJ*. 2002;324(7330):153.
37 Adams AS, Soumerai SB, Lomas J, Ross-Degnan D. Evidence of self-report bias in assessing adherence to guidelines. *Int J Qual Health Care*. 1999;11(3):187-192.
38 Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of vignettes, standardized patients, and chart abstraction. *JAMA*. 2000;283(13):1715-1722.
39 Luck J, Peabody JW. Using standardised patients to measure physicians' practice: validation study using audio recordings. *BMJ Qual Saf.* 2002;325(7366):679.
40 Peabody JW, Luck J, Glassman P, *et al*. Measuring the quality of physician practice by using clinical vignettes: a prospective validation study. *Ann Intern Med*. 2004;141(10):771.
41 Hughes R, Huby M. The application of vignettes in social and nursing research. *J Adv Nurs.* 2002;37(4):382-386.
42 Tiemeier H, De Vries WJ, Van Het Loo M, *et al*. Guideline adherence rates and interprofessional variation in a vignette study of depression. *Qual Saf Health Care.* 2002;11(3):214-218.
43 Sandvik H. Criterion validity of responses to patient vignettes: an analysis based on management of female urinary incontinence. *Fam Med*. 1995;27(6):388-392.
44 Hrisos S, Eccles MP, Francis JJ, *et al*. Are there valid proxy measures of clinical behaviour? A systematic review. *Implement Sci*. 2009;4:37.

45 Jamtvedt G, Young JM, Kristoffersen DT, O'Brien MA, Oxman AD. Audit and feedback: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev.* 2006(2):CD000259.
46 Ophey M, Maas M, Beer J. Onderzoek naar de inhoudsvaliditeit van het performance-assessment in de hoofdfase van de bacheloropleiding fysiotherapie [Study on the content validity of the performance assessment in the main phase of the bachelor physiotherapy]. *Dutch Journal of Medical Education.* 2006;25(2):88-95.
47 Campbell SM, Braspenning J, Hutchinson A, Marshall MN. Research methods used in developing and applying quality indicators in primary care. *BMJ.* 2003;326(7393):816-819.
48 Grol R, Cluzeau FA, Burgers JS. Clinical practice guidelines: towards better quality guidelines and increased international collaboration. *Br J Cancer.* 2003;89 Suppl 1:S4-8.
49 Reeves D, Campbell SM, Adams J, Shekelle PG, Kontopantelis E, Roland MO. Combining multiple indicators of clinical quality: an evaluation of different analytic approaches. *Med Care.* 2007;45(6):489-496.
50 Grant AM, Franklin J, Langford P. The Self-Reflection and Insight Scale: A new measure of private self-consciousness. *Soc Behav Pers.* 2002;30(8):821-835.
51 Durgahee T. Facilitating reflection: from a sage on stage to a guide on the side. *Nurse Educ Today.* 1998;18(2):158-164.
52 Austin Z, Gregory PA, Chiu S. Use of reflection-in-action and self-assessment to promote critical thinking among pharmacy students. *Am J Pharm Educ.* 2008;72(3):48.
53 Roberts C, Stark P. Readiness for self-directed change in professional behaviours: factorial validation of the Self-reflection and Insight Scale. *Med Educ.* 2008;42(11):1054.
54 Twisk J. *Applied Multilevel Analysis.* New York: Cambridge; 2006.
55 Deuning CM. Percentage vrouwelijke fysiotherapeuten 2010 [percentage of female physical therapists in 2010]. 2011. http://www.zorgatlas.nl/zorg/eerstelijnszorg/paramedische-zorg/percentage-vrouwelijke-fysiotherapeuten#breadcrumb. Accessed November 2011.
56 Hingstman L, Kenens RJ. *Cijfers uit de registratie van fysiotherapeuten: peiling 1 januari 2010 [Statistics from the registration of physiotherapists: survey 1 January 2010].* Utrecht: Nivel;2011.
57 Baker R, Camosso-Stefinovic J, Gillies C, *et al*. Tailored interventions to overcome identified barriers to change: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev.* 2010(3):CD005470.
58 Rebbeck T, Maher CG, Refshauge KM. Evaluating two implementation strategies for whiplash guidelines in physiotherapy: a cluster randomised trial. *Aust J Physiother.* 2006;52(3):165-174.
59 Bekkering GE, van Tulder MW, Hendriks EJ, *et al*. Implementation of clinical guidelines on physical therapy for patients with low back pain: randomized trial comparing patient outcomes after a standard and active implementation strategy. *Phys Ther.* 2005;85(6):544-555.
60 Ivers N, Jamtvedt G, Flottorp S, *et al*. Audit and feedback: effects on professional practice and healthcare outcomes. *Cochrane Database Syst Rev.* 2012;6:CD000259.
61 Smeele IJ, Grol RP, Van Schayck CP, Van Den Bosch WJ, Van Den Hoogen HJ, Muris JW. Can small group education and peer review improve care for patients with asthma/chronic obstructive pulmonary disease? *Qual Health Care.* 1999;8(2):92.
62 Dannefer EF, Henson LC, Bierer SB, *et al*. Peer assessment of professional competence. *Med Educ.* 2005;39(7):713-722.
63 Grimshaw J, Eccles M, Tetroe J. Implementing clinical guidelines: current evidence and future implications. *J Contin Educ Health Prof.* 2004;24 Suppl 1:S31-37.
64 Grimshaw J, Eccles M, Thomas R, *et al*. Toward evidence-based quality improvement. Evidence (and its limitations) of the effectiveness of guideline dissemination and implementation strategies 1966-1998. *J Gen Intern Med.* 2006;21 Suppl 2:S14-20.
65 Chow AYM, Lam DOB, Leung GSM, Wong DFK, Chan BFP. Promoting Reflexivity among Social Work Students: The Development and Evaluation of a Programme. *Soc Work Educ.* 2011;30(2):141-156.

66 Carr SE, Johnson PH. Does self reflection and insight correlate with academic performance in medical students? *BMC Med Educ.* 2013;13:113.

67 Lowe M, Rappolt S, Jaglal S, Macdonald G. The role of reflection in implementing learning from continuing education into practice. *J Contin Educ Health Prof.* Summer 2007;27(3):143-148.

68 Gould D. Using vignettes to collect data for nursing research studies: how valid are the findings? *J Clin Nurs.* 1996;5(4):207-212.

69 McGlynn EA, Asch SM, Adams J, *et al*. The quality of health care delivered to adults in the United States. *N Engl J Med.* 2003;348(26):2635-2645.

70 Hofhuis H, Plas M, Ende Evd. *Eindevaluatie van het programma Implementatie Kwaliteitsbeleid Paramedische Zorg (IKPZ): beschrijving van de kwaliteitssystemen van de paramedische beroepsgroepen in 2003 [Final evaluation of the program implementation quality in allied medicine (IKPZ): description of the quality of the allied professions in 2003].* Utrecht: Nivel;2004.

71 Forsetlund L, Bjorndal A, Rashidian A, *et al*. Continuing education meetings and workshops: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev.* 2009(2):CD003030.

72 Farmer AP, Legare F, Turcot L, *et al*. Printed educational materials: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev.* 2008(3):CD004398.

73 Grimshaw JM, Thomas RE, MacLennan G, *et al*. Effectiveness and efficiency of guideline dissemination and implementation strategies. *Health Technol Assess.* 2004;8(6):1-72.

74 Grol R. Improving the quality of medical care: building bridges among professional pride, payer profit, and patient satisfaction. *JAMA.* 2001;286(20):2578-2585.

Chapter 4
Critical features of peer assessment of clinical performance to enhance adherence to a low back pain guideline for physical therapists: a mixed methods design

Marjo Maas
Simone van Dulmen
Greetje Sagasser
Yvonne Heerkens
Cees van der Vleuten
Ria Nijhuis-van der Sanden
Philip van der Wees

BMC Medical Education, 2015;15(1):203

Abstract

Background

Clinical practice guidelines are intended to improve the process and outcomes of patient care. However, their implementation remains a challenge. We designed an implementation strategy, based on peer assessment (PA) focusing on barriers to change in physical therapy care. A previously published randomized controlled trial showed that PA was more effective than the usual strategy "case discussion" in improving adherence to a low back pain guideline. PA aims to enhance knowledge, communication, and hands-on clinical skills consistent with guideline recommendations. Participants observed and evaluated clinical performance on the spot in a role-play simulating clinical practice. Participants performed three roles: physical therapist, assessor, and patient.

Aim

To explore the critical features of the PA program that contributed to improved guideline adherence in the perception of participants.

Setting and participants

Dutch physical therapists working in primary care (n=49) organized in communities of practice (n=6).

Methods

By unpacking the PA program we identified three main tasks and eleven subtasks. After the program was finished, a questionnaire was administered in which participants were asked to rank the program tasks from high to low learning value and to describe their impact on performance improvement. Overall ranking results were calculated. Additional semi-structured interviews were conducted to elaborate on the questionnaires results and were transcribed verbatim. Questionnaire comments and interview transcripts were analyzed using template analysis.

Results

Program tasks related to performance in the therapist role were perceived to have the highest impact on learning, although task perceptions varied from challenging to threatening. Perceptions were affected by the role-play format and the time schedule. Learning outcomes were awareness of performance, improved attitudes towards the guideline, and increased self-efficacy beliefs in managing patients with low back pain. Learning was facilitated by psychological safety and the quality of feedback.

Conclusion

The effectiveness of PA can be attributed to the structured and performance-based design of the program. Participants showed a strong cognitive and emotional commitment to performing the physical therapist role. That might have contributed to an increased awareness of strength and weakness in clinical performance and a motivation to change routine practice.

Background

Clinical practice guidelines are intended to optimize patient care and improve patient outcomes.[1] Guidelines are also increasingly regarded as a part of professional quality systems and policies.[2] However, the uptake of guidelines in physical therapy (PT) practice remains a challenge, despite the variety of implementation strategies that have been developed.[3–5] Professionals are hampered by a lack of commitment to the guidelines, insufficient knowledge and skills related to the guidelines, and limited social and organizational support.[6-8] In addition, a study by Rutten *et al*.[9] on determinants of guideline adherence showed that physical therapists (PTs) do not hold realistic perceptions of the extent to which they adhere to guideline recommendations.

The limited ability of clinicians to accurately self-assess the quality of their professional performance is not new.[10] A compelling body of research evidence shows that the development of adequate self-perception requires both internal and external information about one's professional performance, including appropriate performance standards.[11-15] There is a need for interventions containing feedback that can help to develop realistic self-perceptions of guideline adherent behavior and enhance motivation to change routine practice.

We designed an implementation strategy based on peer assessment (PA) that targets identified barriers to change for PTs in primary care.[16] We tailored an existing PA design that was shown to be effective in undergraduate PT education[17] to the context of professional PT practice and to the purpose of guideline implementation. In a previously published randomized-controlled trial (Table 1), PA was shown to be more effective than the traditional "case discussion" implementation strategy.[18] We analyzed this PA program to determine the critical features of its success.

In PA professionals evaluate or are being evaluated by observing their peers in a role-play that simulates PA practice. They provide each other with performance feedback that might evoke reflection

Table 1 — Overview of the methods and results of a previously published trial (Van Dulmen *et al.*)[18]

Design

A cluster-randomized controlled trial was conducted among 10 communities of practice (CoPs) of Dutch physical therapists (n=90) to compare the effectiveness of two implementation strategies: peer assessment (PA) and case discussion (CD). Both strategies aimed to improve adherence to the clinical practice guidelines for the management of patients with low back pain. The programs consisted of four meetings over a six-month period. Outcomes were measured at baseline and at 6 months follow up.

Randomization and intervention allocation

CoPs showing interest in the program were invited to a plenary meeting in November 2009. They were informed that the study compared two educational strategies, and that both programs required an equal amount of time and effort. All physical therapists regularly treating patients with low back pain were eligible for inclusion. Included CoPs were randomly allocated to the PA group and the CD group resulting in six CoPs for the PA program (n=49) and four CoPs for the CD program (n=41).

Interventions

PA is the process whereby professionals evaluate or are being evaluated by their peers and provide each other with performance feedback. The main difference between PA and CD is that in the PA approach the tasks were structured, with a focus on performance rather than discussion, and participant roles were pre-defined. In the CD approach the tasks were less structured with ample opportunity for in-depth elaboration and discussion, and participant roles were not defined. In PA and CD, participants worked on identical cases concerning problem content, but for PA these cases were adjusted to allow for performance of participants in different roles. In PA, written cases were not known in advance but were presented by a coach on the spot, simulating daily clinical practice. For CD groups, written cases were included in the program guide to allow for proper preparation, along with instructions and written questions to guide the discussion process.

Outcome measures

Outcomes were assessed at baseline and at six months. Primary outcome was knowledge and guideline-consistent reasoning, measured with 12 performance indicators using four vignettes that fully covered the patient profiles described in the guidelines. Changes in reflective practice were measured with the Self-Reflection and Insight Scale (Grant *et al.*).[49]

Results

Multilevel analysis showed an increase in guideline-consistent clinical reasoning of 8.4% in the PA groups whereas the control groups showed a decline of 0.1% (estimated group difference 8.7%; [95%CI: 3.9 to 13.4; $P \leq .001$]). No group differences were found for self-reflection.

and identify areas of clinical performance that need improvement.[19,20] Personal assumptions about one's professional competence can be compared with peer views that might compensate for poor self-assessment.[13,14] PA enhances the development of a mutually accepted quality standard of performance by introducing peers to an "assessor" or "auditor" perspective.[23,26] In this respect, PA might be a an effective tool to enhance bottom up quality improvement and accountability of health care.[21]
Research shows that effective PA practices are context-specific and culture dependent,[23,24] and these findings also apply to effective implementation strategies.[25] Thus, to enhance the generalizability of the trial results, and to allow for adequate knowledge transfer, understanding of the causal mechanisms of PA is necessary.[25-27]
The aim of this study was to explore the features of the PA program that were perceived to have a powerful impact on learning and change of routine practice.

Our research question was: *Which elements of the PA program were perceived to have a strong impact on clinical performance improvement consistent with clinical guidelines, and why?*

Methods

Study Design

We conducted a mixed-methods study using questionnaires and semi-structured interviews to explore the critical features of the PA program that contributed to improved guideline adherence.

Setting and participants

The Royal Dutch Society for Physical Therapy offers annual professional development programs for the approximately 800 communities of practice in the Netherlands. Communities of practice are small groups of 5-15 PTs who share the same setting or the same interests. The current study focused on communities of practice (n=6; 44 participants) that participated in a randomized controlled trial (Table 1) and were allocated to the PA-condition.

The Peer Assessment Program

The PA program was launched in February 2010 and finished in September 2010. Its design was built on a mix of theoretical constructs related to learning and professional behavior change, which were assumed to contribute to improved clinical performance.[26] Table 2 shows the theoretical framework, the underlying constructs, and

the operationalization of these constructs in the PA design.
The PA program aimed to enhance clinical performance consistent with guideline recommendations including knowledge, communication, and hands-on clinical skills. Clinical performance was directly observed and evaluated by peers in a role-play that simulated clinical practice. Participants received a PA-manual in advance, containing a description of the PA-procedure, a time schedule for each meeting, and guidelines for receiving and providing constructive feedback. They received a link to the updated guideline "Low back pain for physical therapy and manual therapy" (Staal *et al.*)[28] published by the Royal Dutch Society for Physical Therapy. Four meetings were scheduled over a period of six months. As the PTs were novices in the PA method, and no additional training was provided, the process was supported by a coach (MM or HE). Coaches were experienced PTs, teachers in PT education, and trained in the PA procedure. They facilitated the process of providing and receiving feedback, and they gave additional feedback when needed.
Each participant performed three roles: PT, assessor and simulated patient. In the PT role, participants completed a written assignment that contained a clinical case and brief instructions for diagnosis or treatment. Clinical cases were developed by a team of experienced PTs and guideline experts. The cases fully covered the patient profiles of low back pain described in the guidelines, including red flags. PTs analyzed the clinical cases by reasoning aloud and demonstrated (hands-on) skills relevant to the clinical problem. Afterwards, they reflected on their performance. In the assessor role, peer performance was observed and assessed with a scoring sheet containing performance criteria that could be scored on a 7-point scale (1= much improvement needed, to 7= no improvement needed) and space for written feedback. Performance categories addressed diagnosis, treatment, and evaluation. In the patient role, participants received the clinical case along with written simulation instructions. Simulation instructions consisted of a description of the patient's complaints, including personal factors (e.g., cognitive / emotional), and contextual factors (e.g. family, work) that might be relevant to the patient's problem. Participants were instructed to improvise patient responses and provide feedback from the patient perspective.
Prior to the third session, each participant developed a personal change plan, including an action plan, based on performance feedback and self-assessment. In the third meeting, the group reviewed change plans and provided additional peer feedback. The fourth session was identical to the first two sessions, but the design of the clinical cases was tailored to participants' specific learning needs.

Table 2 — Theoretical framework of the PA program design

Theory	Underlying constructs used	Operationalization of constructs
Social constructivist learning theory[48]	Contextual learning, collaborative learning, active participation, and knowledge construction to enhance attention, storage, and retrieval of knowledge from memory.	Presenting a variety of clinical problems that adequately reflect authentic clinical practice, accounting for the case-specifity of clinical competence. Simulating the context of daily practice in a role-play accounting for the context-specifity of clinical competence. Enhancing active participation of each participant by assigning pre-defined roles, and by using a performance based format.
Self-regulated learning theory[50,51]	Applying metacognitive strategies to guide the professional development process.	
	Self-assessment	Designing an improvement plan based on peer feedback.
	Conscious goal setting and action planning	Discussing the improvement plan with peers.
Situated learning theory[40,52]	Learning in the context of daily practice to bridge the gap between learning context and application context.	Delivering the program within communities of practice that share the same setting or the same interest.
Social cognitive learning theory[33]	Enhancing the development of self-efficacy beliefs, by: Performing the new behavior and experiencing the consequences of that behavior (mastery experience).	Performing the new behavior individually, by reasoning aloud and demonstrating diagnostic and treatment skills relevant to the LBP guide lines.
	Observing the behavior of others and the consequences of that new behavior (vicarious experience).	Observing a peer's performance and providing individualized improvement feedback.
Stages of change theory[53]	Alligning implementation strategies to the stages of change.	Delivering the implementation program within communities of practice. Peers are involved in the professional development process and are capable of tailoring feedback to stages of change.
Theory of planned behaviour[34]	Changing attitudes and subjective norm toward the new behavior. Enhancing the development of self-effecacy beliefs.	Introducing peers to the assessor perspective. In appraising a peers' performance, peer assessors need to develop an understanding and a mutually accepted quality standard to deliver credible performance feedback.

Questionnaires and interviews

Prior to data collection, we unpacked tee PA program and identified three main tasks and eleven subtasks that were assumed to affect guideline adherence. Task analysis was supported by guidelines described by Janssen-Noordman *et al*.[29] An online questionnaire was administered after completion of the PA program in which participants were asked to rank the program tasks from high to low learning value, assigning the highest rank for the most learning value and the lowest rank for the least. Subsequently, they were asked to provide written comments on the three most instructive PA task elements (Appendix).

Emerging questions from the questionnaire comments served as input for conducting semi-structured interviews to obtain more understanding of how the PA program affected professional development. In contrast to a reductionist approach to the data by means of task analysis and task ranking, the interviews had a more holistic approach, focusing on experiences with the PA program as an integrated system. From each peer group, one participant was selected for an interview (n=6). Purposeful selection was based on average and deviant ranking results. An interview guide (Appendix) was designed by MM and PW addressing the three main questions that emerged from the questionnaire data:

1 What did you expect of the peer assessment program?
2 How did you perceive the peer assessment program, and how did it affect your daily practice?
3 In the questionnaire, you indicated that you perceived task X, Y and Z to have the strongest learning value. Can you explain why?

Selected participants were invited by e-mail, and received information about the study's purpose, procedure, the use of the data, and the focus of the interview.

The first interview was conducted by MM and PW face-to-face. The following interviews were conducted by either MM or PW using teleconferencing technology. To enhance the credibility of the results, research assistants AS and GB joined the telephone interviews, taking notes and posing additional questions when needed. Interviews were audiotaped after informed consent was obtained from each participant. Interviews lasted between 45 and 90 minutes. Recordings were transcribed verbatim. An independent check on the transcripts was conducted by AS and GB.

Data analysis

Quantitative analysis

Ranking results were described by calculating mean, median, and sum scores for each learning task using IBM SPSS statistics 20.

Qualitative analyses

A sample of questionnaire comments and interview transcripts was studied and coded by MM and PW independently. The analytic process was guided by template analysis that combines a-priori codes with emerging codes.[30] The PA program as a whole and its tasks and subtasks served as a-priori codes. Additional codes were defined during the analytic process when these seemed relevant regarding the research question. Codes were compared, and some codes were merged into higher-order codes. PW and MM discussed a codebook until consensus was reached. Subsequently, all written comments in the questionnaires and interview transcripts were analyzed line-by-line, using ATLAS-ti v.7 software. Emerging themes were identified by constant comparison of codes and higher order codes. We summarized the results in a matrix that crossed a-priori codes (tasks and subtasks) and emerging themes from the data.[31] Two independent researchers SD (health scientist and PT) and MS (educational scientist) evaluated the analysis process and outcomes. They were not involved in the design or delivery of the PA program. Disagreements were discussed until consensus was reached and we finally agreed that the matrix fully fitted the data.

Ethical aspects

This project received approval of the medical ethical committee of Radboud University Medical Center. All participants volunteered to participate and gave their informed consent. We adhered to the RATS guidelines for qualitative research.[32]

Results

In total, 44 PTs participated in the program. Table 3 shows an overview of the participants' characteristics. Two PTs did not fully complete the ranking procedure and were excluded from quantitative analyses (response rate = 95%). All PTs invited for additional interviews (n=6) agreed to participate.

Table 3 — Peer assessment group characteristics

Physical therapist characteristics	N = 44
Age mean (SD)	40.4 (12.4)
Sex (male/female)	17/27
Working hours per week (SD)	32.5 (9.6)
Treatment of patients with LBP per year	
<25	12
25-50	12
50-75	6
76-100	5
>100	10
Manual therapist	8
Years of experience (SD)	16.5 (11.9)

Table 4 — Results quantitative analysis

Tasks	Subtasks	Mean	Median	Range	Sum
Study manual	Study PA procedure and guidelines	5.09	6.0	10	195
Perform in PT role	Perform clinical task individually	8.05	9.0	10	322
	Receive peer feedback	9.75	10.0	6	389
	Receive external coach feedback	8.48	9.0	10	331
	Receive simulated patient feedback	6.84	7.0	9	253
	Receive written feedback and scores	2.91	2.0	9	102
Perform in assessor role	Observe peer performance	6.46	6.0	9	252
	Provide oral feedback	5.75	5.5	9	230
	Provide written feedback and scores	2.58	2.0	4	44
Design change plan	Design and discuss change plan	6.38	6.00	10	249
Perform in patient role	Simulate patient problem	3.26	3.0	7	98

Results quantitative analysis

Ranking results showed that participants committed the most to subtasks related to task performance in the PT role. Receiving peer feedback was perceived as the most valuable element, followed by receiving external coach feedback, performing the clinical task individually, and receiving simulated patient feedback. Participants varied widely in their preferences for learning in the PT role, but agreed on the superior value of receiving peer feedback. Table 4 shows an overview of the results.

Results qualitative analysis

Five themes emerged from the analysis of the questionnaires comments and the additional interview transcripts. These themes were related to the PA program either as a whole, or related to its specific learning tasks and subtasks: a) general perceptions of the PA program, b) determinants of PA affecting perceptions, c) facilitators for learning, d) learning activities, and e) learning outcomes.
We summarized the results by creating a matrix that crossed a-priori categories (program tasks and subtasks) with emerging themes, leaving empty fields where data were not available (table 5). Program tasks and subtasks in the matrix follow the build-up of the PA program. In the next section, we first discuss the general perceptions of the PA program, determinants of PA affecting these perceptions, and the general outcomes. Second, we discuss the subtasks by following the matrix, including their related learning activities, outcomes, and facilitators for learning. Although we did not explicitly ask participants to comment on tasks that were perceived as less instructive, they often did so spontaneously:

"Receiving feedback from your colleagues provides new insights. You learn from the mistakes you make, or how you can handle them better. I assigned the lowest ranks to 'receiving and providing scores' because I think that scores add nothing to the learning process. Moreover not all aspects of performance can be expressed in scores and scores are not objective" (Q-P8).

We limit the discussion to comments on the most instructive subtasks. Participants' quotes are coded by information source (Questionnaire = Q; Interview Transcript = IT) and by participant number (P1 – P42). Table 5 provides a summary of the of the qualitative analysis results.

Table 5 — Summary of results qualitative analysis

	PA Program tasks and subtasks	Perceptions of the PA program	Determinants of PA affecting perceptions	Facilitators for learning and change	Learning Processes	Learning Outcomes
	PA Program					Change in attitudes toward guidelines. Awareness of professional limitations.
1	Study manual					Update of knowledge.
2	Perform task in PT role	Fear to expose professional competence. Challenge of obtaining performance feedback.	Tight time schedule. Role play format.	Training in the PT role. Group safety	Uncovers weakness. Reinforces strength. Stimulates reasoning aloud, self-assessment and critical reflection.	
3	Receive peer feedback			Peer feedback is concrete, concise, critical and personal. Varied group composition.	Reveals strength and weakness. Shows improvement areas. Reveals new reasoning perspectives and performance alternatives. Stimulates self-assessment and critical reflection.	Awareness of gaps in professional performance.
4	Receive simulated patient feedback				Reveals how interventions are perceived from the patient perspective.	Improved self-confidence in arguing for choices.
5	Receive external coach feedback			External coach poses challenging questions, guides the PA process, facilitates giving and receiving feedback, provides non-judgmental, concise feedback, monitors the time schedule, maintains group safety.	Reveals new reasoning perspectives and performance alternatives. Stimulates self-assessment and critical reflection.	Improved self-efficacy beliefs in managing LBP* patients.
6	Receive written feedback and scores				Stimulates self-assessment and critical reflection.	
7	Observe peer performance			Modeling peer performance.	Reveals new reasoning perspectives and performance alternatives.	Improved self-confidence in managing LBP patients.
8	Provide oral feedback			Training in the assessor role.	Triggers being concrete and concise in reasoning aloud. Elicits discussion over criteria.	Shared quality standards of performance.
9	Provide written feedback and scores					
10	Design change plan				Guides improvement process.	
11	Perform task in Simulated patient role					

*LBP: low back pain

The PA program as a whole

General perceptions

Participants were generally satisfied with the program. They reported that the mix of written cases adequately reflected the problems encountered in daily practice, however, the PA format was new, and was perceived with mixed feelings. Physical therapists were not used to exposing their professional performance for group review. Some participants appraised the PA program as challenging, providing an excellent opportunity to receive performance feedback; others were reluctant to expose their professional competence, triggered by feelings of performance anxiety.

Specific task features (time schedule and role-play format) affected perceived learning opportunities and threats. Participants, who appreciated the task structure, reported that PA allowed them to solve a considerable number of clinical cases in a relatively short time and trained them to be concrete and concise in reasoning aloud in the PT role as well as in the assessor role.

"The strongest feature of PA was the structure of the meetings. The system of PA was interesting ... for example, I appreciated that repeating feedback that was provided by someone else, was not allowed. It's useless to repeat advice." (IT-P41)

Participants who criticized the task structure perceived the timetable as stressful, and as a barrier to in-depth case discussion.

"Yes, time pressure was a weakness of PA … sometimes the performance evaluation raised questions which could not be addressed in-depth, because you had to skip to a new problem. I would prefer to perhaps discuss fewer cases more extensively." (IT-P18)

From the perspective of the assessor, the role-play was appreciated because it allowed implicit behaviors to become explicit. From the perspective of the assessed, the role-play was critically appraised. Some participants believed that it poorly reflected their authentic professional behaviors, and that they underperformed in the PA context.

"It was hard to perform a clinical examination or treatment in this setting; partly, because the patient is a colleague. It is not like in your own working room. In addition, you consciously think about the decisions you make, because your steps will be evaluated." (Q-P8)

General learning outcomes

The PA program resulted in distinct levels of self-reported behavioral change. Although participants studied the updated guidelines prior to the program and were tested on their knowledge with clinical vignettes, they reported that applying knowledge in the context of PA increased their understanding of the guidelines, and facilitated their use in clinical practice.

"Yes, you want to work according to the guidelines. Therefore, you need to master them … I realized that I in fact did not fully understand the guidelines for low back pain. I knew vaguely what the content was, but not exactly. I think I have obtained a better understanding of the classification system of patient profiles, and therefore I apply them more frequently in my work." (IT-P18)

Participants noticed that working with the guidelines in the context of the PA program changed their attitudes towards the guidelines. In their view, guidelines are often considered as too theoretical and of limited applicability in daily practice.

"I also noticed that some colleagues perceived the guidelines as less annoying or boring." (IT-P18)

Although participants did not explicitly report changes in their management of patient problems, they did report changes in their professional identity and awareness of the limitations of their profession.

"What clearly emerged from the cases we discussed [in the PA program] was that as a PT we like to help people and it remains questionable if that is always justified? We somehow suffer from an irrepressible desire to help … we're inclined to always give care, whereas in some cases restraint would be better." (IT-P14)

Performing the PT role

Performing the clinical task individually

Although some participants initially felt reluctant to move out of their "comfort zone", they considered exposure of their routine practice as a necessity for quality improvement. They pointed out that the four PA sessions allowed them to cope with anxiety triggers by training in the PT role.

"Yes, but you need to push yourself sometimes. I mean … I think it's threatening, it's not pleasant at all … but I also know that it is important to bare your buttocks, and look where you go wrong. No pain no gain, that's a bit of the rationale." (IT-P15)

Performance in the PT role necessitated reasoning aloud, triggered underpinning clinical decisions, and stimulated the transfer of research evidence to the context of a particular clinical problem. Participants explained that arguing aloud resulted in improved self-confidence in decision-making. They became more aware of their strengths and weaknesses, either by "reflection in action" or by "reflection on action".
Exposing professional performance in the PT role was facilitated by perceived group safety.

"Your colleagues are the people who know you well and who know what your strengths and your weaknesses are. So they may well shoot at you." (IT-P18)

Receiving peer feedback
Although PTs organized in communities of practice discuss clinical cases on a regular basis, they do not have a culture of asking for and providing performance feedback. The opportunity to receive peer feedback was therefore embraced. Participants felt strengthened in areas of clinical performance they mastered, and felt challenged to appraise areas that needed improvement.

"Receiving peer feedback clearly revealed my strengths and weaknesses. I immediately understood what I needed to work on. And because my strengths were noticed, it was easier to face my weaknesses." (Q-P7)

Learning from peer feedback was facilitated by its quality. Participants preferred personalized feedback, that showed involvement with their development process and their personal learning needs, but feedback should also focused.

"I don't mind when someone criticizes me … of course I like to know if I'm doing right, but I'd rather know what I can improve, and how." (IT-P18)

Another facilitating factor was the heterogeneity in group composition. Differences in age and specialization allowed for different approaches to health problems and different models of reasoning. Because

feedback providers were encouraged to clarify improvement feedback with clear examples of desired behavior, they discovered new reasoning perspectives and performance alternatives.

"For example, we have a specialist in haptonomy in our team, and he brings in new perspectives on health problems … I profit from his views in my daily practice. For example, I try to keep the global overview instead of focusing on a single vertebra. As a manual therapist I tend to focus on the details and lose the whole picture." (IT-P14)

Receiving external coach feedback

In contrast to peer feedback, participants attributed the value of coach feedback to its objectivity, conciseness, and perceptiveness, rather than to its involvement with individual peers.

"Well, the coach had an objective approach. The feedback was very practical and well summarized. Nothing more, nothing less and because the coach was new, feedback was perceived to be more objective. I also noticed that the coach was able to discover strengths in all participants." (IT-P2)

However, from the PT-role perspective, the presence of the coach raised performance stress in some cases.

"We also needed to get used to her [coach]. At least, that applied to me. You need to feel a kind of safety with each other to show openly what you think and what you do. We share this safety in our group, and that allows us not to mince words. But with a strange person here, the threshold is higher, at least in my opinion." (IT-P1)

Facilitating behaviors from the coach included posing critical questions rather than giving straightforward answers, fostering a safe learning environment, monitoring the structure and the time-schedule of the PA process, facilitating peer feedback delivery, and strengthening group learning. Participants rejected too much interference of the coach and judgmental coach feedback.

Receiving simulated patient feedback

Participants varied in their appreciation of simulated patient feedback, referring to the limitations of role-play. Despite its limitations, participants valued the different perspective of patient feedback.

"While performing the assignment, I noticed that I was not always providing clear information … I previously never thought about that … I have learned now that I need to communicate more carefully, for instance when giving bad news." (Q-P12).

Performing the assessor role

Observing a peer's performance

Participants reported that the role of assessor allowed them to mirror and model the observed performance to their own intended performance.

"I found observing a peer's performance very instructive because you often imagine how you would handle the situation. When you see how your colleague deals with a problem, you critically reflect on your own choices." (Q-P19)

Appraising the performance of a peer was not a common practice. Participants would rather discuss than assess the observed behaviors. Giving instructive feedback (according to the feedback guidelines) was perceived as difficult. It required clear reasoning strategies, arguing for quality standards of performance, and the courage to be critical.

"Your own feedback should be carefully considered. You must clearly explain why you do or don't agree with the feedback of your colleagues." (Q-P20)

Discussion

This study aimed to explore the critical features of a PA program that was shown to be effective in a previously published randomized controlled trial. The results clearly show that participants committed the most to learning tasks related to performance in the therapist role: performing the task, receiving peer feedback, external coach feedback, and simulated patient feedback. Participants varied widely in the perceived learning value of subtasks related to performing the PT role, but agreed on the superior value of receiving peer feedback. In the next section, we will elaborate on these results.
These results point to the importance of exposing observable behavior (PA) rather than expressing intended behavior (case discussion). Although exposure was associated with feelings of discomfort and performance stress, its impact on awareness of professional

development was not questioned. This raises the question of how feelings of discomfort and stress can affect learning and change in professional practice.

In the PT role, participants needed to make the transfer from implicit reasoning to explicit reasoning and from intentional behavior to observable behavior to allow for assessment and feedback. Bandura's social cognitive theory emphasizes that exposure is conditional to the development of mastery experiences, and mastery experiences are the most important source of information for the development of self-efficacy beliefs. In turn, self-efficacy beliefs contribute significantly to performance improvement and motivation to change.[33] This notion is supported by the theory of planned behavior.[34] Bandura also points to the importance of the peer group in strengthening self-confidence through "vicarious" experiences provided by social models. The impact of modeling on perceived self-efficacy is strongly influenced by perceived similarity to the models (peers) and is considered to be more powerful than performance feedback.[35] Increased self-confidence might have helped participants to approach difficult tasks as challenges to be mastered rather than as threats to be avoided.

The foregoing explains how PA participants succeeded in raising self-efficacy beliefs despite feelings of performance stress, but does not explain why they showed superior test results on clinical vignettes in the trial (Table 1). High arousal levels are generally considered to have a negative impact on the quality of performance according to the Yerkes-Dodson law,[36] and PA participants' experiences supported that, as they contended that they had underperformed in the PA context. However, they must have processed the information in a way that enhanced retrieval and transfer of knowledge to the context of clinical vignettes. Studies addressing the influence of emotion on cognitive processing provide an explanation for this apparent contradiction. McConnel & Eva[37] conducted a literature review on the impact of emotion on the transfer of clinical knowledge and skills. They conceptualized emotion by two dimensions: *valence* and *arousal*. Valence refers to the emotional state (e.g. positive or negative). Arousal refers to the level of activation. One of the findings was that emotional experiences are more likely to be mulled over than non-emotional experiences. This unintentional retrieval of emotional events might have strengthened memory traces of PA participants and facilitated the transfer to new clinical problems. Another view is presented by regulatory focus theory,[38] which contends that receptiveness to feedback depends on "emotional arousal" rather than "emotional valence". Summarizing these

considerations, the critical feature of PA might be attributed to the emotional involvement (either negative or positive) with performing the PT role. As feelings of failure do not contribute to the development of self-efficacy beliefs,[33] successful PA implementation should allow for coping with performance stress within or between the sessions. Training in the PT role and a safe learning environment might be crucial to enable the coping process.

Performance in the assessor role was perceived as a less powerful learning experience. However, it should be noted that the assessor role and the PT role cannot be considered as independent. Observing peer performance allowed observers to model the observed behavior, which might have contributed to reducing performance stress and triggering performance improvement. On a more unconscious level, participants might have profited from the activity of the mirror neuron system that is capable of shaping the observed behavior to a virtual image of their intended behavior.[39] In appraising their peers' performance, assessors needed to reason aloud, compare personal views with group views, and discuss performance standards. This may have provided peer assessors with the missing data for informed self-assessment.[20]

Regarding the role of the external coach in providing feedback, participants ranked peer feedback higher than coach feedback although coach feedback was valued because of its objectivity, its conciseness, and its receptiveness. A comparable study on PA in undergraduate PT education, in which students were asked to rank similar learning tasks, showed that students preferred teacher feedback to peer feedback.[17] Professionals did not question the quality of peer feedback compared to coach feedback, but emphasized the importance of peers being involved in their professional development process. This finding is supported by situated learning theory,[40,41] which contends that the transfer of knowledge is hampered by the gap between the learning context and application context. Delivering the implementation program within communities of practice allows for co-constructing and tailoring knowledge to the personal learning needs.[41] In this respect, the coach remained an outsider.

Although the PA program was successful regarding its aim, the adoption of the program for knowledge transfer purposes should be carefully considered.

Firstly, some participants argued that the role-play format did not adequately reflect their authentic professional behaviors. This view is understandable, but compared to passive guideline dissemination, role-play aims to facilitate the transfer of scientific evidence to clinical

practice, which it did, according to participant reports. As regards the use of peer role-play (low fidelity simulation) compared to standardized patients (high fidelity simulation), research in undergraduate education shows that both tools provide a psychological safe area of practice, where mistakes are not critical.[42] Studies on student perceptions show that standardized patients are perceived as more effective than peers.[43,44] However, research evidence on learning outcomes remains inconclusive.[44,45] Compared to direct observation (work-place based assessment), the role-play format allows for standardizing the content of interest, creating an adequate case mix, and describing the key-features of health problems relevant to the guidelines.[46] Considering constraints in time and costs, peer role-play is the most feasible method. This conclusion is supported by a systematic review undertaken by Overheem *et al.*[47], who evaluated the feasibility and effectiveness of six methods to assess physician performance.

Secondly, some participants perceived the tight time schedule as stressing and preventing in-depth elaboration of the clinical problems. The PA program was designed to enhance the transfer from the learning context to the application context, as the transfer from one problem to another problem.[48] Yet, in an attempt to solve all the presented problems within time limits, the approach to learning might have been too superficial.

Thirdly, performance in the PT role was perceived as challenging and sometimes even threatening. When conditions of psychological safety are not met, the effectiveness of PA might be questioned.[14]

Strengths and limitations

This study provided rich data and convincing results. Because we clearly described the program design, its underlying theoretical constructs, and the critical features of successful guideline implementation, future program designers may profit from our results.

It can be argued that a limitation of the PA approach is the role-play of peers simulating patients. Although the choice of peers instead of standardized patients was defensible as argued above, and although the results show that their feedback was valued, additional training in the patient role might have increased the fidelity of the peers' performance.

Another limitation concerns the questionnaire and the interview guide. Questionnaire comments were reduced by the three tasks with the highest-ranking results. We compensated for this limitation by interviewing participants with contrasting ranking results. Nevertheless, because we did not focus on less instructive tasks

in our interviews, we might have lost information that would have underpinned our results.
Finally, the generalizability of our results might be limited because all participants in this study were Dutch. Research shows that effective PA practices are culture dependent.[23,24]

Conclusions

The effectiveness of PA can be attributed to the structured and performance-based design of the program. Participants showed a strong cognitive and emotional commitment to performing the tasks related to the physical therapist role. That might have contributed to an increased awareness of strengths and weaknesses, and a motivation to change routine practice in the management of patients with low back pain.
Conditional to successful implementation is an environment where mistakes can easily be made, but in which the self-confidence of participants remains undamaged. Adjustment of the tight time schedule and the number of cases, providing more time to elaborate on problems and to recuperate from experiences, might improve the PA task design. However, attempts to improve the effectiveness of PA should not be limited to the modification of the PA tool. We recommend a shift in the feedback culture of PTs in primary care, from avoiding performance feedback to actively seeking feedback. Future research should address the feasibility of PA as a tool to enhance bottom-up quality improvement and accountability to external stakeholders of PT care.

Abbreviations

PW = Philip van der Wees

MM = Marjo Maas

SD = Simone van Dulmen

MS = Margaretha Sagasser

References

1 Grol RP, Wensing M, Eccles MP, Davis DA, (Eds). *Improving Patient Care: The Implementation of Change in Health Care*. 2nd ed. Chichester, West Sussex: John Wiley & Sons, Inc.; 2013.

2 Van der Wees PJ, Moore AP, Powers CM, Stewart A, Nijhuis-van der Sanden MWG, de Bie RA. Development of clinical guidelines in physical therapy: perspective for international collaboration. *Phys Ther*. 2011;91(10):1551-1563.

3 Bekkering GE, Tulder MW Van, Hendriks EJ, *et al*. Implementation of clinical guidelines on physical therapy for patients with low back pain : randomized trial comparing patient outcomes after a standard and active implementation strategy. *Phys Ther.* 2005;85(6):544-555.

4 Bekkering GE, Hendriks EJ, van Tulder MW, *et al*. Effect on the process of care of an active strategy to implement clinical guidelines on physiotherapy for low back pain: a cluster randomised controlled trial. *Qual Saf Health Care.* 2005;14(2):107-112.

5 Van der Wees PJ, Jamtvedt G, Rebbeck T, de Bie RA, Dekker J, Hendriks H. Multifaceted strategies may increase implementation of physiotherapy clinical guidelines: a systematic review. *Aust J Physiother.* 2008;54(4):233-241.

6 Harting J, Rutten GM, Rutten ST, P KS. A qualitative application of the diffusion of innovations theory to examine determinants of guideline adherence among physical therapists. *Phys Ther.* 2009;89(3):221-232.

7 van Bodegom-Vos L, Verhoef J, Dickmann M, *et al*. A qualitative study of barriers to the implementation of a rheumatoid arthritis guideline among generalist and specialist physical therapists. *Phys Ther.* 2012;92(10):1292-1305.

8 Dannapfel P, Peolsson A, Nilsen P. What supports physiotherapists' use of research in clinical practice? A qualitative study in Sweden. *Implement Sci.* 2013;8:31.

9 Rutten GM, Kremers S, Rutten ST, Harting J. A theory-based cross-sectional survey demonstrated the important role of awareness in guideline implementation. *J Clin Epidemiol.* 2009;62(2):167-176. http://www.sciencedirect.com/science/article/pii/S0895435608001546. Accessed January 12, 2014.

10 Epstein RM. Self Monitoring in Clinical Practice. *J Contin Educ Health Prof.* 2008.

11 Davis DA, Mazmanian PE, Fordis M, Harrison R Van, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence. A systematic review. *JAMA*. 2006;296(9):1094-1102.

12 Eva KW, Regehr G. "I' ll mever play professional football" and other fallacies of self-assessment. *J Contin Educ Health Prof.* 2008;28(1):14-19.

13 Sargeant J, Eva KW, Armson H, *et al*. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ.* 2011;45(6):636-647.

14 Eva KW, Armson H, Holmboe E, *et al*. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ Theory Pract.* 2012;17:15-26.

15 Mann K, van der Vleuten CP, Eva KW, *et al*. Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. *Acad Med*. 2011;86(9):1120-1127.

16 Baker R, Camosso-Stefinovic J, Gillies C, *et al*. Tailored interventions to overcome identified barriers to change: effects on professional practice and health care outcomes. *Cochrane Database Syst Rev.* 2014;(3).

17 Maas MJM, Sluijsmans DM, van der Wees PJ, Heerkens YF, Nijhuis-van der Sanden MWG, van der Vleuten CPM. Why peer assessment helps to improve clinical performance in undergraduate physical therapy education: a mixed methods design. *BMC Med Educ.* 2014;14(1):117.

18 van Dulmen SA, Maas MJ, Staal B, *et al*. Effectiveness of peer-assessment for implementing a Dutch physical therapy low back pain guideline: a cluster randomized controlled trial. *Phys Ther.* 2014;94(10):1396-1409.

19 Mann K, Gordon J, MacLeod A. Reflection and reflective practice in health professions education: a systematic review. Adv Health Sci Educ Theory Pract. 2009;14(4):595-621.
20 Epstein RM, Siegel DJ, Silberman J. Self-monitoring in clinical practice: a challenge for medical educators. *J Contin Educ Health Prof.* 2008;28(1):5-13.
21 Pronovost PJ, Hudson DW. Improving healthcare quality through organisational peer-to-peer assessment: lessons from the nuclear power industry. *BMJ Qual Saf.* 2012;21(10):872-875.
22 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA.* 1993;269(13):1655-1660. http://www.ncbi.nlm.nih.gov/pubmed/8240483.
23 Lin SSJ, Liu EZF, Yuan SM. Web-based peer assessment : feedback for students with various thinking-styles. *J Comput Assist Learn.* 2001;17:420-432.
24 Topping KJ. Methodological quandaries in studying process and outcomes in peer assessment. *Learn Instr.* 2010;20(4):339-343.
25 Brehaut JC, Eva KW. Building theories of knowledge translation interventions: use the entire menu of constructs. *Implement Sci.* 2012;7:114.
26 Colquhoun HL, Brehaut JC, Sales A, *et al*. A systematic review of the use of theory in randomized controlled trials of audit and feedback. *Implement Sci.* 2013;8(1):66.
27 Ivers NM, Sales A, Colquhoun H, *et al*. No more "business as usual" with audit and feedback interventions: towards an agenda for a reinvigorated intervention. *Implement Sci.* 2014;9:14.
28 Staal BJ, Hendriks EJ, Heijmans M, *et al*. KNGF Richtlijn Lage-Rugpijn voor fysiotherapie en manuele therapie [Guideline low back pain for physical therapy and manual therapy]. Royal Dutch Society for Physical Therapy. http://www.fysionet-evidencebased.nl/index.php/component/kngf/richtlijnen. Published 2010.
29 Janssen-Noordman AMB, Merriënboer JJG, van der Vleuten CPM, Scherpbier AJJA. Design of integrated practice for learning professional competences. *Med Teach.* 2006;28(5):447-452.
30 King N, Cassel CM, Symon G. Using templates in the thematic analysis of texts. In: Cassell C, Symon G, eds. *Essential Guide to Qualitative Methods in Organizational Research*. 1st ed. London: Sage Publications; 2004:256-270.
31 Huberman AM, Miles MB, Denzin NK, Lincoln YS. Data management and analysis methods. In: *Handbook of Qualitative Research*. Sage Publications; 1994:428-444.
32 Qualitative research review guidelines – RATS. BioMed Central. http://www.biomedcentral.com/authors/rats.
33 Bandura A, Locke E a. Negative self-efficacy and goal effects revisited. *J Appl Psychol.* 2003;88(1):87-99.
34 Ajzen I. Nature and operation of attitudes. *Annu Rev Psychol.* 2001;52:27-58.
35 Usher EL, Pajares F. Sources of Self-Efficacy in School: Critical Review of the Literature and Future Directions. *Rev Educ Res.* 2008;78(4):751-796.
36 Teigen K. Yerkes-Dodson: a law for all seasons. *Theory Psychol.* 1994;4(4):525-547.
37 McConnell MM, Eva KW. The role of emotion in the learning and transfer of clinical skills and knowledge. *Acad Med.* 2012;87(10):1316-1322.
38 Higgins TE. Beyond pleasure and pain. *Am Psychol.* 1997;52:1280-1300.
39 Iacoboni M. Mirroring People: *The New Science of How We Connect with Others.* 2nd ed. (Farrar S and G, ed.). New York: Macmillan; 2009.
40 Lave J, Wenger E. *Communities of practice*. Cambridge: Cambridge university press; 1999.
41 Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Evolution of Wenger's concept of community of practice. *Implement Sci.* 2009;4(1):11. doi:10.1186/1748-5908-4-11.
42 McCaghie WC, Issenberg BS, Petrusa ER, Scalese RJ. A critical review of simulation-based medical education research: 2003-2009. *Med Educ.* 2010;44:50-63.
43 Munshi F, Lababidi H, Alyousef S. Low- versus high-fidelity simulations in teaching and assessing clinical skills. *J Taibah Univ Med Sci.* 2015;10(1):12-15.

44 Bosse HM, Nickel M, Huwendiek S, Jünger J, Schultz JH, Nikendei C. Peer role-play and standardised patients in communication training: a comparative study on the student perspective on acceptability, realism, and perceived effect. *BMC Med Educ.* 2010;10:27.

45 Mounsey AL, Bovbjerg V, White L, Gazewoord J. Do students develop better motivational interviewing skills through role-play with standardised patients or with student colleagues? *Med Educ.* 2006;40:775-780.

46 Farmer EA, Page G. A practical guide to assessing clinical decision-making skills using the key features approach. *Med Educ.* 2005;39(12):1188-1194.

47 Overheem K, Faber MJ, Onyebuchi AA, *et al*. Doctor performance assessment development in daily practise: does it help doctors or not? A systematic review. *Med Educ.* 2007;41(11):1039-1049.

48 Norman G, Bordage G, Page G, Keane D. How specific is case specificity*? Med Educ.* 2006;40(7):618-623.

49 Grant AM, Franklin J, Langford P. The self-reflection and insight scale: a new measure of private self-consciousness. *Soc Behav Pers.* 2002;30(8):821-836.

50 Schön D. The Reflective Practitioner: *How Professionals Think in Action.* San Francisco: Jossey-Bass Inc; 1983.

51 Greene J, Azevedo R. A theoretical review of Winne and Hadwin's model of self-regulated learning: new perspectives and directions. *Rev Educ Res.* 2007;77(3):334-372.

52 Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Use of communities of practice in business and health care sectors: a systematic review. *Implement Sci.* 2009;4:27.

53 Prochaska JO, Redding CA, Evers KE. Health behavior and health education. In: Glanz K, Rimer B k, Viswanath K, eds. *Health Behavior and Health Education: Theory, Research, and Practice.* 4th ed. Wiley & Sons; 2008:97-121.

Appendix

Online questionnaire

The PA program consisted of several parts. The overview below shows the distinct learning tasks and subtasks. Please rank the eleven subtasks as presented from high to low learning value (1 = most learning value, 11 = least learning value).

Overview of tasks and subtasks

Tasks		Subtasks		Rank
Prepare task	Study manual	1	Study PA procedure and guidlines	
Perform task	Perform in PT role	2	Perform clinical task individually	
		3	Receive peer feedback	
		4	Receive external coach feedback	
		5	Receive simulated patient feedback	
		6	Receive written feedback and scores	
	Perform in assessor role	7	Observe peer performance	
		8	Provide oral feedback	
		9	Provide written feedback and scores	
	Perform in patient role	10	Simulate patient problem	
Evaluate task		11	Design and discuss change plan	

Please motivate your choice for the three most instructive learning tasks

Rank	Comment*
1	
2	
3	

*Characters unlimited

Interview guide

1 What did you expect of the Peer Assessment (PA) program?

Did you have personal learning goals? If so, can you describe them?

To what extent this program has met your expectations? Please explain.

The PA program aimed to enhance clinical performance of physical therapists in primary care. What are the strengths and weaknesses of PA, and why?

2 How did you perceive the PA program, and how did it affect your daily practice?

How did you perceive the PA sessions?

Can you remember a particular event that impressed you? If so, please describe.

When you look back on the PA process, did it affect your professional practice? If so, can you explain what has changed?

Do you think the PA process affected the professional practice of your colleagues? If so, can you explain what has changed?

3 Which elements of the PA program have the strongest learning value in your opinion?

The PA program consisted of several parts. In the questionnaire you were asked to rank eleven subtasks as presented in the overview, from high to low learning value. You indicated that you perceived task X to have the strongest learning value. Can you explain why? Can you proceed to do the same for task Y and Z?

Chapter 5
An innovative peer assessment approach to enhance guideline adherence in physical therapy: a single-masked cluster-randomized controlled trial

Marjo Maas
Philip van der Wees
Carla Braam
Jan Koetsenruijter
Yvonne Heerkens
Cees van der Vleuten
Ria Nijhuis-van der Sanden

Physical Therapy Journal, 2015;95(4):600-612

Abstract

Background

Clinical practice guidelines are not readily implemented in clinical practice. One of the impeding factors is that physical therapists do not hold realistic perceptions of their adherence to clinical practice guidelines. Peer Assessment (PA) is a implementation strategy that aims at improving guideline adherence by enhancing reflective practice, awareness of professional performance, and attainment of personal goals.

Objective

To compare the effectiveness of PA with the usual Case Discussion (CD) strategy on adherence to clinical practice guidelines for physical therapy management in patients with upper extremity complaints.

Design

Single-masked cluster-randomized controlled trial with pre-post-test design.

Intervention

Twenty communities of practice (n=149 physical therapists) were randomly assigned to the PA or CD program, both consisting of four sessions over six months. PA and CD groups worked on identical clinical cases relevant to the guidelines. PA focused on individual performance observed and evaluated by peers; CD focused on discussion.

Outcomes

Guideline adherence was measured with clinical vignettes, reflective practice was measured with the Self-Reflection and Insight Scale (SRIS), awareness of performance was measured via the correlation between perceived and assessed improvement, and attainment of personal goals was measured with written commitments to change.

Results

The PA groups improved more on guideline adherence compared with the CD groups (effect: 22.52, 95% confidence interval [CI]:2.38 – 42.66, P=.03). SRIS scores did not differ between PA and CD groups. Awareness of performance was greater for PA groups (r=0.36) than for CD groups (r=0.08), (effect: 14.73, 95%CI:2.78-26.68, P=.01). PA was more effective in attaining personal goals (effect: 0.50, 95% CI: 0.04 – 0.96, P=.03).

Limitations
Limited validity of clinical vignettes as a proxy measure of clinical practice.

Conclusions
PA was more effective than CD in improving adherence to clinical practice guidelines. Personal feedback may have contributed to its effectiveness. Future research should address the role of the group coach.

Background

Clinical practice guidelines are designed to facilitate evidence-based practice and to improve the quality of health care.[1] The purpose of guidelines is to enhance transparency of care, to reduce unwarranted variability in practice, and to increase accountability to external stakeholders.[2] Despite a multitude of implementation strategies, research has demonstrated unambiguously that clinical practice guidelines are not readily implemented in everyday clinical practice.[3,4] The main bottlenecks for practitioners are attributable to knowledge, attitudes, and factors concerning social, organizational, and societal support.[5] Because education is assumed to be the first step to behavioral change in clinical practice, a variety of educational interventions have been designed to address knowledge, skills, and attitudes.[6] Systematic reviews studying the effectiveness of educational strategies, however, have shown little to moderate effects in improving evidence-based practice.[7] Rutten *et al*.[8] assessed the effectiveness of a quality improvement program aimed at professional and organizational behavioral change in physical therapist (PT) practice. Guideline adherence was assessed by clinical vignettes in a one-group pre- and post-test design. They found 3.1% increase in adherence. Wensing *et al*.[6] reported a mean effect of 5% on different aspects of clinical practice, irrespective of the type of educational intervention. Research showed that the effectiveness of educational strategies might improve when the intervention addresses small groups and allows for active participation and social interaction.[9] In addition, change may be more likely if strategies are specifically chosen to address identified barriers to change.[10] Bekkering *et al*.[11] showed moderate improvement of adherence to clinical practice guidelines by PTs in the Netherlands through active educational strategies (discussion, role playing) compared to standard passive methods of guideline dissemination in physical therapy.

Guideline adherence of PTs depends on levels of awareness of guideline consistent behavior. Rutten[12] used clinical vignettes to compare self-reported with externally assessed adherence. Realistic perceptions of adherence to clinical practice guidelines were found in 38.5%, while 25.2% overestimated and 36.4% underestimated their adherence. These differences in levels of awareness interfered with other determinants of guideline adherence, such as motivation to change. Research showed that health care professionals have a limited ability to accurately assess their own level of competence,[13,14] which they systematically over- or underestimate.[16] The development of adequate self-perception requires both internal and external information about one's professional performance as well as knowledge of appropriate performance standards.[17] This finding is supported by studies showing that the effect of educational strategies on evidence-based practice increases when they are combined with other strategies, such as audit and feedback.[3,18]
Yet audit and feedback have not consistently been found effective to change practice. A systematic review of Ivers *et al*.[19] showed a mean improvement of compliance with desired practice of 4.3% (dichotomous outcomes) and 1.3% (continuous outcomes). Whether feedback is accepted and used to change professional practice depends on a multitude of variables.[20,21] Clinicians struggle with accepting feedback when it is incongruent with their self-assessment or threatens their self-confidence.[17,22] Feedback appears to be more acceptable20 when it is provided in an environment of trust and mutual respect, and it is likely to be rejected when the provider is not perceived to be a credible and trustworthy source of information[17,21] or when it conflicts with personal or group norms and values.[23] Acceptance may be enhanced when feedback is tailored to the stages of change as described by Prochaska,[25] and when it closely connects to the context of daily practice.[5,25]
Situated learning theory, based on studies by Lave and Wenger[26] and Li *et al*.[27], shows that professional knowledge acquired in a certain 'situation', transfers only to similar situations.[26,27] Their studies support the assumption that feedback provided within communities of practice (CoPs) has greater impact on the improvement of clinical practice than feedback provided by 'outsiders'. Moreover, the involvement of CoP participants in each other's professional development process may facilitate acceptance of feedback and alignment with personal learning needs and goals.[28-30]
Drawing on these considerations, we introduced peer assessment (PA) as a new implementation strategy for clinical guidelines within existing CoPs. PA is the process whereby professionals evaluate

or are being evaluated by their peers and provide each other with performance feedback. The positive impact of PA on learning and change has been well researched in higher education[31–33] and health care professional education.[34–37] However, Topping[38] argues that generalizations to professional practice should be made with caution, because successful PA implementation depends on variables such as the context of peers, the nature of the PA intervention, and the outcomes assessed. Lack of specified knowledge about the PA practices, impedes the transfer of results.[38]

During the implementation of the Dutch guideline for PT management in patients with nonspecific low back pain,[39] PA showed promising results. In a randomized controlled trial conducted by van Dulmen and collegues,[40] PA was significantly more effective in improving guideline adherence (measured by clinical vignettes) than the usual implementation strategy Case Discussion (CD). We redesigned this PA program for the implementation of a newly developed guideline for CANS (complaints of arm, neck and shoulder)[41] and a new evidence statement for Subacromial Complaints.[42] We also included the appraisal of patient records as a new element. Record keeping is an important quality indicator for physical therapy care and patient records offer authentic assessment material that reflects clinical practice.[43,44]

PA and CD are implementation strategies informed by several, sometimes overlapping theoretical constructs concerning learning and behavior change: principles of social constructivist learning theory[45], such as contextual learning, collaborative learning, and active knowledge construction, and principles of self-regulated learning theory, such as conscious goal setting and reflection.[29,46] In addition, the PA-approach builds on principles of social-cognitive learning theory (concrete experience with – and performance of desired behavior)[47], and stages of change theory (tailored feedback).[29,30] Moreover PA targets the development of a mutual accepted quality standard of performance by introducing peers with an 'assessor' perspective.[48,49]

The objective of this study was to compare the effectiveness of PA with the casual CD strategy on adherence to clinical practice guidelines for physical therapy management in patients with upper extremity conditions.

Following social cognitive theory, our hypothesis was that the performance based approach of PA, combined with giving and receiving personal performance feedback, would be a more powerful tool than the CD approach for uncovering areas in personal clinical practice that need improvement. Based on self-directed learning theory and stages of change theory, we also posited that PA would

provide a stronger trigger for reflective practice, would develop greater awareness of guideline-consistent behavior in daily practice, and would be more effective in guiding self-directed change toward personal learning goals than CD.
The effectiveness of PA and CD was tested on four outcome measures: 1) guideline adherence, 2) reflective practice, 3) awareness of performance, and 4) attainment of personal goals.

Method

Design

This study was a single-masked cluster-randomized controlled trial with a pre-and post-test design comparing the effectiveness of two implementation strategies.

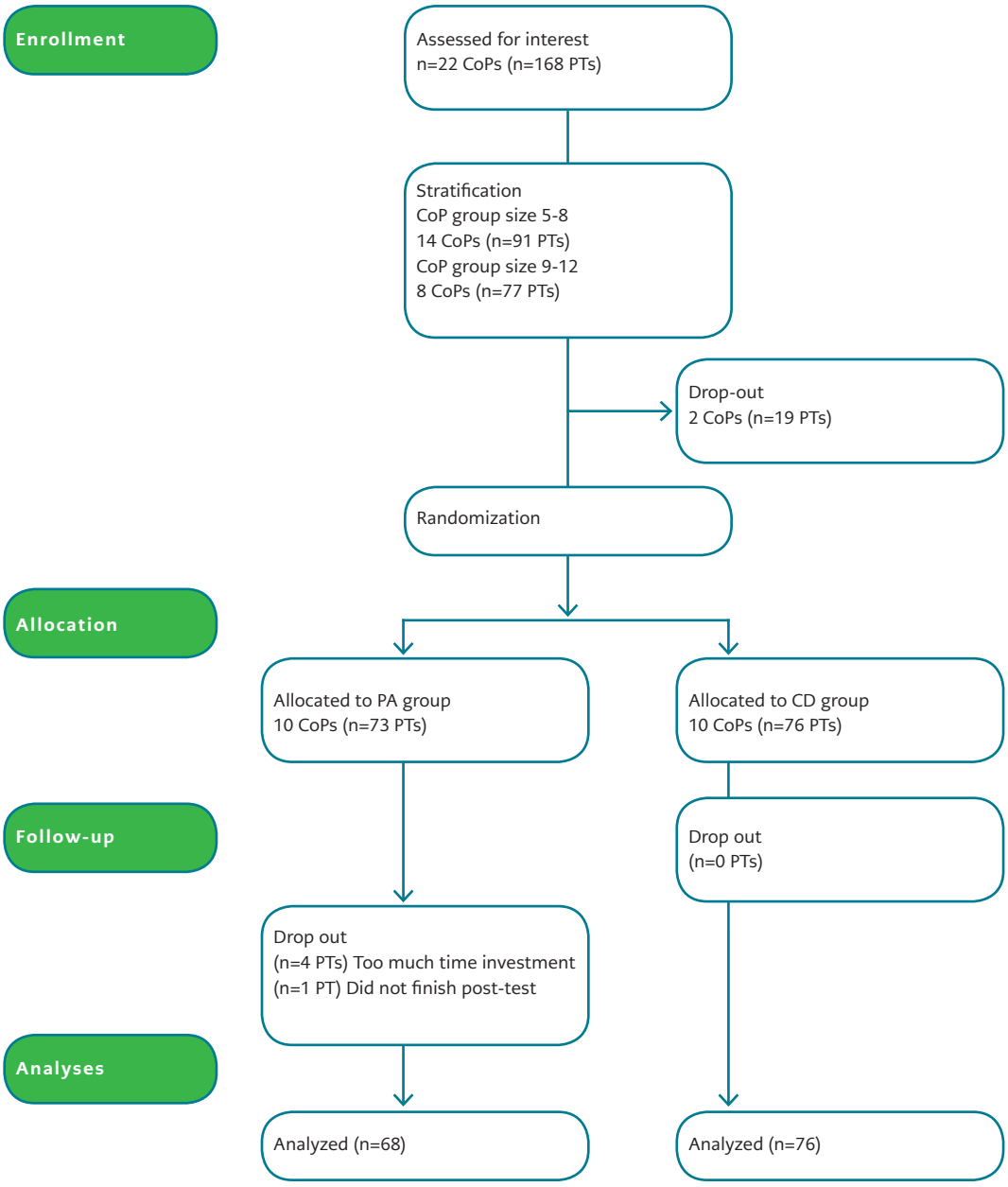
Setting and participants

Participants were physical therapists (PTs), organized in Communities of Practice (CoPs), which are small groups of 5–15 professionals who share the same setting or the same interests and who work together on the improvement of the quality of care in yearly provided post-graduate training programs provided by the Royal Dutch Society for Physical Therapy (KNGF). CoPs can register with the KNGF to participate in such a programs. The aim of the program under study was to implement two newly developed guidelines for physical therapy management in patients with upper extremity complaints. In November 2011 formal contact persons of CoPs were invited by an electronic newsletter to a joint introduction meeting on the training program. CoPs that showed interest in participating received an information letter containing details of the training program, randomization procedure, time investments, risks and advantages. Participation was awarded with continuing education credits for the Dutch quality register. All CoPs that showed interest were eligible for inclusion. We conducted a sample-size calculation based on an estimated difference between the two interventions of 5% (power: 80%, p=.05), with an anticipated intra-class correlation (ICC) of 0.10, and 10% loss to follow-up. This resulted in the required inclusion of n=110 PTs in 22 clusters with at least 5 PTs per cluster.[50]

Randomization

In December 2011, 22 CoPs showed interest in our study. Before randomization (January 2012), 2 CoPs withdrew because they felt the program would take too much time. A flowchart of the study sample is presented in the figure. Because we expected that the

Figure 1


Enrollment
Assessed for interest
n=22 CoPs (n=168 PTs)
Stratification
CoP group size 5-8
14 CoPs (n=91 PTs)
CoP group size 9-12
8 CoPs (n=77 PTs)
Drop-out
2 CoPs (n=19 PTs)
Randomization
Allocation
Allocated to PA group
10 CoPs (n=73 PTs)
Allocated to CD group
10 CoPs (n=76 PTs)
Follow-up
Drop out
(n=4 PTs) Too much time investment
(n=1 PT) Did not finish post-test
Drop out
(n=0 PTs)
Analyses
Analyzed (n=68)
Analyzed (n=76)

size of the group would affect its learning,[31,35,38] we aimed at a balanced distribution of large and small CoPs between CD and PA groups. The 20 CoPs were stratified by the number of participants into two blocks of groups with 5-10 and 11-15 participants and were randomly assigned to the intervention or the control group by using randomization software.[51] This procedure resulted in 10 PA groups (n=73 PTs) and 10 CD groups (n=76 PTs). The CoPs were masked for the intervention because PA and CD were presented as alternative interventions. The primary researcher (MM) was not masked for the allocation of CoPs, because she participated in conducting the intervention program. To reduce the risk of bias, she was masked for the outcomes until the data sampling was completed and the pre-post-test differences were calculated.

Interventions

Before the start of the program, both PA and CD groups received a link to the KNGF guidelines and a link to the pretest questionnaires. All participants received by email a program guide tailored to the intervention providing detailed information about learning objectives, learning content, training schedule, didactic format and procedure. The program for both groups consisted of four 3-hour sessions and was launched in February 2012. Table 1 shows a detailed program overview and time schedule. In sessions 1, 2 and 4, the participants worked on written cases that fully covered the patient profiles described in the guidelines. Session 3 consisted of a review of patient records using a set of quality indicators derived from the KNGF guidelines on record keeping.[52]

The main difference between the two interventions is that in the PA approach, the tasks are structured, with a focus on performance rather than discussion, and roles are pre-defined. Each participant performed 3 roles: *physical therapist, assessor*, and *simulated patient*. Because PTs were complete novices in the PA method, the process was supervised by a group coach. In the CD approach tasks are less structured with ample opportunity for in-depth elaboration and discussion, and participant roles are not defined. In PA and CD, participants worked on identical cases concerning problem content, but for PA these cases were adjusted to allow for performance of participants in different roles. In PA, written cases were not known in advance but were presented by a coach on the spot, simulating daily practice. Participants were provided with ground rules for providing and receiving constructive feedback and for creating a safe learning environment. In the role of PT, they analyzed the case by reasoning aloud and demonstrated (hands-on) diagnostic

Table 1 — Intervention Programs in both groups

	Period	Peer Assessment (PA)	Case Discussion (CD)
Pretest	Feb-2012	Online Test based on four clinical vignettes (TCV)	
		Online questionnaire Commitment to Change Statements (CTCS)	
		Online questionnaire Self Reflection and Insight Scale (SRIS)	
	Feb-2012	Program manual by email	
Session 1	Feb-2012	PA of individual performance	Case-based group discussion
Session 2	March-2012	PA of individual performance	Case-based group discussion
Session 3	April-2012	Review of personal patient records	Review of personal patient records
Session 4	June-2012	PA of individual performance	Case-based group discussion
	July-2012	Answering key to clinical cases	
Posttest	July-2012	TCV	
		CTCS	
		SRIS	
	Sept-2012	Personal knowledge of results by email	

and treatment skills. Peer performance was assessed by using a global scoring sheet designed to support peer assessors in giving constructive feedback. It contained three performance categories (planning, performance and evaluation) that were scored on a 5-point Likert scale (1 = much improvement needed to 5 = no improvement needed).

Accordingly qualitative oral improvement feedback was given. The complete PA program guide, including assessment criteria, is accessible online.[53] Three group coaches (HE, HN and VV) were trained by MM in the PA procedure, supported by a coaching manual. They were experienced tutors in problem-based learning, and they were instructed to encourage the group in providing tailored performance feedback, and not to serve as an information source themselves. To reduce the risk of bias, the group coaches were not involved in the development of clinical vignettes. For CD groups, written cases were included in the program guide to allow for proper preparation, along with instructions and written questions to guide the discussion process. After completion of the program in July 2012, and before the posttest, all participants received an email with model answers to all the cases, that were discussed during the program to control for unintended differences in knowledge development between and within groups, due to the influence of the group coach.

Outcome measures

1 Guideline adherence

Participants completed an online test based on four clinical vignettes one week before the start of the program and within two weeks after completion of the program. A previous study of Rutten[54] showed that vignettes have acceptable validity to measure PTs' adherence to clinical practice guidelines and these results were consistent with studies by Peabody and colleagues.[55-57] Clinical vignettes require factual knowledge of clinical practice guidelines as well as clinical reasoning consistent with clinical practice guidelines in the context of a clinical problem. Four clinical vignettes were based on upper-extremity disorders in the context of direct physical therapy access.[58] Three vignettes adequately covered the patient profiles described in the guidelines, and the fourth vignette did not because of 'red flags'. The vignettes and test items were constructed by a team containing 2 PT scientists involved with guideline development, 5 PT practitioners specializing in upper extremity conditions and 1 PT educational scientist specializing in assessment development. Each vignette was accompanied by 11 response categories derived from the guidelines: 1) clinical pattern, 2) impairments and disabilities, 3) onset risk factors, 4) impeding recovery factors, 5) patient profile according to guidelines, 6) measurement instruments, 7) diagnostic clinical tests, 8) main treatment goals, 9) treatment approach, 10) information and advice, and 11) expected recovery time. Each response category contained a set of test-items in the form of statements. Vignette 1, 2 and 3 each contained 119 items; vignette 4 consisted of fewer items (n=31) because quality indicators addressing additional diagnosis and treatment were not applicable. The statements could be scored on a 3-point scale: D = disagree, D/A = neither disagree nor agree, A = agree. Because clinical evidence is limited and guidelines cannot inform all clinical decisions, the option D/A was offered to reflect the way information is processed in the context of uncertainty.[59] The group of eight experts evaluated and adjusted the vignettes and test items. All experts completed the final test informed by the guidelines. The scoring method took variability of reasoning among experts into account as long as differences were limited to two alternatives (D and D/A; or D/A and A). Items with contradictory answers (D and A) were reviewed. The alternative that was chosen by the majority (>4) was assigned two points, and equal distribution was assigned one point for each alternative. A majority opting for alternative D/A did not occur. The final scoring key was discussed among 4 experts until consensus was reached. The maximum score was 737 points

(some answers received 1 point). The Appendix shows an example of a test item and its scoring key. The scores for each vignette were added up, and mean total scores on the 4 clinical vignettes were perceived of as a measure 'guideline adherence'.

2 **Reflective practice**
Participants completed the validated questionnaire 'Self-Reflection and Insight Scale' (SRIS), developed by Grant.[60] It aims to measure the readiness for purposeful behavior change and has been shown responsive to change in the context of continuing professional education.[61] The SRIS has been validated by Roberts & Stark[62] and modified for the medical education context. It contains three subscales: the *engagement* with reflection, the *need* for reflection, and the insights obtained by reflection. Engagement and need refer to the practice of inspecting and evaluating one's own thoughts, feelings and behavior; insight refers to understanding them. Sum scores for each subscale were computed and mean total scores were conceived of as a measure of 'reflective practice'.

3 **Awareness of performance**
Awareness was conceived of as the association between perceived improvement and assessed improvement. At posttest, participants were asked to indicate how much guideline knowledge they had at pretest and how much at posttest on a 5-point scale from 1 = no knowledge to 5 = much knowledge. The pretest-posttest difference was conceived of as a measure of perceived improvement. Assessed improvement was the difference between pretest and posttest scores on clinical vignettes.

4 **Attainment of personal goals**
At pretest all participants were asked to formulate 3 learning goals, ordered on personal importance conform the concept of Commitments to Change Statements.[29,63] Conscious goal setting belonged to the intervention strategy to enhance self-directed learning, and progression through the stages of change.[30] They also served as an outcome measure.[63] Before the posttest all participants were emailed a reminder of their personal goals at pretest. At posttest they were asked to indicate the extent to which their goals were achieved on a 3-point scale from 1 = not achieved, 2 = partly achieved, to 3 = achieved. Achievement scores for each personal goal were added, and mean total scores were conceived of as a measure of goal attainment.

Statistical Analysis

IBM SPSS, version 20 was used for statistical analysis. For baseline characteristics (age, gender, clinical setting, specialization), pretest scores on clinical vignettes and SRIS of PTs were described and tested for differences between the PA and CD groups using chi-square tests and unpaired t-tests. Internal consistency of the clinical vignettes was tested by Cronbach alpha. Outcome differences between PA and CD groups were described and tested by multilevel linear regression to adjust for clustering within CoPs. For each outcome measure the intraclass correlation coefficient (ICC) was calculated to test clustering at the CoP level. Baseline characteristics were included as covariate when differences between groups were statistically significant.

Pretest and posttest sum scores and mean total scores were calculated for each vignette. The intervention effect for guideline adherence was estimated with *posttest scores* on vignettes as dependent variable and *intervention* and *pretest scores* as covariates. In the same way mean pretest and posttest SRIS scores were calculated. The intervention effect for reflective practice was tested with *posttest scores* as dependent variable and *intervention* and *pretest scores* as covariates. Mean posttest sum scores were calculated for each personal objective and total scores. Differences in attainment of personal goals were tested with *personal goals* as dependent variable and *intervention* as covariate. Mean assessed improvement scores and mean perceived improvement scores on clinical vignettes were calculated and correlations were computed with *assessed improvement* as dependent variable and *perceived improvement* as independent variable. Differences in awareness were estimated with *assessed improvement* as dependent variable and the interaction between the variables *intervention* and *perceived improvement* as covariate.

Role of funding source

This was a study initiated by researchers and funded by the Royal Dutch Society for Physical Therapy (KNGF). The KNGF had no role in the conduct of this study, analysis or interpretation of data, or preparation of the manuscript.

Results

The pretest response was 100%. The post-test response of PA was 93,2% (n=68) and of CD was 100% (n=76). A flow chart is presented in figure 1. Baseline characteristics of the participating PTs are

Table 2 — Physical Therapists characteristics

Intervention (n)		PA[a] (n=73)	CD[b] (n=76)
Mean age (Y) (SD)	45,15 (11,03)	44,76 (9.74)	
Mean working experience (Y) (SD)	20,42 (11,37)	20,86 (9.71)	
Gender male / female*	39 / 34	27 / 49	
Specialization	No specialization / %	47 / 64.4%	46 / 61.3%
	Manual Therapist / %	7 / 9.6%	8 / 10.7%
	Other Specialization / %	19 / 26.0%	21 / 28.0%
Clinical Setting	Primary Care / %	56 / 76.7%	58 / 76.3%
	Hospital or nursing home / %	17 / 23.3%	18 / 23.7%

[a] Peer Assessment; [b] Case-based Discussion, * Significant difference: P<.05

Table 3 — Multi-level analyses for Guideline adherence, Reflective practice, and Attaining personal goals

	PA[a] group		CD[b] group					
Measure	**Pretest N=73**	**Posttest N=68**	**Pretest N=76**	**Posttest N=76**	**Estimated Difference**	**ICC[c]**	**95% CI[d]**	**P**
Vignettes					22.52	0.079	2.38 – 42.66	.03*
Means	474.26	501.99	472.54	482.03				
SD	65.85	65.07	56.09	62.19				
SRIS[e]					0.06	0.048	-2.79 – 2.65	.96
Means	82.71	85.10	83,47	85.33				
SD	9,80	9.32	8.08	8.90				
CTCS[f]					0.50	0.002	0.04 – 0.96	.03*
Means		7.44		6.90				
SD		1.44		1.29				

[a] Peer Assessment; [b] Case-based Discussion, [c] Intraclass correlation coefficient, [d] 95% Confidence Interval, [e] Self Reflection and Insight Scale (reflective practice), [f] Commitment to Change Statements (attainment of personal goals), * Significant difference P<.05

presented in table 2. We found differences between PA and CD for gender (P=.028), so we controlled for this confounder in multilevel linear regression. Internal consistency between scores across clinical vignettes (n=4) was good (pretest α=.82; posttest α=.86).
Table 3 presents the results of the outcome measures guideline adherence, reflective practice, and attainment of personal goals. Results of awareness of performance are presented separately.

Concerning guideline adherence, table 3 shows that mean pretest scores on vignettes were comparable between PA and CD groups. At posttest the PA and CD groups showed significant improvement: PA groups=29.82 (SD=63.97), *P*<.001 and CD groups=9.49 (SD=40.52), *P*<.001. Percent improvement was 5.8% and the PA groups and 2.0% for the CD groups. Multilevel linear regression analysis, controlling for sex, showed that the difference between PA and CD groups was statistically significant in favor of the PA groups (estimated effect= 22.52 points; 95% CI=2.38–42.66; *P*=.031).
Mean pretest scores on the SRIS showed no difference between PA- and CD groups. At posttest, scores were significantly improved in both PA and CD groups: PA=2.34 (SD=8.69), *P*<.001 and CD=1.85 (SD=7.05), *P*<.001. Percent improvement was for PA=2.8% and CD=2.2%. The difference between groups was not statistically significant (estimated effect= -0.06 points, 95% CI: -2.79-2.65, *P*=.96).
The results related to attainment of personal goals, showed that scores were significantly higher for PA groups than CD groups (estimated effect= 0.50; 95% CI: 0.04–0.96; *P*=.03).
At posttest participants in the PA group showed greater awareness of their professional performance. The correlation between 'perceived improvement' and 'assessed improvement' was r= 0.36 for PA groups, *P*=.002, and r=0.08, *P*=.50 for CD groups. The difference was statistically significant (estimated effect=14.73; 95% CI=2.78–26.68, *P*=.01).

Discussion

This study evaluated the effect of 2 implementation strategies for the implementation of Dutch clinical practice guidelines. It showed that PA was more effective in improving 'guideline adherence' measured by clinical vignettes, than CD. Moreover PA groups were more effective in 'attaining personal goals' and showed higher levels of 'awareness of performance'. The strength of this study is that we offered PA and CD groups high quality programs. Program evaluation showed that the perceived instructional value of PA and CD was comparable between groups (results not presented). The outcome measures were equally facilitated by both interventions. First, PA- and CD groups had equal access to the guidelines, worked on solving identical clinical problems, and had equal access to the model answers of each problem. Second, neither of the two interventions included tasks, such as writing reflection reports and improvement plans that explicitly aimed to facilitate the outcomes

reflective practice, awareness of performance or *attainment of personal goals*. Any pre-test effect of the SRIS or the CTCS would have applied to both interventions.

We showed that a tailored, multifaceted intervention that addresses specific barriers to change,[10] such as 'awareness of performance' as identified by Rutten,[12] is effective and these findings are in line with the existing research evidence on implementation strategies.[4,10,19,64] We observed high baseline scores and moderate, but statistically significant, improvement scores for continuous outcomes of clinical vignettes (PA=5.8%; CD=2.0%). High baseline scores can be attributed to the fact that participants received the guidelines before the pretest and were allowed to study them beforehand. Several studies have shown that the intervention effect on desired practice increases when baseline performance is low.[19,65]

Rutten *et al*.[8] observed 3.1% guideline adherence increase for the low back pain guideline using clinical vignettes that assessed the effectiveness of their program. This program however, involved interventions addressing professional as well as organizational determinants of guideline adherence, so the results cannot be compared. We did not find studies that assessed comparable content and constructs concerning the improvement of the uptake of clinical practice guidelines except for the study of Van Dulmen *et al*.,[40] which showed that PA was more effective in the implementation of the low back pain guideline than CD and that is in line with our findings. Given the notion that intervention programs aimed at enhancing the transfer of research evidence into clinical practice are very heterogeneous and the generalizability of the effects is limited,[18,66] we explored the key-differences between PA and CD informed by theory, which may contribute to the generalizability of results. First, the PA-task is highly structured and necessitates strong involvement of each participant. Individual contributions in learning groups may vary widely when conditions such as shared responsibility, inter-dependency, mutual trust and psychological safety are not met.[32,67] Discussion tasks do aim at active participation, but the task structure does not control for individual contributions to group learning. Second, in contrast to CD, PA focuses on performance that can be observed and evaluated. PA group participants performed in pre-defined roles that forced the transfer of knowledge and skills in order to fulfill this role convincingly. In the role of PT, participants needed to make the transfer from 'implicit reasoning' to 'explicit reasoning' and from 'intentional behavior' to 'observable behavior'. The transferred knowledge and skills became transparent and this new information became accessible for group review.[68] The variety

of feedback that PA-participants obtained about their performance may have helped them to become aware of areas in professional practice that need improvement and may have supported them in attaining personal goals. In the assessor role, participants needed to make a transfer from 'implicit appraisal' to 'explicit appraisal'. Supported by predefined performance criteria, peer assessors revealed their personal norms about the quality of the observed behavior. Personal standards could be compared to group standards. Research has revealed that the availability of both internal as well as external data about one's performance is conditional on the development of correct self-perceptions (awareness)[49,50], which may explain why PA groups outperformed CD groups in this respect. A different perspective on why PA groups showed more improvement on guideline adherence, is the 'testing effect'. Recent insights in cognitive psychology show that tested information is better stored and retrieved from memory than information that is not.[70,71] Because PA is based on assessment (unlike CD), PA participants were repeatedly challenged to reproduce and apply newly acquired knowledge of clinical practice guidelines. That may have strengthened awareness of deficiencies, and facilitated retrieval of information from memory at posttest.

Although PA was more effective in 3 outcome measures, we could not explain these results by differences in 'reflective practice'. Both the PA and CD showed comparable improvement scores on the SRIS. These scores reflect perceptions of 'conscious' reflective practice[60,62] and conscious reflective practice was apparently enhanced by both interventions. Professional behavioral change however does not necessarily depend on conscious reflection but might also occur spontaneously through informal learning, such as concrete experience, role modeling,[72] and action observation.[73] PA involved concrete experience with guideline recommendations, including hands-on clinical skills. This might have prompted spontaneous (unintended) learning experiences more than the cognitive directed approach of CD. Studies by Bandura[74] show that experience is the strongest source of information for the development of self-efficacy beliefs and self-efficacy beliefs contribute significantly to motivation for behavioral change.

A third difference between PA and CD groups is the presence of the group coach. Peer groups contained experienced healthcare practitioners, but they were absolute novices in the PA method. Research has revealed that the acceptability of peer feedback highly depends on its perceived reliability[32,68] and that reliability and validity of peer feedback improves by training and experience.[31] It is possible that

peers have used the group coach as a tool to justify feedback because they did not fully rely on their peers' judgment. We assume that the effect of PA may increase when groups have more training in giving and receiving peer feedback and when standards for the quality of physical therapy care are internalized and mutually shared.[48,49] We also assume that successful PA practices depend on commitment of PTs to the PA procedure. The role of the group coach might be important in this respect.
On the other hand, it should also be noted that CD groups might have performed better when guided by a coach.
Finally, it should be noted that research has shown that improved guideline adherence is associated with improved process of care, but not always with improved patient outcomes.[5,11,75]

Limitations

First, clinical vignettes remain a proxy measure of clinical practice. Direct observation or audio or video recording might be measures that better reflect authentic practice, but a systematic review by Hrisos *et al*.[76] suggests that such measures may lack reliability and validity as well, because the behavior of interest cannot be standardized beforehand, and generalizations of the inferences are hard to make. Standardized (simulated) patients are generally considered to be an acceptable substitute, but these measures are costly and were not feasible given the sample size. Moreover standardized patients do not provide a sufficiently broad case-mix compared to clinical vignettes. Based on these considerations and the existing validity evidence,[55-57] we opted for clinical vignettes.
A second limitation is the involvement of the primary researcher MM in conducting the intervention program. To reduce risk of bias MM was masked for the outcomes until pre-post-test scores had been described and between- group differences were calculated. The primary researcher was involved in additional multilevel analyses supervised by JK who was masked for the intervention.
Third, the involvement of the group coaches should also be addressed. We controlled for differences in knowledge development between and within groups by emailing each participant before the post-test the model answers for all the clinical cases. Outcomes on all outcome measures did not show significant difference between group coaches MM, HE, HN or VM (results not presented). However, we could not control for implicit effects of the group coaches on motivation for change such as role modelling effects, increased self-efficacy beliefs, improved attitudes toward guidelines[24,32] and shared quality standards of performance.[49]

Fourth, the reliability of the test-scores should be considered. The test contained a considerable number of test-items (n=388). Although each participant fully completed the test within time-limits (2 hours) at pretest and posttest, cognitive overload caused by time-on-task may have biased test results. The effect however applied to both PA and CD groups, so it does not affect the validity of the inferences made about between-group differences.
Finally, we address the generalizability of our results. Studies report cultural differences in attitudes towards PA, for example reluctance of peers in giving face-to-face feedback.[28,32] External validity might be limited because the sample contained only Dutch PTs.

Conclusions

PA is more effective in guideline implementation than CD. PA participants showed higher improvement scores on clinical vignettes, showed more awareness of guideline consistent behavior and were more successful in attaining personal goals. The focus on individual performance, allowing for concrete experience with the guideline and obtaining personal performance feedback, probably contributed to its effectiveness. Moreover, performance in the assessor role necessitates critical appraisal of the observed behavior as well as critical self-appraisal.
We recommend PA for guideline implementation within CoPs. Further research should address the role of the group coach on the intervention effect and should explore the feasibility of replacing the group coaches by trained CoP members. They could play in important role in future bottom-up quality improvement initiatives addressing evidence-based practice and unwarranted variability in physical therapy care.

Abbreviations

HE = Henk van Enck
HN = Henk Nieuwenhuijzen
VV = Volcmar Visser
MM = Marjo Maas

References

1 Field MJ, Lohr KN (Eds). *Guidelines for Clinical Practice: from development to use.* Washington DC: National Academies Press; 1992.

2 Burgers J, Smolders M, Wollersheim H, Grol R. Richtlijnen als hulpmiddel bij de verbetering van de zorg [guidelines as a tool to improve patient care]. In: *Implementatie: Effectieve verbetering van de patiëntenzorg [Implementationa: Effective improvement of patient care].* 4th ed. Amsterdam: Reed Business; 2011:155–189.

3 van der Wees PJ, Jamtvedt G, Rebbeck T, de Bie RA, Dekker J, Hendriks H. Multifaceted strategies may increase implementation of physiotherapy clinical guidelines: a systematic review. *Aust J Physiother.* 2008;54(4):233–41.

4 Grimshaw J, McAuley L, Bero L, *et al*. Systematic reviews of the effectiveness of quality improvement strategies and programmes. *Qual Saf Health Care.* 2003;12(4):298–303.

5 Grol R, Wensing M, Bosch M, Hulscher M, Eccles M. *Theorieën over implementatie. In: Implementatie: effectieve verbetering van de patientenzorg.* 4th ed. Amsterdam: Reed Business; 2011:43–68.

6 Wensing M, Grol R, Fluit C. *Educatieve strategieën. In: Implementatie: effectieve verbetering van de patiëntenzorg.* 4th ed. Amsterdam: Reed Business; 2011:326–340.

7 Forsetlund L, Bjørndal A, Rashidian A, *et al*. Continuing education meetings and workshops : effects on professional practice and health care outcomes. *Cochrane Database Syst Rev.* 2009;2(3):CD003030.

8 Rutten GM, Harting J, Bartholomew LK, Schlief A, Oostendorp R a B, de Vries NK. Evaluation of the theory-based Quality Improvement in Physical Therapy (QUIP) programme: a one-group, pre-test post-test pilot study. *BMC Health Serv Res.* 2013;13(1):194.

9 Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Use of communities of practice in business and health care sectors: a systematic review. *Implement Sci.* 2009;4:27.

10 Cheater F, Baker R, Gillies C, *et al*. Tailored interventions to overcome identified barriers to change : effects on professional practice and health care outcomes (Review). *Cochrane Database Syst Rev.* 2009;(4):CD005470.

11 Bekkering GE, Hendriks EJ, van Tulder MW, *et al*. Effect on the process of care of an active strategy to implement clinical guidelines on physiotherapy for low back pain: a cluster randomised controlled trial. *Qual Saf Health Care.* 2005;14(2):107–112.

12 Rutten GM, Kremers S, Rutten ST, Harting J. A theory-based cross-sectional survey demonstrated the important role of awareness in guideline implementation. *J Clin Epidemiol.* 2009;62(2):167–176.

13 Adams AA, Soumerai S, Lomas J, Ross-Degnan D. Evidence of self-report bias in assessing adherence to guidelines. *Int J Qual Heal Care.* 1999;11(3):187–192.

14 Davis DA, Mazmanian PE, Fordis M, *et al.* Accuracy of physician self-assessment compared with observed measures of competence. A systematic review. *JAMA.* 2006;296(9):1094–1102.

15 Eva KW, Regehr G. " I' ll mever play professional football" and other fallacies of self-assessment. *J Contin Educ Health Prof.* 2008;28(1):14–19.

16 Regehr G, Eva KW. Self-assessment, self-direction, and the self-regulating professional. *Clin Orthop Relat Res.* 2006;449:34–38.

17 Sargeant J, Eva KW, Armson H, *et al*. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ.* 2011;45(6):636–647.

18 Brehaut JC, Eva KW. Building theories of knowledge translation interventions: use the entire menu of constructs. *Implement Sci.* 2012;7:114.

19 Ivers N, Jamtvedt G, Flottorp S, *et al*. Audit and feedback: effects on professional practice and health care outcomes (Review). *Cochrane Database Syst Rev.* 2012;(7):1–227. Eva KW, Armson H,

20 Holmboe E, *et al.* Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ Theory Pract.* 2012;17:15–26.

21 Mann K, van der Vleuten CP, Eva KW, *et al*. Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. *Acad Med*. 2011;86(9):1120–1127.
22 Bandura A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol Rev.* 1977;84(2):191–215.
23 Ajzen I. Nature and operation of attitudes. *Annu Rev Psychol.* 2001;52:27–58.
24 Prochaska J, DiClemente C, Norcross J. In search of how people change. Applications to addictive behaviors. *Am Psychol.* 1992;47(9):1102–14.
25 Burke L, Hutchins H. Training Transfer: An Integrative Literature Review. *Hum Resour Dev Rev.* 2007;6(3):263–296.
26 Lave J, Wenger E. *Communities of Practice. Cambridge: Cambridge university press;* 1999.
27 Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Evolution of Wenger's concept of community of practice. *Implement Sci.* 2009;4(1):11.
28 Le May A. Introducing communities of practice. In: le May A, ed. *Communities of practice in health and social care.* Oxford: Wiley-Blackwell; 2008:3–16.
29 Mazmanian PE, Mazmanian PM. Commitment to change: Theoretical foundations, methods, and outcomes. *J Contin Educ Health Prof.* 1999;19(4):200–207.
30 Prochaska JO, Redding CA, Evers KE. Health behavior and health education. In: Glanz K, Rimer B k, Viswanath K, eds. *Health behavior and health education: theory, research, and practice.* 4th ed. Wiley & Sons; 2008:97–121.
31 Van Zundert M, Sluijsmans D, van Merriënboer J. Effective peer assessment processes: research findings and future directions. *Learn Instr.* 2010;20(4):270–279.
32 Van Gennip NA, Seger MS, Tillema HH. Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learn Instr.* 2010;20(4):280–290.
33 Strijbos J-W, Sluijsmans D. Unravelling peer assessment: methodological, functional, and conceptual developments. *Learn Instr.* 2010;20(4):265–269.
34 Arnold L, Shue CK, Kalishman S, *et al*. Can there be a single system for peer assessment of professionalism among medical students? A multi-institutional study. *Acad Med.* 2007;82(6):578–86.
35 Dannefer EF, Henson LC, Bierer SB, *et al*. Peer assessment of professional competence. *Med Educ.* 2005;39:713–722.
36 Norcini JJ. Peer assessment of competence. *Med Educ.* 2003;37(6):539–543.
37 Wenghofer EF, Way D, Moxam RS, Wu H, Faulkner D, Klass DJ. *Effectiveness of an Enhanced Peer Assessment Program: Introducing Education into Regulatory Assessment*. 2006;26(3):199–208.
38 Topping KJ. Methodological quandaries in studying process and outcomes in peer assessment. *Learn Instr.* 2010;20(4):339–343.
39 Staal BJ, Hendriks EJ, Heijmans M, *et al*. KNGF Richtlijn Lage-Rugpijn voor fysiotherapie en manuele therapie [Guideline low back pain for physical therapy and manual therapy]. *R Dutch Soc Phys Ther.* 2010. Available at: http://www.fysionet-evidencebased.nl/index.php/component/kngf/richtlijnen.
40 Van Dulmen SA, Maas MJ, Staal B, *et al*. Effectiveness of peer-assessment for implementing a Dutch physical therapy low back pain guideline: a cluster randomized controlled trial. *Phys Ther.* 2014.
41 Heemskerk M, Staal J, Bierma-Zeinstra S, *et al*. KNGF-richtlijn Klachten aan de arm, nek en/of schouder (KANS) [KNGF-guideline complaints of arm, neck and/or shoulder (CANS). 2010;(1). Available at: http://www.fysionet-evidencebased.nl/index.php/component/kngf/richtlijnen.
42 Jansen M, Brooijmans F, Geraets J, *et al*. KNGF Evidence Statement Subacromiale klachten [Evidence Statement Subacromial complaints]. 2011;(1):1–14. Available at: http://www.fysionet-evidencebased.nl/index.php/component/kngf/richtlijnen.
43 Oostendorp RA, Pluimers DJ, Nijhuis-van der Sanden MW. Fysiotherapeutische verslaglegging: de Achilleshiel voor Evidence-based Practice (EBP)? [Record Keeping in Physical Therapy: The Achilles Heel for Evidence Based Practice (EBP)?]. *Ned Tijdschr voor Fysiother.* 2006;116(3):56.

44 van Dulmen SA, Calsbeek H, Cruijsberg J, Koetsenruijter J, Braspenning J. *Kwaliteitsindicatoren Eerstelijns Fysiotherapie [Quality Indicators Physical Therapy Primary Care]*. Nijmegen; 2011. Available at: http://www.iqhealthcare.nl/nl/kennisbank/rapporten/k/kwaliteitsindicatoren-eerstelijns-fysiotherapie-(kwaliefy)/#.Uz8IGPmKVcY.
45 Norman GR, Schmidt HG. The Psychological Basis of Problem-based Learning: A Review of the Evidence. *Acad Med.* 1992;67(9):557–365.
46 Greene J, Azevedo R. A Theoretical Review of Winne and Hadwin's Model of Self-Regulated Learning: New Perspectives and Directions. *Rev Educ Res.* 2007;77(3):334–372.
47 Bandura A. *Self-efficacy: The exercise of control.* (Anonymous, ed.). Freeman; 1997:604.
48 Epstein RM, Siegel DJ, Silberman J. Self-monitoring in clinical practice: a challenge for medical educators. *J Contin Educ Health Prof.* 2008;28(1):5–13.
49 Pronovost PJ, Hudson DW. Improving healthcare quality through organisational peer-to-peer assessment: lessons from the nuclear power industry. *BMJ Qual Saf.* 2012;21(10):872–5.
50 Maas CJ, Hox JJ. Sufficient Sample Sizes for Multilevel Modeling. *Methodology*. 2005;1(3):86–92.
51 Dallal GE. Randomization.com. 2008. Available at: http://www.randomization.com. Accessed February 2, 2012.
52 Heerkens YF, Hendriks H, De Graaf-Peters VB. KNGF-richtlijn Fysiotherapeutische verslaglegging [KNGF-guideline record keeping in Physcial Therapy]. 2011. Available at: https://www.fysionet-evidencebased.nl/index.php/richtlijnen/richtlijnen/fysiotherapeutische-verslaglegging-2011.
53 Available at: http://www.han.nl/onderzoek/kennismaken/revalidatie-arbeid-sport/lectoraat/arbeid-en-gezondheid/projecten/.
54 Rutten GMJ, Harting J, Rutten STJ, Bekkering GE, Kremers SPJ. Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *J Eval Clin Pract.* 2006;12(5):491–500.
55 Peabody JW, Luck J, Glassman P, Dresselhaus TR, Lee M. Comparison of Vignettes , Standardized Patients , and Chart Abstraction. *JAMA*. 2000;283(13):1715–1722.
56 Peabody JW, Dresselhaus TR, Luck J, Bertenthal D. Improving Patient Care Measuring the Quality of Physician Practice by Using Clinical Vignettes : A Prospective Validation Study. *Ann Intern Med.* 2004;141(10):813–814.
57 Dresselhaus TR, Peabody JW, Luck J, Bertenthal D. An evaluation of vignettes for predicting variation in the quality of preventive care. *J Gen Intern Med.* 2004;19(10):1013–8.
58 Childs JD, Whitman JM, Sizer PS, Pugia ML, Flynn TW, Delitto A. A description of physical therapists' knowledge in managing musculoskeletal conditions. *BMC Musculoskelet Disord*. 2005;6(32):1–7.
59 Charlin B, van der Vleuten CPM. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Eval Health Prof.* 2004;27(3):304–319.
60 Grant AM, Franklin J, Langford P. The Self-reflection and Insight Scale: A New Measure Of Private Self-consciousness. *Soc Behav Pers.* 2002;30(8):821–836.
61 Grant AM. Personal life coaching for coaches-in-training enhances goal attainment , insight and learning. *Coach An Int J Theory, Res Pract.* 2008;1(1):54–70.
62 Roberts C, Stark P. Readiness for self-directed change in professional behaviours: factorial validation of the Self-Reflection and Insight Scale. *Med Educ.* 2008;42(11):1054–63.
63 Wakefield J, Herbert CP, Maclure M, *et al*. Commitment to change statements can predict actual change in practice. *J Contin Educ Health Prof.* 2003;23(2):81–93.
64 Grol R, Bosch M, Wensing M. Ontwikkeling of selectie van strategieen voor verandering [Development or selection of strategies for change]. In: *Implementatie: Effectieve verbetering van de patiëntenzorg [Implementation: effective improvement of patient care]*. 4th ed. Amsterdam: Reed Business; 2011:281–323.
65 Jamtvedt G, Young J, Kristoffersen D, O'Brien M, Oxman A. Audit and feedback: effects on professional practice and health care outcomes (Review). *Cochrane Database Syst Rev.* 2007;(4):CD000259.

66 Scott SD, Albrecht L, O'Leary K, *et al*. Systematic review of knowledge translation strategies in the allied health professions. *Implement Sci.* 2012;7(1):70.
67 Van Gennip NAE, Segers MSR, Tillema HH. Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educ Res Rev.* 2009;4(1):41–54.
68 Sargeant JM, Mann K V, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract.* 2009;14(3):399–410.
69 Ramsey PG, Wenrich MD, Carline JD, Inui TS, Larson EB, LoGerfo JP. Use of peer ratings to evaluate physician performance. *JAMA.* 1993;269(13):1655–1660.
70 Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci.* 2006;17(3):249–255.
71 Kromann CB, Jensen ML, Ringsted C. The effect of testing on skills training. *Med Educ.* 2009;43(1):21–27.
72 Simons P, Ruijters M. Learning professionals (ed): towards an integrated model. In: Boshuizen H, Bromme R, Gruber H. (eds). *Professional learning: Gaps and transitions on the way from novice to expert*. Dordrecht: Kluwer; 2004:207–229.
73 Iacoboni M. Mirroring People: *The new science of how we connect with others*. New York: Macmillan; 2009.
74 Bandura A, Locke EA. Negative self-efficacy and goal effects revisited. *J Appl Psychol.* 2003;88(1):87–99.
75 Bekkering GE, Tulder MW Van, Hendriks EJ, *et al*. Implementation of clinical guidelines on physical therapy for patients with low back pain : Randomized trial comparing patient outcomes after a standard and active implementation strategy. *Phys Ther.* 2005;85(6):544–555.
76 Hrisos S, Eccles MP, Francis JJ, *et al*. Are there valid proxy measures of clinical behaviour? A systematic review. *Implement Sci*. 2009;4:37.

Appendix 1
Example of a clinical vignette with two exemplary test items

Vignette 3
Part 1: case history

Personal information: female, 38 years old, married, two daughters (21 and 14 years old).
Work: nurse, three days a week
Hobbies: reading, playing cards
Sports: nothing specific except for making long walks

Marianne visits the physical therapist without referral of a physician. She has suffered from intense right shoulder pain for a week. She reports existing neck complaints and vague shoulder pain for a few months. Shoulder complaints were conspicuous in activities such as changing a drip or moving patients in their bed. She was also hindered in driving her car, because of limited rotation of the cervical spine. These problems, however, did not have serious impact on her daily life. Complaints seriously aggravated when she went swimming with her two nieces aged 5 and 7. The girls repeatedly climbed on her shoulders to subsequently dive into the water. Shoulder pain gradually developed the same evening and increased the following days. Because of the pain, she is currently unable to lift her right arm without the support of her left arm, which hinders her in activities such as dressing, undressing, and other activities involving raising her right arm. She locates the pain at the antero-lateral aspect of her right shoulder and upper arm. Currently, she is unable to perform her nursing tasks which she perceives very annoying, because there is much failure of nursing staff lately.
At night, Marianne has difficulty in sleeping. She easily wakes up when she rests on her painful shoulder. She uses paracetamol and ibuprofen as medication, and that relieves the pain, but not enough to perform her daily tasks.
Marianne still smokes after repeated attempts to quit, and her youngest daughter uses every opportunity to comment on her smoking habit. Besides a hypofunction of the thyroid, for which she receives medication, Marianne has no other health problems. Apart from the last few months, she never had shoulder problems before. Marianne is right handed.

Example of a response category and related test items

Which risk factors may have played a role in the onset of the symptoms in this health problem according to the guideline? D = disagree D/A = disagree nor agree A = agree	**Distribution of expert scores (n=8)**			**Rewarded points**		
	D	**D/A**	**A**	**D**	**D/A**	**A**
Age	0	0	8	0	0	2
Gender	0	4	4	0	1	1
Medication use	7	1	0	2	1	0
Neck complaints	0	0	8	0	0	2
Postures related to work	0	4	4	0	1	1
Movements related to work	0	0	8	0	0	2
Amount of working hours	8	0	0	2	0	0
Arm dominance	0	2	6	0	1	2
Coping with stress	0	2	6	0	1	2
Smoking	6	2	0	2	1	0
Hormonal changes	8	0	0	2	0	0

Chapter 6
Feasibility of peer assessment and clinical audit to self-regulate the quality of physical therapy services: a mixed methods study

Marjo Maas
Ria Nijhuis-van der Sanden
Femke Driehuis
Yvonne Heerkens
Cees van der Vleuten
Philip van der Wees

BMJ Open. 2017;7:1-10

Abstract

Objectives

To evaluate the feasibility of a quality improvement program aimed to enhance the client-centeredness, effectiveness, and transparency of physical therapy services by addressing three feasibility domains: 1) acceptability of the program design, 2) appropriateness of the implementation strategy, and 3) impact on quality improvement.

Design

Mixed methods study

Participants and setting

64 physical therapists working in primary care, organized in a network of communities of practice in the Netherlands

Methods

The program contained: 1) two cycles of online self- and peer assessment of clinical performance using client records and video-recordings of client communication followed by face-to-face group discussions, and 2) clinical audit assessing organizational performance. Assessment was based on pre-defined performance indicators which could be scored on a 5-point Likert scale. Discussions addressed performance standards and scoring differences. All feasibility domains were evaluated qualitatively with two focus groups and 10 in-depth interviews. In addition, we evaluated the impact on quality improvement quantitatively by comparing self- and peer assessment scores in cycle 1 and 2.

Results

We identified critical success features relevant to program development and implementation, such as clarifying expectations at baseline, training in peer assessment skills, prolonged engagement with video-assessment, and competent group coaches. Self-reported impact on quality improvement included awareness of clinical and organizational performance, improved evidence-based practice and client-centeredness, and increased motivation to self-direct quality improvement. Differences between self-scores and peer scores on performance indicators were not significant. Between cycle 1 and cycle 2, scores for record keeping showed significant improvement, however not for client communication.

Conclusions

This study demonstrated that bottom-up initiatives to improve healthcare quality can be effective. The results justify ongoing evaluation to inform nationwide implementation when the critical success features are addressed. Further research is necessary to explore the sustainability of the results and the impact on client outcomes in a full scale study.

Introduction

Healthcare professionals and provider organizations have an ethical and professional obligation to strive for continuous quality improvement of services. When healthcare professionals are able to self-regulate and account for the quality of their services, they perceive control of the quality improvement strategies and the outcome measures used, in contrast to external regulations. Professionals often resist external audits; they fear a deterioration of their professional identity and an increase of administrative burden. Moreover, external regulations can potentially be effective, but the evidence is not convincing regarding the sustainability of the results and the strategy might induce unwanted consequences such as under-treatment of clients with multi-morbidity or disparities in healthcare delivery.[1-3]

Research has shown that bottom-up quality improvement initiatives, such as communities of practice and professional networks focusing on collaborative learning, might hold better and more sustainable results than external, top-down regulations[4-6] because shared social and professional norms are important predictors for behavior change.[7,8] Conditional to successful self-regulation is that professionals share the quality standards of their services and demonstrate the willingness and ability to critically appraise their own and their colleagues' performance.[4,6,9] Literature showed that quality improvement programs targeting self-regulation should not be limited to individual healthcare professionals, but also involve teams and provider organizations to align the desired processes and outcomes.[10]

Clinical governance has been introduced as a multi-level approach to quality improvement, bridging the gap between managerial and clinical approaches. The approach allows for early spotting of poorly performing clinicians, teams, and organizations to support self-regulated quality improvement.[4,11,12] Following this approach, we developed and tested the feasibility of a program combining self-assessment, peer assessment (PA) and clinical audit as a

strategy to improve the quality of physical therapy services. Self-assessment is the process whereby professionals reflect on their clinical and organizational performance according to quality indicators.[13] In PA – also known as peer review – professionals evaluate or are being evaluated by their peers and provide each other with performance feedback.[14] The aim of PA is to guide the self-directed quality improvement process towards desired mutual accepted performance standards.[15] A cornerstone of the PA strategy is to raise awareness of clinical performance informing self-assessment,[16] and to develop a critical attitude towards the process and outcomes of healthcare by introducing professionals with an 'assessor' or 'auditor' perspective.[17] The results of a systematic review by Fox *et al.*[18] on PA practices in healthcare demonstrated that peer review was associated with measurable performance improvement of healthcare professionals on several outcome levels and in a variety of competency domains. Clinical audit is a common strategy for quality improvement at the level of provider organizations. Aspects of the structure, processes, and outcomes of care are selected and systematically evaluated against explicit criteria by trained colleagues. The method has proved its effectiveness in primary care.[15,19] Although PA and clinical audit are well-studied strategies,[20-22] our program design is innovative because it focuses on assessment of authentic clinical behaviors, and integrates clinical performance and organizational performance assessment.
We used the framework of the Medical Research Council 2015 to develop, test and implement the program.[23] The framework recommends a feasibility and piloting phase to allow for optimizing the program design and implementation prior to evaluating effectiveness in a larger study. The appendix shows the details of the design process including the development of performance indicators.
This study addresses the evaluation of three feasibility domains: 1) acceptability of the program for quality improvement purposes including strengths and weaknesses, 2) appropriateness of the implementation strategy to execute the program as intended including barriers and facilitators, and 3) impact of the program on quality improvement and professional behavior change.[23,24]

Aim

The aim of this study was to evaluate the feasibility of a quality improvement program aimed to enhance the effectiveness, client-centeredness, and transparency of physical therapy services to allow for optimizing the program design and implementation prior to more rigorous evaluation and nationwide implementation.

Methods

Design

Program feasibility was evaluated with mixed methods using qualitative and quantitative data.[24]

Subjects and setting

We tested our program with physical therapists working in primary care clinics, organized in a regional network of communities of practice in the Netherlands. In communities of practice professionals share the same interests, setting, or specialization. Formally registered physical therapist networks were invited by the Royal Dutch Society for Physical Therapy (KNGF) via a digital newsletter. Participation was voluntary and was awarded with 30 accreditation points for the quality register. To facilitate the implementation process, we used two knowledge brokers (FT and HK) as the linking pin between researchers and participants,[25-27] trained PA coaches (n=5) to support the PA process, and trained auditors (n=3). The knowledge brokers were leaders of the professional network and took part of the stakeholder group. Coaches and auditors were members of the professional network recruited by the knowledge brokers.

Program content

Assessment addressed three performance domains related to the three quality domains: client-centeredness, effectiveness (including evidence-based practice), and transparency of physical therapy services: 1) record keeping, 2) client communication, and 3) organization and management. The program contained two cycles of PA with an interval of 4-6 weeks, followed by one cycle of clinical audit. Assessment was based on pre-defined performance indicators which could be scored online on a 5-point Likert scale. The appendix presents the performance indicators and their relationships with the three quality domains. Assessment was based on pre-defined performance indicators which could be scored on a 5-point Likert scale.
Assessment of clinical performance included online self-assessment and PA of 1) client records and 2) video-recordings, followed by 3) face-to-face discussion of the results, supported by a trained group coach. Participants were assigned to upload one electronic client record and one video-recording of client communication – limited to the discussion of the diagnosis and treatment plan – in both cycle 1 and 2. Before assessing their peers, the participants self-assessed their performance using the same indicators. Peers provided online scores and written improvement feedback if relevant. Discrepancies in scores were used as input for the subsequent

discussions. After the first PA session, participants designed an online personal improvement plan; during the second session they reflected on the improvements made. Participants who objected to uploading videotapes were allowed to choose for role-playing instead; they simulated the client conversation and their performance was assessed on the spot.
PA coaches received a program guide and two additional training sessions conducted by professional trainers (MM and PW) using samples of client records, video-recordings, and role-play to train the process of providing, receiving and using feedback.
Clinical audit was scheduled after all PA activities were completed. A convenience sample of four private clinics was invited to participate. These clinics provided online management information according to the program guidelines and self-assessed their organizational performance using an online scoring sheet (appendix). Clinical audit included inspection of the property, assessment of two at random selected client records, and discussion of self-assessment scores for management and organization. Afterwards a report was written according to a structured reporting format. Participants were invited to comment on the clinical audits report before it was finalized. Auditors received a program guide and a training conducted by professional auditors (MA and MB) using worked samples and role-play to train the auditing process.

Website

We developed a web-based assessment system that allowed for 1) downloading program guides and instruction manuals, 2) uploading assessment materials such as client records, video-recordings, management information, and improvement plans, 3) online scoring, 4) downloading assessment results, and 5) storing and exporting qualitative and quantitative data. See Appendix for additional information.

Program delivery

A research team member (FD) was the program manager. She provided participants of a program guide including a manual for uploading and downloading assessment material and guidelines for providing and using quality improvement feedback.[16,22,28-31]

Ethical issues

All participants gave their online informed consent. Clients providing video-recordings and client records gave their written informed

consent. This study was approved by the Medical Ethical Committee Arnhem Nijmegen (CMO): 2015-1797.

Evaluation of program feasibility

All three feasibility domains were explored with focus groups, in-depth individual interviews, and written coach reports aiming for saturated information from multiple perspectives to optimize the credibility and transferability of the results.[32] In addition, we evaluated the impact on quality improvement quantitatively by comparing self-assessment and PA scores, and cycle 1 and cycle 2 scores.

Data sampling and analyses of interviews

We aimed to bring together coaches, clinic visitors and knowledge brokers in separate focus groups to explore their experiences in performing the same role. Individual participants were purposely sampled for in-depth interviews including all participants in clinic clinical audits. The website provided us with data to identify and select little, moderate, and very active PA participants. They were approached by email. An interview guide was designed by the research team (MM, FD, PW, RN) addressing the three feasibility domains tailored to the participant role. Focus groups lasted 90-100 minutes and were conducted face-to-face by MM and PW using open-ended questions allowing for group discussion and knowledge construction. In-depth interviews permitted us to explore thoughts and feelings that might not easily be shared with colleagues, but relevant to understanding participant's behavior; they were conducted by MM or FD (MSc, health scientist, quality of healthcare research) using teleconferencing technology and lasted 50-60 minutes. Interviews of all participants, including verbal consent, were audio-taped and transcribed verbatim afterwards. The analytic process was guided by 'template analysis' that combines a-priori codes informed by the research questions with emerging codes from the interview data.[33] PW and MM independently studied and coded five transcripts. Differences in coding were discussed and a code book was created based on consensus. Subsequently, all transcripts were analyzed line-by-line, using ATLAS-ti v.7 software. Codes were compared and some codes were merged into higher-order codes. Emerging themes were identified by constant comparison of codes and higher order codes. Finally, we summarized the results relevant to ongoing program development and implementation.[34] To increase the credibility of the results, a peer debriefing and member checking procedure was conducted with research group members FD and RN and with knowledge brokers and stakeholders.

Data sampling and analyses of scores

Online scores for record keeping and client communication in the first and second assessment cycle were imported in IBM-SPSS Statistics 22. Indicators that were scored as 'not relevant' or 'not applicable' were treated as missing values.
The mean and median indicator scores for each performance indicator and for each performance domain were calculated for self-scores and PA scores as well as the percentages of missing values. We used the Wilcoxon Signed Rank test to calculate differences between self-assessment and PA scores, and between cycle 1 and cycle 2 scores, including *P* values for statistical significance.

Results

In total 64 physical therapists took part in the program. Twelve peer groups were formed based on specialization, each containing 4-6 participants. Eleven peer groups participated in online PA; one group used printed scoring sheets. Three group clinics and one solo clinic participated in clinical audits. Table 1 shows an overview of participants' characteristics.

Qualitative results

We conducted 2 focus groups and 10 in-depth interviews reaching data saturation. Results are discussed using pre-defined categories including references to quotes labelled by number and participant's role (KB=knowledge broker, C=coach, V=visitor, P=participant). Quotes are presented in table 2. We identified strengths and weaknesses of the program design, implementation barriers and facilitators, the impact on quality improvement, and critical success features relevant to program development and implementation. In table 3 the results are summarized.

Acceptability of the program design

General perceptions

At the beginning participants were skeptical regarding the feasibility of the program aims and procedures. Frustrated by the quality demands of health insurers, they were not seeking for an extra administrative burden. However, their views changed along with the program (Q1-P7). Looking back participants were positive about the program, because it focused on their core-business and uncovered 'what happens behind closed doors' (Q2-P4).
Despite the guidelines for constructive feedback, providing and

Table 1 — Participant demographics and characteristics

Individual characteristics (n^1=64)		National (N^2=17.802)
Mean age in years (SD)	50 (10.1)	42
Gender: woman %	50%	56%
Communities of practice characteristics and number of participants		
General conditions	26	
Respiratory conditions	13	
Cardiovascular conditions	10	
Psychosomatic conditions	4	
Neurologic conditions	5	
Geriatric conditions	10	
Total	*68**	

n^1 = number of participants;

N^2 = Number of physical therapists in the Netherlands working in primary care;

* Four physical therapists participated in two groups.

receiving feedback were not self-evident. Participants struggled with critically appraising their peers. Being insecure about their own performance, they were cautious in providing critical feedback resulting in 'halo marking' (too high) as supported by the quantitative data (Q3-P6). Experiencing a safe setting, allowing to make mistakes, was perceived as conditional to critical peer appraisal. Regarding receiving feedback, some participants faced difficulties with adequately responding to it (Q4-P9). Participants were unanimous in their view that feedback should be critical to enable meaningful improvement. Compliance to program guidelines and shared responsibility for group learning was perceived as critical to program efficacy (Q5-KB). Although participants generally accepted the program for quality improvement purposes, some of them reported that both client records and videotapes might present an overly optimistic picture of clinical practice because they were self-selected (Q6-P8).

Assessment of client communication

Some participants objected to making video-recordings, unwilling to put an unnecessary load on their clients, worrying about client privacy, and assuming that they would not consent. Although clients rarely objected to it – in contrast to participant's pre-assumptions – online assessment activities in cycle 1 were limited (Q7-C). Initial reluctance disappeared when participants personally experienced

the added value of video assessment by simply 'doing it', or by watching others 'doing it'; worked samples enhanced the acceptability. It became clear that peer groups needed time and deliberate practice to get used to video-assessment and to feel safe enough to expose their clinical performance. Participants who preferred video-assessment to client records, argued that this instrument allows to observe what physical therapists 'do' instead of what they 'say they do'. They agreed that 'taking a look inside' provided valuable information, such as attitudes becoming observable (Q8-P6). On the one hand, video-recordings allowed for modelling professional behaviors of skilful colleagues, on the other hand, unwanted behavior became transparent triggering suggestions for alternative behaviors, especially regarding the efficiency of chronic disease management (Q9-P9).

Although online peer scores did not always reflect the ability or willingness of participants to critically appraise their peers, during the sessions feedback quality increased by comparing self-perceptions with peer perceptions and by discussing quality standards of performance. Participants who consciously selected their best videotape, could be confronted with different views on quality indicators (Q10-P1).

Participants who preferred assessment of client records to video-recordings, argued that they felt uncomfortable with the knowledge that their conversation was recorded ('audience effect') or that a 'snapshot' poorly represents the process of patient management (Q11-P2).

Assessment of client records

Assessment of record keeping was valued because client records present the process of patient management unlike video-tapes, allowing to assess clinical reasoning and decision making such as the application of clinical practice guidelines, the use of client reported outcome measures and performance outcome measures. Here again face-to-face discussions were critical to an in-depth understanding of quality indicators e.g. for evidence based practice. For example, one of the knowledge brokers noticed that 'clicking on the guideline button' in the electronic record system, indicating the use of a particular guideline, was no guarantee for adequately 'applying' the guideline in the specific context of the patient problem (Q12-KB). In contrast to feedback provided by professional auditors, peer feedback was perceived as a good vehicle to self-direct improvement (Q13-P3).

Clinical audit

Participating private clinics all appreciated clinical audits. They reported that they were 'pretty nervous' in advance, but valued the safe setting allowing for discussion of strengths and weaknesses, providing them feedback to guide improvement of management and organization towards its quality standards and giving them back responsibility and ownership (Q14-P4; Q15-P5). They all agreed with their audit reports, providing minor comments, although reports were perceived as more formal than audits.

Appropriateness of the Implementation strategy

Motivational issues

At baseline, participants were poorly informed about the program aims, intended outcomes and consequences. Frustrated by the dominant role of insurers in quality control, participants were suspicious about whose interests were being served by their extra efforts affecting their motivation to participate. Alignment of expectations, might have prevented false cognitions and enhanced motivation to invest time and effort (Q16-P7).

Communication technology support

Although the website has been improved continuously throughout the program, it was not perceived as user-friendly causing feelings of frustration (Q17-P10). Despite the supply of a user manual, peer support and learning by doing turned out to be more effective.

The role of knowledge brokers, group coaches and auditors

Although the knowledge brokers were involved in writing the program guide, they didn't succeed in adequately inform the participants. Their role as linking pin between researchers and clinicians required advanced communication and leadership skills (Q18-KB).

The role of the coach was perceived as crucial in facilitating critical reflection and an in-depth understanding of quality standards. However, some group coaches had to deal with the 'wait and see' attitude of some participants who did not provide online materials in time. They lacked the coaching skills to support active participation and shared responsibility for group learning (Q19-C).

Some clinic auditors struggled with their role identity. They were trained to communicate what they observed regarding the quality indicators; as such they felt competent to provide information on 'what' can be improved (feedback), but not on 'how' (feed forward) appealing to a counsellor role rather than an auditor role (Q20-V).

Table 2 — Quotes of participants

Acceptability of the program design	
General perceptions	
Q1-P7	*"Some of my colleagues were very critical, but now their views are changed. In particular because the program was 'again' about quality standards that we must meet (…). First the health insurers with their audits and now this (…). We don't want more paperwork."*
Q2-P4	*"I think the system is appropriate. In fact nobody evaluates you this way. No one comes so close; (…) no one comes into your room and that's how it felt somehow. Yes, perhaps a trainee, or you consult a colleague to look at your patient's problem, but you never ask your colleague to evaluate you; (…) we are very much loners in this respect."*
Q3-P6	*"I think it (critical appraisal of peer performance) needs time to develop. I think it will come by doing it a number of times (…). You need to feel safe enough to trust."*
Q4-P9	*"In the beginning – I think we needed to get used to it – I saw some participants instantly responding by defending themselves. But I also observed – probably because there was enough empathy and respect for each other – that there was no need to. I saw that (responding to feedback) gradually improved."*
Q5-KB	*"Because there are always two or three early adopters and the rest is lagging behind. (…) I think it's that sense of responsibility that you need as colleagues to get these things right."*
Q6-P8	*"Well, I would do the same. I always say: 'When someone comes to have a look into your kitchen, you make sure that it is cleaned'."*
Assessment of client communication	
Q7-C	*"The first time we (the group majority) chose to role-play, but we also watched two videotapes. After that, everybody said: 'we'll go for the videotapes the next time, these are far more interesting'. And by doing that, we already made some improvements."*
Q8-P6	*"Well, actually it was enjoyable (…). It is quite surprising to see how your colleagues, that's how you know them, how they interact with their client. That provides a lot of information to reflect on. I think attitudes are very important and client records are such a long stories - of course important – but in particular those videos were interesting. Although I also noticed that everyone had more trouble with it.*
Q9-P9	*"We saw a COPD-patient [on the video]. It didn't become clear how long she intended to treat this patient?' You often miss some kind of timeline in chronic disease management. I understand that it is not easy, but you can help yourself by setting an evaluation moment."*
Q10-P1	*"When they saw my videotape they commented that it was 'big' … I am a bit wordy, that's what I am, that's what I have been doing for thirty years now. Some said it was OK, but one said: 'actually, I scored a 2' [for patient centeredness], meaning that much improvement was needed. I was shocked, because I scored myself with a 4, I thought I was not bad (…). The patient [on the video] said 'yes..yes' all the time. I thought the patient was agreeing, but I should have asked. That was really confronting for me."*
Q11-P2	*"… it was great, it was fun, but the tapes were pretty short (…), snapshots of six minutes. I think an electronic client record provides much more information when it comes to critical appraisal."*
Assessment of record keeping	
Q12-KB	*"I think we need to address [clinical practice] guidelines. Put them on the table, show them. We know that they exist, but little effort has been made to applying them (…). Everybody has them on their book shelf, but no one knows the content, well … that might be an exaggeration".*

Q13-P3	*"Yes, it was insightful, it confirms what you do and what you don't do. Nobody ever taught me how to keep my records, yes … I once took a course, but that was twenty years ago… you keep records according to your best knowledge; you don't receive feedback until you are audited. Now your colleagues can guide you, I perceived that as very helpful."*
Assessment of management and organization	
Q14-P4	*"Well, I think it's very appropriate. In this way – unlike the health insurer - your colleagues come to visit you, it feels more like feedback … because it allows you to create real improvements for yourself"*
Q15-P5	*"It is very important that people really feel that they can improve, instead of being challenged. And that's the basis on which people dare to do this."*
Appropriateness of the implementation strategy	
Q16-P7	*"The minister [Public Health] has cooked up all this and meanwhile the health insurers laugh themselves to death … I think it was fun, we had a nice group and we definitely learnt from each other, but it feels incredibly bad that we only invest and never get something back."*
Q17-P10	*"We all struggled with the website somehow. For me, I'm not skilled in computer work. Therefore, I asked my colleague for help in the beginning. But once you start working with it, it becomes more clear. For instance, I succeeded better in uploading the second case than the first one."*
Q18-KB	*"I never realised that when you want to run such a project, that it takes so much effort to keep PTs focused."*
Q19-C	*"This morning Karen told me: 'In my view, the most difficult thing of coaching was to activate my group', so, looking back, I think that we should have addressed this issue more extensively during our training'.*
Q20-V	*"It was difficult to see that C. was very insecure (…). I thought: 'I am not a monster? Shall I comfort her by saying that she faces the same problems as I do, and that making mistakes is human?' You actually want to help, by saying 'do this, or try that' (…). It's difficult to stay neutral."*
Impact on quality improvement and professional behaviour change	
Q21-P2	*"Yes, according to the guidelines, an EPR includes a baseline-measurement, an in-between measurement, and a final evaluation. That is how it should be, but I don't always meet that standard. I promised to improve that."*
Q22-P9	*"I need to make clear what we are going to do [treatment] and when are we going to stop. Especially since I am dealing with clients with a chronic condition."*
Q23-P1	*"… my record keeping was alright, but regarding patient communication … I explained a lot, but I didn't check to see if my message was understood. That's an improvement I need to make. I try now to ask my patient: 'What did you learn about what I explained just now?' I consciously address that issue. Moreover, I pay more attention to their personal goals. I can have a plan, but that plan might not be in line with their expectations … I might be too dominant in this respect because I think I know what they need, but I should not think for them."*
Q24-KB	*"You know, we saw clinical practice and science bump into each other, and now I think that we have finally come together, pushing, pulling and rolling the same cart."*

Q=quote; P=participant; C=Coach; KB=knowledge broker; V=visitor

Table 3 — Summary of findings

Acceptability of the program design		
Strengths	**Weaknesses**	**Critical success features**
General perceptions Focus on the core-business of physical therapists. Uncovers 'what happens behind closed doors'.	Limited validity client records and videotapes because they are self-selected. Limited validity of online scores due to unwillingness or incompetence to adequately apply performance indicators.	Training in critical performance appraisal to support self-directed quality improvement. Time to build a safe setting allowing to make mistakes. Face-to-face discussions of discrepancies in online scores to compare self-perceptions with peer perceptions. Active participation. Compliance to program guidelines. Safe setting.
Assessment of client communication Shows what physical therapists 'do' instead of what they 'say they do'. Uncovers undesired attitudes. Allows for modelling desired behavior.	Reluctance to expose clinical performance to an 'audience'. Snapshot, poorly representing the process of patient management.	Using worked samples of video-assess ment to enhance its acceptability. Extended engagement with video-assessment.
Assessment of record keeping Presents the process of patient management allowing to assess clinical reasoning and evidence based practice.		
Assessment of management and organization Provides guidance to self-direct improvement.		

Appropriateness of the implementation strategy		
Barriers	**Facilitators**	**Critical success features**
Program aims, expected efforts, and desired outcomes insufficiently clarified at baseline. Dominant role of insurers in quality control causing doubts about the stakeholders in the quality improvement program. Poor program efficacy beliefs.	Learning by 'doing' or by watching others doing it (role models). Emphasis on learning and improvement instead of judgment.	Discussion of program aims, desired results and consequences on the long term to clarify and align expectations. Shared responsibility for group learning and quality improvement program outcomes.
Absence of financial incentives. Complex website design.	Awarding efforts with credits. Peer coaching in using communication technology.	User friendly website design.
Limited skills of group coaches to enhance shared responsibility for group learning and results.		Competent group coaches.

Impact on quality improvement and professional behavior change		
Individual level	**Organizational level**	**Network level**
Awareness of clinical performance.	Awareness of organizational performance. Increased self-efficacy and program-efficacy beliefs.	Increased self-efficacy beliefs and motivation for ongoing PA activities. Commitment to ongoing PA activities and clinical audits.
New insights in the application of clinical practice guidelines, the use of client reported outcomes and performance measures. Improved client involvement in goal setting and treatment planning. Improved peer assessment skills.		

Table 4 — Differences between self-assessment and peer assessment scores and differences between cycle 1 and cycle 2 scores tested with non-parametric Wilcoxon signed Ranks Test (Likert scale 1-5).

							Differences between SA[1] and PA[2] scores		Differences between cycle 1 and cycle 2 scores	
	N	Mean	M/R/A[3]	Median	Min	Max	Md[4]	*P*-value	Md[4]	*P*-value
SA[1] client communication cycle 1	9	3.52	5.6%	3.60	2.33	4.75			0.10	.674
PA[2] client communication cycle 1	11	3.79	3.9%	3.77	3.25	4.78	0.27	.263	- 0.09	.386
SA Record keeping cycle 1	26	3.43	3.2%	3.50	2.00	5.00			0.20	.007**
PA Record keeping cycle 1	31	3.41	1.6%	3.60	1.25	4.50	-0.02	.760	0.15	.002**
SA client communication cycle 2	40	3.62	5.8%	3.67	2.20	5.00				
PA client communication cycle 2	45	3.70	6.4%	3.80	2.21	4.50	0.08	.274		
SA record keeping cycle 2	48	3.62	2.8%	3.70	1.67	5.00				
PA record keeping cycle 2	63	3.75	3.3%	3.75	2.08	4.58	0.13	.269		

[1]SA = Self-assessment; [2]PA = Peer assessment

[3]M/R/A = Mean percentage of Missing / perceived not Relevant / perceived not Applicable indicator scores.

[4]Md = Mean difference

** Significant at a 01 level.

Impact on quality improvement and professional behavior change

The program impacted on different levels of professional practice, providing feedback to individuals, peer groups, and clinics. Regarding professional development, positive feedback enhanced self-efficacy beliefs and motivation to participate in continuing PA activities. Intentions to behavior change focused on guideline adherence, performance measurement and client reported outcome measurement.(Q21-P2; Q22-P9; Q23-P1). On the level of organization and management, participants reported improved awareness of strengths and weaknesses and increased beliefs in the change capacity of the program (Q14-P4; Q15-P5).

The collaboration between the research team and the network of participating physical therapists resulted in context-specific knowledge, relevant to ongoing quality improvement activities. The network committed to continue with PA and clinical audits, intending to address its critical success features (Q24-KB).

Quantitative results

Table 4 presents the results of the online uploaded data on the website showing that online activities varied widely and that

participation in cycle 1 was substantially lower than in cycle 2. Perceived barriers to online activities are reported in the qualitative results section. Except for record keeping in cycle 1, peer scores were higher than self-scores but differences were not significant. Since participants' online activities were low in cycle 1, data on the improvements made are limited. As shown in the blue printed area of table 4, differences between cycle 1 and 2 were not significant for client communication, but significant for record keeping, especially regarding the lower performers at baseline. Note that these differences relate only to the limited number of participants who were active in both cycles.

Discussion

This study focused on the feasibility of a quality improvement program aiming to enhance the client-centeredness, effectiveness, and transparency of physical therapy services. The qualitative results showed that participants viewed the program as an *acceptable* intervention for quality improvement purposes, allowing for stepwise self-directed quality improvement unlike the one-shot assessments of external auditors. We identified its critical success features such as training in performance appraisal and time to build a safe setting. Regarding the *appropriateness* of the implementation strategy to execute the program as intended, participants reported several facilitators and barriers, allowing us to identify critical success features for broader implementation such as adequate communication of program aims and intended outcomes at baseline, user-friendliness of the website design, and competent group coaches. The weaknesses of the program design and the barriers to program implementation affected the impact on quality improvement and behavior change. However, we identified meaningful self-reported results including awareness of clinical and organizational performance, improved evidence-based practice and client-centeredness, and increased motivation to self-direct quality improvement. The assessed (quantitative) results showed that online activities were low in cycle 1, providing limited data on the improvements made in cycle 2. Despite the limited data, we observed significant improvement of self-scores and peer scores for record keeping.
When we look at program acceptability, participants' views on the validity and the learning value of video-recordings and client records differed. We suggest that the acceptability of videos could be improved. Instead of using two single video clips, perceived as 'snap-

shots', several video recordings would provide more valid information as showed by a study of Ram *et al*.[35] However, that involves additional time and costs and might threaten long-term feasibility. Assuming that each instrument to assess professional performance has its advantages and disadvantages (standardized clients, direct observation, multi-source feedback) and that there is no single best measure as shown by a systematic reviews of Overheem *et al*.[36], the use of multiple measures is justifiable and even desirable for the purpose of gathering valid and reliable information on clinical competence.[37]

Regarding program implementation, we assume that the socio-political context – the dominant role of health insurers in quality assurance – impacted heavily on commitments to change and outcome expectancies.[38] Although PA aimed to provide formative feedback, emphasizing learning and improvement, it was viewed as summative assessment as physical therapists questioned the interest of stakeholders in their personal efforts. Improved communication at baseline might have enhanced participant's motivation and adherence to program guidelines. Moreover external, top down empowerment, a trade-off between trust and control, might be critical to successful outcomes on the long term as recognition of professional accomplishment and innovation is a strong motivator of improvement.[1]

Looking at the impact on quality improvement, we observed that peer scores for client communication were consistently higher than self-scores, demonstrating that participants either underestimated their own performance or overestimated their peers. In contrast to the literature on self-assessment showing that physicians generally overestimate themselves,[39] we assume that feelings of insecurity underlie both over- as underestimation in this case as supported by the qualitative data. Extended exposure to critical appraisal and reinforcement of constructive feedback practices could strengthen self-efficacy beliefs according to Bandura's cognitive learning theory.[40] The results also showed that the program was more effective in enhancing record keeping skills, than communication skills. Apparently, it took more time or effort to develop communication skills within the time span of the program. This assumption is supported by feedback intervention theory[31,41] explaining that the effectiveness of performance feedback is lower when the 'task novelty' and 'task complexity' is higher. Trained by audits of health insurers, participants were more familiar with assessment of record keeping. Moreover, the literature showed that clinical competency is content and context specific, meaning that competent (complex) behavior in one case (cycle 1) is a poor predictor for another case (cycle 2),[42,43] and this also applies to communication skills.[44] Although this program was not intended to

produce generalizable scores, we suggest that prolonged engagement with video-assessment would yield better outcomes.

Strengths and limitations

This study evaluated what physical therapists do in their day-to-day practice by assessing client records, video-recordings and management information. The quality improvement program was systematically developed, and theory-based and evidence-based. Stakeholders and end-users were actively involved in program development and implementation and their experiences provided meaningful information on its critical success features. Participants did not fully adhere to the program guidelines resulting in limited sample sizes threatening internal and external validity of the quantitative results. It should also be noted that we could not distinguish between 'missing' and 'not relevant or not applicable' indicator scores of active PA participants which might have biased the results. Although generalizability of the quantitative results is limited regarding the specific population of Dutch physical therapists in primary care, we think that the qualitative results related to the acceptability and the implementation of the quality improvement program are learning points for a broader group of healthcare professionals.

Conclusions

This study demonstrated that bottom-up quality improvement initiatives can be effective in improving healthcare quality. The results justify more rigorous evaluation to inform nationwide implementation when its critical success features are addressed. Crucial is the willingness of professionals and organizations to provide access to the confidential areas of their clinical practice. However, this information is vulnerable to summative judgment and should be protected by all stakeholders in healthcare quality. Further research is necessary to explore the sustainability of the results and the impact on client outcomes in a full scale study.

Abbreviations

MM = Marjo Maas

FD = Femke Driehuis

RN = Ria Nijhuis-van der Sanden

PW = Philip van der Wees

References

1 Institute of Medicine. *Crossing the quality chasm: a new health system for the 21st century.* Washington, DC: National Academy Press 2001.

2 Eijkenaar F. Pay-for-performance for healthcare providers. Design, performance measurement, and (unintended) effects. *Health Policy (New York)* 2013;110:115–30.

3 Casalino LP, Elster A, Eisenberg A, *et al*. Will pay-for-performance and quality reporting affect health care disparities? *Health Aff* 2007;26.

4 Grol R. Quality improvement by peer review in primary care: a practical guide. *Qual Health Care* 1994;3:147–52.

5 Butterfield R, McCormick B, Anderson R, *et al*. Quality of NHS care and external pathway peer review. 2012. Available at: https://www.chseo.org.uk/downloads/report3-peerreview.pdf. (accessed October 2016)

6 Pronovost PJ, Hudson DW. Improving healthcare quality through organisational peer-to-peer assessment: lessons from the nuclear power industry. *BMJ Qual Saf* 2012;21:872–5.

7 Prochaska JO, Redding CA, Evers KE. Health behavior and health education. In: Glanz K, Rimer B k, Viswanath K, eds. *Health behavior and health education: theory, research, and practice.* Wiley & Sons 2008. 97–121.

8 Ajzen I. Nature and operation of attitudes. *Annu Rev Psychol* 2001;52:27–58.

9 Maas MJM, van Dulmen SA, Sagasser MH, *et al*. Critical features of peer assessment of clinical performance to enhance adherence to a low back pain guideline for physical therapists: a mixed methods design. *BMC Med Educ* 2015;15:203.

10 Ferlie EB, Shortell SM. Improving the quality of health care in the United Kingdom and the United States: a framework for change. *Milbank Q* 2001;79:281–315.

11 Australian commission on safety and quality in health care. Review by peers. A guide for professional, clinical and administrative processes. 2010. Available at: https://www.safetyandquality.gov.au/wp-content/uploads/2012/01/37358-Review-by-Peers1.pdf. (accessed April 2016)

12 Buetow SA, Roland M. Clinical governance: bridging the gap between managerial and clinical approaches to quality of care. *Qual Heal Care* 1999;8:184–90.

13 Eva KW, Regehr G. Self-assessment in the health professions: a reformulation and research agenda. *Acad Med* 2005;80:S46-54.

14 Norcini JJ. Peer assessment of competence. *Med Educ* 2003;37:539–43.

15 Hofhuis H, van den Ende C, de Bakker D. Effects of visitation among allied health professionals. *Int J Qual Heal Care* 2006;6:397–402.

16 Mann K, van der Vleuten CP, Eva KW, *et al*. Tensions in informed self-assessment: how the desire for feedback and reticence to collect and use it can conflict. *Acad Med* 2011;86:1120–7.

17 Maas MJM, van der Wees PJ, Braam C, *et al*. An innovative peer assessment approach to enhance guideline adherence in physical therapy: a single-masked, cluster-randomized controlled trial. *Phys Ther* 2015;95:600–12.

18 Phillips Fox D. Peer review of health care professionals: a systematic review of the literature. Melbourne: 2009. Available at: https://www.safetyandquality.gov.au/wp-content/uploads/2012/01/25738-LitReview.pdf. (accessed April 2016)

19 Van den Hombergh P, Grol R, van den Hoogen HJ, *et al*. Practice visits as a tool in quality improvement: acceptance and feasibility. *Qual Health Care* 1999;8:167–71.

20 Ivers N, Jamtvedt G, Flottorp S, *et al*. Audit and feedback: effects on professional practice and health care outcomes (Review). *Cochrane Database Syst Rev* 2012;1–227.

21 Payne VL, Hysong SJ. Model depicting aspects of audit and feedback that impact physicians' acceptance of clinical performance feedback. *BMC Health Serv Res* 2016;16.

22 Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. *Med Care* 2009;47:356–63.

23 Moore GF, Aurdrey S, Barker M, *et al*. Process evaluation of complex interventions: Medical Research Council guidance. *Br Med J* 2015;350.
24 Bowen DJ, Kreuter M, Spring B, *et al*. How we design feasibility studies. *Am J Prev Med* 2010;36:452–7. doi:10.1016/j.amepre.2009.02.002.
25 Hoens A, Li L. The knowledge broker's 'fit'in the world of knowledge translation. *Physiother Canada* 2014;66:223–7.
26 Ward V, House A, Hamer S. Knowledge Brokering : the missing link in the evidence to action chain? *Evid Policy* 2009;5.
27 Li LC, Grimshaw JM, Nielsen C, *et al*. Use of communities of practice in business and health care sectors: a systematic review. *Implement Sci* 2009;4:27.
28 Eva KW, Armson H, Holmboe E, *et al*. Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ Theory Pract* 2012;17:15–26.
29 Archer JC. State of the science in health professional education: effective feedback. *Med Educ* 2010;44:101–8.
30 Sargeant JM, Mann K V, van der Vleuten CP, *et al*. Reflection: a link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract* 2009;14:399–410.
31 Kluger AN, Denisi A. The effects of feedback interventions on performance : a historical review, a meta-analysis , and a preliminary feedback intervention theory. *Pscyological Bull* 1996;119:254–84.
32 Tong A, Sainsbury P, Craig J. Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus group. *Int J Qual Heal Care* 2007;19:349–57.
33 King N, Cassel C, Symon G. Using templates in the thematic analysis of texts. In: *Essential guide to qualitative methods in organisational research*. London: Sage; 2004. 256–70.
34 Huberman AM, Miles MB, Denzin NK, *et al*. Data management and analysis methods. In: Handbook of Qualitative Research. *Sage Publications* 1994. 428–44.
35 Hobma S, Ram P, Muijtjens A, *et al*. Effective improvement of doctor-patient communication: A randomised controlled trial. *Br J Gen Pract* 2006;56:580–6.
36 Overheem K, Faber MJ, Onyebuchi AA, *et al*. Doctor performance assessment development in daily practise: does it help doctors or not? A systematic review. *Med Educ* 2007;41:1039–49.
37 van der Vleuten CPM, Schuwirth LWT, Driessen EW, *et al*. Twelve tips for programmatic assessment. *Med Teach* 2015;37:641–6.
38 Weiner B. A theory of organisational readiness for change. *Implement Sci* 2009;4:67.
39 Eva KW, Cunnington JPW, Reiter HI, *et al*. How can I know what I don't know? Poor self assessment in a well-defined domain. *Adv Health Sci Educ Theory Pract* 2004;9:211–24.
40 Bandura A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol Rev* 1977;84:191–215.
41 Hysong SJ, Kell, Harrison J, Petersen LA, Campbell BA, *et al*. Theory-based and evidence-based design of audit and feedback programms: examples from two clinical intervention studies. *BMJ Qual Saf* 2016;bmjqs-2015.
42 Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ* 2005;39:98–106.
43 Durning S, Artino AR, Pangaro L, *et al*. Context and clinical reasoning: Understanding the perspective of the expert's voice. *Med Educ 2011*;45:927–38.
44 Baig LA, Violato C, Crutcher RA. Assessing clinical communication skills in physicians: are the skills context specific or generalizable. *BMC Med Educ* 2009;9:22.

Appendix

Development of the quality improvement program

In the developmental phase of the program we applied participatory action research principles to facilitate relationship building and cross-sharing of knowledge and best practices between program developers and end-users.[1] The program was developed from September 2014 to January 2015 by the research group (n=4) in collaboration with experts in QI research (n=4) and a stakeholders group (n=9) containing four members of the Royal Dutch Society for Physical Therapy (KNGF), one professional auditor, and four physical therapists working in primary care including two leaders of networks of end-users. Six meetings were scheduled with the stakeholder group; experts were consulted more frequently.

We conducted the following steps to build the program: 1) exploring the existing literature on PA of clinical and organizational performance, including the measurement instruments being used, 2) identifying existing theory on QI, 3) designing the program including the development of performance indicators and testing procedures, 4) developing software to support the PA and clinic visitation process, and 5) developing of a program guide.

Exploring the existing literature

A scoping review of the scientific literature and grey literature was conducted to identify relevant PA practices focusing on competency assessment in primary care. Relying on prior research of our research group on the effectiveness of PA[2,3] and supported by literature on effective QI interventions[4-6], the QI was developed.

Identifying existing theory

We mapped relevant theories, the underlying constructs and explicated how we operationalized these constructs in our program design. Box 1 shows an overview.

Designing the program including the development of performance indicators and testing procedures

The development of performance indicators was informed by the Dutch professional profile for physical therapists[18] which includes a set of competency domains and corresponding global performance indicators according to the CanMEDS physician competency framework,[19] and in line with the quality indicators of the Institute of Medicine.[20] We selected three domains of professional performance, strongly related to client-centeredness, effectiveness, and transparency of physical therapy care: 1) record keeping, 2) client

Box 1 — Theories used to build and to implement the QI program

Theory	Underlying constructs used	Operationalization of constructs
Social constructivist learning theory[7]	Contextual learning, collaborative learning, active participation, and knowledge construction to enhance attention, storage, and retrieval of knowledge from memory.	Presenting authentic clinical problems (client records and video-recordings) to approach clinical practice as much as possible. Enhancing active participation by using a performance based QI strategy. Using face-to-face discussion to share knowledge and deepen understanding.
Self-regulated learning theory[8,9]	Applying meta-cognitive strategies to enhance readiness for change and to guide the professional development process.	Conscious goal setting based on self-assessment and peer assessment results.
Situated learning theory[10,11]	Learning in the context of daily practice to bridge the gap between learning context and application context.	Delivering the program within communities of practice that share the same setting or the same interest.
Social cognitive learning theory[12]	Enhancing the development of self-efficacy beliefs by performing the new behavior and experiencing the consequences of that behavior (mastery experience).	Exposing professional behaviors for critical appraisal.
	Enhancing the development of self-efficacy beliefs by observing peer behavior and the consequences of that behavior (vicarious experience).	Observing a peer's performance.
Feedback Intervention Theories addressing performance improvement in health care[13-15]	Providing 'feedback' (knowledge of results) and 'feed forward' (guidance for self-regulated improvement) based on standards of performance.	
Theory of planned behavior[16]	Changing attitudes and subjective norms toward the new behavior and enhancing the development of self-efficacy beliefs.	Introducing peers to the assessor perspective. In appraising a peer's performance, assessors need to develop an understanding and a mutually accepted quality standard to deliver credible performance feedback.
Diffusions of innovations theory[17]	Aligning the QI program to the context of end-users.	Using knowledge brokers to bridge the gap between program developers and program users.

communication, and 3) organization and management of private physical therapy clinics. Informed by the professional profile of the physical therapists[18] we developed global performance indicators for client communication, record keeping, and clinic organization and management. The performance indicators were compared to existing indicators and scoring criteria collected by the scoping review, audit criteria of health insurers in the Netherlands, KNGF guidelines for record keeping, KNGF toolkits for communication , physical setting and equipment, privacy and safety. It should be noted that evidence based practice is the cornerstone of clinical reasoning and decision making in physical therapy practice. The KNGF guideline for Record Keeping states that professionals who do not adhere to clinical practice guidelines if applicable, should motivate that in their client records. The guideline also states that 'testing for the use of clinical practice guidelines is addressed in peer review and clinical audit'.
A series of consensus meetings with experts were organized to discuss the three sets of global performance indicators and scoring criteria. Finally the stakeholder group approved all the quality indicators.

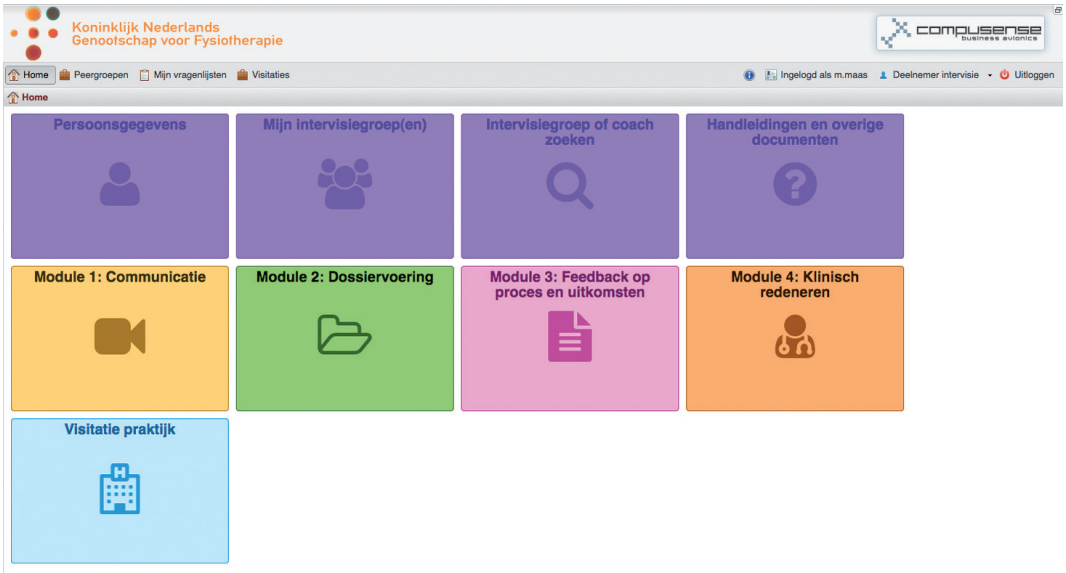
Developing software to support the PA and clinic visitation process

Together with te software company Compusense Business Avionics B.V.[21] we developed a web-based performance assessment system. Figure 1 shows a screen shot of the introduction page. The icons give access to different functions.

Developing a program guide

The program guide included a description of the background of the QI program, its aims and procedures, and information on participation requirements regarding accreditation. Each step of the PA and visitation process was elaborated to allow for sufficient preparation including all performance indicators and references to relevant documents. We provided evidence based guidelines for providing constructive performance feedback and for enhancing feedback acceptance including tips for responding to feedback.[13,15,22-24]

Figure 1 — Web-based performance assessment system (screenshot introduction page)



Self-assessment and peer assessment form client communication

Instruction
1–5: 1 = much improvement needed; 5 = no improvement needed
When improvement is needed, please provide written feedback and tips for improvement.
N = not relevant/not applicable

Performance indicators and corresponding quality domains	1	2	3	4	5	N	Feedback and tips
1[a] Is the help request clarified?	O	O	O	O	O	O	
2[a] Are the findings of the intake and clinical examination clearly communicated in understandable, client-friendly language?	O	O	O	O	O	O	
3[a,b,c] Are the patient reported outcomes and performance outcomes used to develop a treatment plan in dialogue with the client?	O	O	O	O	O	O	
4[a,b,c] Are the outcome expectancies of therapist and client aligned?	O	O	O	O	O	O	
5[a,b,c] Are the outcome expectancies formulated SMART (specific, measurable, acceptable, realistic, time contingent)?	O	O	O	O	O	O	
6[a] Are the interventions clearly communicated in dialogue with the client?	O	O	O	O	O	O	

[1] quality domain 'client-centeredness'
[2] quality domain 'effectiveness' including evidence based practice.
[3] quality domain 'transparency'

Additional remarks

Self-assessment and peer assessment form record keeping

Instruction
1–5: 1 = much improvement needed; 5 = no improvement needed
When improvement is needed, please provide written feedback and tips for improvement.
N = not relevant/not applicable

Performance indicators and corresponding quality domains			1	2	3	4	5	N	Feedback and tips
1[3]	Readability	Is the record written in plain language and is reporting concise?	O	O	O	O	O	O	
2[3]	Completeness	Does the record adhere to the KNGF guideline for record keeping 2016?	O	O	O	O	O	O	
3[3]	Transparency	Is the process of clinical reasoning and decision making transparent?	O	O	O	O	O	O	
4[1,2,3]	Consistency	Are the different steps in the process of diagnosis, treatment, and evaluation consistent with each other (are there no contradictory steps)?	O	O	O	O	O	O	
5[1,2,3]	Client reported outcome measures (questionnaires)	Is the use of client reported outcome measures (if relevant) adequate?	O	O	O	O	O	O	
6[1,2,3]	Performance measures (clinical tests)	Is de use of performance measures (if relevant) adequate?	O	O	O	O	O	O	

[1] quality domain 'client-centeredness'
[2] quality domain 'effectiveness' including evidence based practice
[3] quality domain 'transparency'

Additional remarks

Instruction
1–5: 1 = much improvement needed; 5 = no improvement needed
When improvement is needed, please provide written feedback and tips for improvement.
N = not relevant/not applicable

Performance indicators and corresponding quality domains		1	2	3	4	5	N	Feedback and tips
Indicator 1[1,2,3] Quality management	The practice conducts an active policy focusing on continuous improvement and accountability for the quality of the organization, staff, care and services. **How to demonstrate** *Quality policy* – Statement of mission, purposes, goals, and procedures. – Written plan for continuous improvement of quality of care and performance of services. – Written annual quality report. *Staff development* – Written plan that provides for appropriate and ongoing staff development. – BIG Registry[25] KNGF quality register.[26] – Active participation in peer assessment activities. *Quality of care and services* – Client reported outcomes (PROMS). – Client reported experiences (PREMS). – Complaints procedures for clients. – Treatment averages reported by condition.	O	O	O	O	O	O	
Indicator 2[2,3] Client management	Client records adhere to the KNGF guideline record keeping 2016[27] and demonstrate. transparency in the process of clinical reasoning and decision making according to the KNGF professional profile of the physical therapist.[18] **How to demonstrate?** – Evidence of participation in peer assessment activities. – Improvement plans. – Assessment of a random sample of client records.	O	O	O	O	O	O	
Indicator 3[3] Communication and collaboration	The practice has an appropriate system of internal communication and collaboration and complies with the NGF/KNGF guidelines for information exchange.[28] **How to demonstrate?** *Practice meetings* – Records and notes *Inter professional meetings* – Records and notes – Reports to referring physicians	O	O	O	O	O	O	

Continuation appendix — see next page

Continuation appendix 'Audit form organization and management'

Performance indicators and corresponding quality domains		1	2	3	4	5	N	Feedback and tips
Indicator 4[1] Physical setting	**Criterium** The physical setting is designed to provide a safe and accessible environment. The equipment is safe and appropriate to achieve the purposes and goals of physical therapy. **How to demonstrate?** – Quality policy – KNGF Toolkit 'client information' – KNGF Toolkit 'clinic design requirements' – KNGF Toolkit 'patient safety'	O	O	O	O	O	O	
Privacy and safety	The clinic conducts an active policy to safeguard privacy and safety. Physical therapists comply with Professional standards of ethical conduct. **How to demonstrate?** – Quality policy – KNGF Toolkit 'client information' – KNGF Toolkit 'clinic design requirements' – KNGF Toolkit 'patient safety' – KNGF Toolkit 'security of client information'	O	O	O	O	O	O	
Innovation and entrepreneurship	The clinic develops and implements innovations to respond to societal changes and to connect to new developments in healthcare. **How to demonstrate?** The clinic determines how innovation and entrepreneurship is demonstrated.	O	O	O	O	O	O	

[1]quality domain 'client-centeredness'
[2]quality domain 'effectiveness' including evidence based practice.
[3] quality domain 'transparency'

Additional remarks

Appendix references

1 Minkler M, Wallerstein N. *Community-Based Participatory Research for Health: From Process to Outcomes*. 2nd ed. (Minkler M, Wallerstein N, eds.). San Fransico: Jossey-Bass Inc; 2008.

2 Maas MJM, van Dulmen SA, Sagasser MH, *et al*. Critical features of peer assessment of clinical performance to enhance adherence to a low back pain guideline for physical therapists: a mixed methods design. *BMC Med Educ.* 2015;15(1):203.

3 van Dulmen SA, Maas MJ, Staal B, *et al.* Effectiveness of peer-assessment for implementing a Dutch physical therapy low back pain guideline: a cluster randomized controlled trial. *Phys Ther.* 2014;94(10):1396-1409.

4 Grimshaw JM, Eccles MP, Lavis JN, Hill SJ, Squires JE. Knowledge translation of research findings. *Implement Sci.* 2012;7(1):50.

5 Ivers N, Jamtvedt G, Flottorp S, *et al.* Audit and feedback: effects on professional practice and health care outcomes (Review). *Cochrane Database Syst Rev*. 2012;(7):1-227.

6 Colquhoun HL, Brehaut JC, Sales A, *et al*. A systematic review of the use of theory in randomized controlled trials of audit and feedback. *Implement Sci.* 2013;8(1):66. doi:10.1186/1748-5908-8-66.

7 Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Med Educ.* 2006;40(7):618-623.

8 Schön D. The Reflective Practitioner: *How Professionals Think in Action*. San Francisco: Jossey-Bass Inc; 1983.

9 Greene J, Azevedo R. A theoretical review of Winne and Hadwin's model of self-regulated learning: new perspectives and directions. Rev Educ Res. 2007;77(3):334-372. doi:10.3102/003465430303953.

10 Wenger E. Communities of practice and social learning systems. *Organization*. 2000;7(2):225-246. doi:10.1177/135050840072002.

11 Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Use of communities of practice in business and health care sectors: a systematic review. *Implement Sci*. 2009;4:27. doi:10.1186/1748-5908-4-27.

12 Bandura A. Self-efficacy: toward a unifying theory of behavioral change. *Psychol Rev.* 1977;84(2):191-215. doi:10.1037/0033-295X.84.2.191.

13 Kluger AN, Denisi A. The effects of feedback interventions on performance : a historical review, a meta-analysis , and a preliminary feedback intervention theory. *Pscyological Bull.* 1996;119(2):254-284.

14 Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. *Med Care.* 2009;47(3):356-363. doi:10.1097/MLR.0b013e3181893f6b.

15 Hysong SJ, Kell HJ, Petersen LA, Campbell BA, Trautner BW. Theory-based and evidence-based design of audit and feedback programmes: examples from two clinical intervention studies. *BMJ Qual Saf.* 2016;0(6):1-12.

16 Ajzen I. Nature and operation of attitudes. *Annu Rev Psychol.* 2001;52:27-58.

17 Ward V, House A, Hamer S. Knowledge Brokering : the missing link in the evidence to action chain? *Evid Policy.* 2009;5(3).

18 de Vries C, Hagenaars L, Kiers H, Schmitt M. KNGF Beroepsprofiel fysiotherapeut 2014. https://www.kngf.nl/binaries/content/assets/kngf/onbeveiligd/vakgebied/vakinhoud/beroepsprofielen/2014-01_kngf_beroepsprofiel-ft_20131230_2.pdf. Accessed October 1, 2016.

19 Frank J, Jabbour M, Tugwell P, Boyd D, Fréchette D, Labrosse J. The CanMEDS 2005 Physician Competency Framework. http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/the_7_canmeds_roles_e.pdf. Published 2005.

20 Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press; 2001

21 Compusense Business Avionics. http://www.compusense.nl/.

22 Sargeant J, Eva KW, Armson H, *et al.* Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ.* 2011;45(6):636-647.
23 Eva KW, Armson H, Holmboe E, *et al.* Factors influencing responsiveness to feedback: on the interplay between fear, confidence, and reasoning processes. *Adv Health Sci Educ Theory Pract.* 2012;17:15-26.
24 Sargeant JM, Mann K V, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract.* 2009;14(3):399-410.
25 Ministerie van Volksgezondheid en Welzijn BIG register. https://www.bigregister.nl/. Accessed February 27, 2016.
26 KNGF kwaliteitsregister. https://www.kngf.nl/vakgebied/kwaliteit/ckr.html. Accessed October 1, 2016.
27 KNGF-richtlijn Fysiotherapeutische dossiervoering 2016. https://www.fysionet-evidencebased.nl/index.php/richtlijnen/richtlijnen/fysiotherapeutische-dossiervoering/praktijkrichtlijn/a-aan-het-dossier-toe-te-voegen-gegevens-per-fase-van-het-methodisch-handelen/a-5-dossiergegevens-behandelplan-fase-5. Published 2016.
28 NHG-KNGF-richtlijn gestructureerde informatie- uitwisseling tussen huisarts en fysiotherapeut. https://www.nhg.org/themas/artikelen/richtlijn-informatie-uitwisseling-huisarts-en-fysiotherapeut Accessed October 1, 2016.

Chapter 7
The impact of self-assessment and peer assessment on clinical performance of physical therapists in primary care: a cohort study

Marjo Maas
Femke Driehuis
Guus Meerhoff
Yvonne Heerkens
Cees van der Vleuten
Ria Nijhuis-van der Sanden
Philip van der Wees

Accepted for publication
Physiotherapy Canada

Abstract

Aim

To evaluate the impact of a quality improvement program based on self- and peer assessment to justify nation-wide implementation.

Subjects and setting

Four professional networks of physical therapists in The Netherlands (n=379).

Methods

The program comprised two cycles of online self-assessment and peer assessment using video-recordings of client communication and clinical records. Assessment was based on performance indicators that could be scored on a 5-point Likert scale. Online assessment was followed by face-to-face feedback discussions. After cycle 1, participants developed personal learning goals.
Personal goals were analysed thematically. Goal attainment was measured with a questionnaire. Performance improvement was tested with multilevel regression analyses, comparing self-assessment and peer assessment scores in cycle 1 and 2.

Results

In total 364 (94%) participants were active in online self-assessment and peer assessment. However, online activities varied between cycle 1 and 2, and between client communication and record keeping. Personal goals addressed: client-centered communication (54%), record keeping (24%), performance- and outcome measurement (15%), other (7%). Goals were completely attained (29%), partly (64%), and not (7%). Self-assessment and peer assessment scores improved significantly for both client communication (self-assessment=11%; peer assessment=8%) and record keeping (self-assessment=7%; peer assessment=4%).

Conclusions

Self-assessment and peer assessment are effective in enhancing commitment to change and improving clinical performance. Nation-wide implementation is justified. Future studies should address the impact on client outcomes.

Introduction

People seeking the help of a physical therapist deserve the best possible care provided by up-to-date educated professionals who can take responsibility for the quality of their services. In the Netherlands, the quality of physical therapy services is defined by four quality domains: effectiveness, patient-centeredness, transparency and safety according to the quality framework of the Institute of Medicine (IOM).[1] Several Dutch authorities determine whether the best possible care is delivered, representing different interests and focusing on different quality domains. The Dutch government regulates quality by maintaining a national register based on certification,[2] and the health insurers by sampling and benchmarking process and outcome data, and conducting client satisfaction surveys and clinical audits. The literature shows that external regulations can potentially be effective, but the results are short-term and the strategy might induce unwanted consequences such as under-treatment of clients with multi-morbidity or disparities in healthcare delivery.[1,3] Professionals often resist external regulations by health insurers because these might challenge their professional identity and autonomy as explained by self-determination theory.[4] A study of Scholte *et al.* on the impact of a Dutch performance feedback system for physical therapists, based on indicator scores extracted directly from electronic health records, showed that financial incentives by health insurers negatively affected the use of feedback reports for quality improvement (QI). A lack of 'belief' in the QI system and 'distrust' among physical therapists towards health insurers were the major barriers to feedback use.[5]

If physical therapists themselves take responsibility for the quality of their services, they have the opportunity to design the QI interventions and outcome measures in such a way that they are in line with their professional norms and values, cover the complexity of their professional roles, and adequately reflect their day-to-day practice. It gives them the possibility to look forward in anticipating change rather than to look back.[6] Looking forward might address the changing roles of physical therapists in providing client-centered care which requires sophisticated client communication skills and advanced collaboration in healthcare teams. It also implies adequate process and outcome measurement including record keeping, to meet the increasing demand for transparency, accountability, and access to information.[7]

Research has shown that self-regulation might be more effective on the long-term than external regulations[8,9] because mutually

shared social and professional standards of performance are critical to professional behavior change.[10-12] In contrast to external regulations, self-regulation allows for providing feedback to raise awareness of actual performance and feedforward to anticipate desired future performance.[13]

Conditional to successful self-regulation is that professionals develop shared quality standards of their services, and the willingness and ability to critically appraise their own and their colleagues' performance.[8,9,14,15] Ideally, self-regulation systems should address all competency domains and professional roles. This implies that self-regulation should not be limited to individual healthcare professionals, but also involve teams and provider organizations as captured in the concept of clinical governance.[16] Supported by the Royal Dutch Society for Physical Therapy we developed a QI program based on self- and peer assessment as an integral part of a comprehensive national quality assurance system: "Quality In Motion" which includes clinical audit and benchmarking patient reported experiences (PREMs) and outcomes (PROMs).[17]

In self-assessment, professionals reflect on their own professional practice according to pre-defined quality indicators.[18] Research on the validity and reliability of self-assessment showed that clinicians are poor self-assessors and that information of external resources – such as peer feedback – is needed to build up an adequate self-concept.[19-21]

In peer assessment, participants critically appraise their peers' performance and provide each other with constructive feedback, allowing for strengthening adequate performance and early spotting poor performance. Peer assessment aims to develop a critical attitude towards clinical performance by introducing professionals with an 'assessor' or 'auditor' perspective.[14] Two randomized controlled trials comparing the effectiveness of 'peer assessment' with 'group discussions' as a strategy for guideline implementation, showed that peer assessment was more effective and associated with significantly higher levels of self-awareness than group discussions.[22,23] An evaluative study on the critical success features of this implementation strategy demonstrated the strength of a performance-based program, triggering both cognitive, emotional, and social commitment to the assessment task.[24] Regarding effective feedback for improvement, a meta review of Ivers *et al*.[25] indicated that "feedback may be more effective when the source is a supervisor or colleague, when it is provided more than once, when it is delivered in both verbal and written formats, and when it includes both explicit targets and an action plan".[25] Research on the accept-

ance and use of performance feedback showed that the credibility of the feedback source impacts on feedback acceptance.[26,27] In short, feedback aiming at QI should be provided by a 'trustworthy' and 'credible' source.

We developed a theory- and evidence-based QI program as advocated by the literature, aiming to enhance and ultimately self-regulate the quality of physical therapy care.[28,29] The program was developed, implemented and evaluated informed by the 4-stage framework of the Medical Research Council which involves: 1) program development, 2) assessing feasibility, 3) assessing impact on performance outcomes, 4) implementation.[30] The results on program development and feasibility testing are reported in a previous study.[31] This study reports of the third step, prior to implementation. Our research questions address the impact of the QI program on commitments to professional behavior change and performance improvement.

Objectives

To evaluate the impact of a QI program based on self- and peer assessment to improve the client-centeredness, effectiveness, and transparency of physical therapy care, and to justify nationwide implementation.

Methods

Design

The impact of the QI program was evaluated in a non-controlled cohort study with a before-after design using mixed methods.[32]

Subjects and setting

Participants were physical therapists working in primary care organized in four existing professional networks in the Netherlands. In primary care, physical therapists work in mono-disciplinary or multi-disciplinary private clinics. They are accessible with or without the referral of a physician. The participants were invited by the Royal Dutch Society for Physical Therapy (KNGF) via a digital newsletter. Participation was voluntary and awarded with continuing education credits for the quality register.[33] We used knowledge brokers[34] as the linking pin between researchers and participants to enhance program implementation and trained coaches to support feedback acceptance and use in face-to-face discussions.[26]

Knowledge brokers were physical therapists who were board members of the participating professional networks and coaches were members of these networks, trained to perform this role.

Program development

We selected two 'competency domains' for performance assessment, both closely related to the IOM[1] quality domains client-centeredness, effectiveness, and transparency: 1) client communication and 2) record keeping. For each quality domain, six performance indicators were developed, guided by the Dutch professional competency profile[35] which was developed according to the CanMeds physician competency framework.[36] Indicators aimed to support self- and peer assessors in the process of providing improvement feedback and to guide the process towards its intended outcomes.[18] Appendix 1a-b presents the performance indicators and their relationship with the distinct quality domains. We developed a web-based assessment system that allowed participants to a) upload assessment materials such as client records, video-recordings and improvement plans, b) score themselves and their peers, and c) download the assessment results in the form of narrative peer feedback and scores. The website allowed the researchers for a) upload program guides and instruction manuals for participants, b) monitor progress of participants, and c) store and export qualitative and quantitative data. Access to information was limited and regulated according to the participant's role. The software company Compusense Business Avionics[37] adhered to all the regulations relating to privacy of personal data for both clients and participants.

Program content

The program contained two cycles of online self- and peer assessment followed by face-to-face peer group discussions. Participants were provided with a program guide describing the program aims, content, procedures, intended results, and expected investment of time and effort. Included was a manual for uploading and downloading assessment material and guidelines for providing and using feedback informed by the best available evidence.[25,26,28,38-41] An introduction meeting was scheduled addressing all program issues including any perceived barriers to participation.

In cycle 1, participants uploaded a self-recorded video of a client-interview and a corresponding client record. Participants were assigned to limit the video-recording to the summarized discussion of the diagnosis and treatment plan. Self- and peer assessment comprised studying the uploaded materials, scoring performance

indicators on a 5-point Likert scale ranging from 1=much improvement needed to 5=no improvement needed (see appendix 1), and providing written improvement feedback. Participants had to first self-assess their performance before they were given access to assess their peers to avoid self-assessment bias. Discrepancies between indicator scores were used as input for the subsequent face-to-face discussions. Coaches had special access to the results of the peer groups they were coaching to monitor their online activities and prepare the face-to-face group discussions. After the first cycle, participants were encouraged to reflect on their peer feedback and to formulate personal goals according to the concept of 'Commitments to Change'.[42]
In cycle 2 (4 - 6 weeks later), the process of self-assessment and peer assessment was repeated and personal goals were evaluated. Appendix 2 shows the planned program activities and its intended impact on clinical performance.

Program delivery

Peer group coaches were trained in three sessions by members of the research team who were physical therapists and experienced trainers (MM, PW, GM, FD). They used samples of client records, video-recordings and role-play to train the process of providing, receiving, and using performance feedback. Participants received a program guide tailored to their role in the assessment process. The program was managed by two research team members (FD and AB) functioning as linking pins between the research team and others involved to allow for early spotting and solving implementation problems.

Outcome measures

The impact on commitments to change was explored by thematic analysis of the content of personal goals after cycle 1 and exploring the attainment of personal goals with an online questionnaire after cycle 2.[42] Participants were asked to indicate on a 3-point Likert scale to what extent their personal goals were achieved (1=not achieved; 2=partly achieved, 3=completely achieved).
The impact on performance improvement was assessed by comparing indicator scores for each competency domain and for each corresponding performance indicator between cycle 1 and 2.

Data sampling and analysis

We took a descriptive approach to the content analysis of personal goals. MM, FD, and AB independently studied and coded a sample of personal goals (formulated after cycle 1) of each physical therapist. The analysis was guided by 'template analysis' that combines a-priori

codes (client communication and record keeping) and emerging codes from the data.[43] We discussed differences in coding and created a code book based on consensus. Subsequently, all personal goals were coded using Microsoft Excel 2013 software. Codes were compared and some codes were merged into higher-order codes. Data were reduced by constant comparison of codes allowing us to identify themes and subthemes in personal goals. The frequency with which these themes and subthemes were mentioned has been counted and described as a percentage of the total of personal goals to identify participants' major learning needs. Questionnaire data on the attainment of personal goals as reported after cycle 2 were entered in IBM SPSS Statistics 22 and the results were described.
Online scores for record keeping and client communication in the first and second assessment cycle were imported in IBM-SPSS Statistics 22 and statistically analyzed. We calculated and described the proportion of 'not relevant/not applicable' scores, and treated them as missing values in the analyses. Mean indicator scores for each competency domain were calculated for self-assessment and peer assessment. 'Difference indicator scores' were calculated by subtracting mean cycle 1 scores from cycle 2 scores. Because of the hierarchical structure of our study (subjects nested within peer groups) we performed a multilevel (mixed model) analysis.[44] We used a random intercept model with the 'difference indicator score' as outcome variable and the peer group as a random factor. Difference indicator scores were estimated as mean difference with 95% confidence intervals. We described the differences as median improvement percentages. A standard of 5% improvement was determined based on the results of a meta review on effects of audit and feedback on professional practice by Ivers *et al*.[25]

Results

Four networks of physical therapists participated in the program (n= 379). We trained 38 coaches to support 73 peer groups. Table 1 shows participants' characteristics.

Impact on commitment to change

In total 303 participants uploaded their personal goals after cycle 1 (80%). We analyzed the content of personal goals of all participants and identified three major themes and 16 subthemes. The themes and subthemes demonstrated great similarity to the performance indicators for client communication and record keeping (Appendix

Table 1 — Participants' characteristics

Network		1	2	3	4	Total
Participants		68	87	148	76	379
Online active		65 (95%)	85 (98%)	143 (97%)	71 (93%)	364 (96%)
Peer groups		14	17	30	12	73
Coaches		13	8	11	6	38
Mean age (SD[1])		39.25 (10.88)	41.46 (12.23)	44.35 (12.27)	48.40 (12.34)	43.66 (12.40)
Mean experience (SD)		15.21 (10.25)	16.49 (12.03)	20.39 (11.8)	25.93 (11.8)	19.74 (12.14)
Gender	Male	24 (35%)	27 (31%)	60 (42%)	27 (38%)	140 (38%)
Specialisation	Generalist	38 (46%)	34 (39%)	81 (55%)	35 (52%)	197 (52%)
	Specialist	30 (44%)	53 (61%)	67 (45%)	32 (48%)	182 (48%)

[1]SD: Standard Deviation

1a-b). Table 2 shows an overview of all themes and subthemes including the frequencies in which these themes and subthemes have been mentioned by the participants.
The major learning needs emerging after cycle 1 addressed 1) client–centered communication including shared decision making (54%), followed by 2) record keeping including measurable goal setting (24%), and 3) performance- and patient reported outcome measurement (15%), and other themes (7%). The results on goal attainment show that 29% of the participants completely, 64% partly, and 7% did not attain their personal goals.

Impact on clinical performance

In total 364 (94%) participants were active in online self-assessment and peer assessment. However, online activities varied between cycle 1 and 2, and between client communication and record keeping. The mean impact of self-assessment and peer assessment on the improvement of client communication and record keeping is only calculated for participants who were active in both cycles as presented in table 3. The results show that self-assessment scores for both client communication and record keeping were consistently lower than peer assessment scores, but the differences became smaller in cycle 2. Self-assessment and peer assessment scores significantly improved in cycle 2 for both communication and record keeping, although self-assessment results improved more than peer assessment results. Table 3 shows that the median percentage change was higher than 5% except for peer assessed record keeping. Self-assessed communication 11%, peer assessed

Table 2 — Self-reported goal attainment, themes and subthemes of personal goals

General information		Number	%
Participants who uploaded personal goals		303	80
Participants who completed the questionnaire		242	64
Mean number of personal goals per physiotherapist (SD)	3,93 (0.97)		
Self-reported goal attainment		**Number**	**%**
Not realized		18	7
Partly realized		154	64
Completely realized		70	29
Themes	**Subthemes**	**Number**	**%**
Client-centered communication and shared decision making.	Clarify request for help.	74	7
	Allow for more dialogue.	40	4
	Convey clear en concise information, avoid technical terms.	94	9
	Structure and summarize information and verify if information is heard and understood.	55	5
	Pay more attention to non-verbal behaviours.	14	1
	Involve client in goal setting and treatment planning.	68	7
	Discuss PROMs results and use them as an aid to set and evaluate measurable goals.	95	9
	Clearly communicate prognoses, align mutual expectancies and share responsibilities.	110	11
	Subtotal on communication	**550**	**54**
Record keeping	Improve conciseness	24	2
	Improve completeness	60	6
	Improve SMART[1] goal setting aligned with the request for help.	92	9
	Familiarize with software programme.	16	2
	Improve transparency in clinical reasoning.	49	5
	Subtotal on record keeping	**241**	**24**
Performance and client reported outcome measurement.	Select and apply appropriate performance – and outcome measures.	72	7
	Improve regular monitoring with performance and outcome measures.	79	8
	Subtotal on performance and outcome measurement.	**150**	**15**
Other	Guideline adherence / Video-recording / Training protocols / Critical performance appraisal.	78	7
Total		**1019**	**100%**

[1] SMART = specific, measurable, acceptable, realistic, time contingent

communication 8%, self-assessed record keeping 7%, and peer assessed record keeping 4%. Appendix 3 shows that significant improvements were made on each individual indicator for client communication and record keeping and for self- and peer assessment. Appendix 3 also shows that the highest self-rated improvements were consistent with the highest peer rated improvements; for client communication indicator 3: '*Are the patient reported outcomes and performance outcomes used to develop a treatment plan formulated in dialogue with the client*' and for record keeping indicator 6: '*Is the use of performance measures (clinical tests) adequate?*'

Discussion

We evaluated the impact of an innovative QI program based on self-assessment and peer assessment on clinical performance in physical therapy practice. Our hybrid program containing both online and face-to face learning, allowed for powerful learning experiences. The results showed that the program was successful in supporting participants in achieving their personal goals and improving their performance on all quality indicators for client communication and record keeping. The results on commitments to change show that the majority was focused on client communication and to a lesser extent to record keeping and performance- and outcome measurement. Because a major part of physical therapists in the Netherlands is familiar with the evaluation of clinical records (including the measurements used) but unfamiliar with the evaluation of client communication, this outcome was no surprise. Research on client-communication in the physical therapy domain shows considerable room for improvement in this respect.[45,46] We also learned that personal goals showed strong agreement with the distinct performance indicators for client communication and record keeping. Apparently, the performance indicators triggered a need for change in routine practice and guided the QI process towards its intended outcomes as personal goals addressing other areas were scarce (7.6%). In this respect, the use of performance indicators was effective. On the other hand, indicator scores in cycle 1 may have encouraged participants to take a reductionist (short-cut) approach to performance appraisal in cycle 2, narrowing the scope on areas that need improvement. Encouraging feedback providers to underpin their online scores with narrative feedback and encouraging feedback receivers to reflect on both quantitative and qualitative feedback, might trigger participants to broaden their scope.

Table 3 — Differences in mean indicator scores between cycle 1 and 2 using a 5-point Likert scale

	Cycle 1 n	Min	Max	Mean	Med	SD	NA[1]
Client communication							
SA[6] scores	351	1	5	3.69	3.76	0.70	4.9%
PA[7] scores	351	2	5	4.05	4.08	0.50	2.2%
Record keeping							
SA[6] scores	345	1	5	3.79	3.83	0.66	1.8%
PA[7] scores	351	2	5	4.24	4.30	0.42	1.7%

* Significant on a .01 level, [1] Proportion of perceived 'not-applicable' or 'not relevant' indicator scores,
[2] Number of participants active in both cycle 1 and 2, [3] Mean Difference, [4] Confidence Interval,
[5] Intraclass correlation coefficient, [6] Self-assessment, [7] Peer assessment

The results on performance improvement show that the median percentage ranged from 4% to 11% and that approximates the findings of the meta review of Ivers *et al*.[25] Self-assessment scores were lower than peer assessment scores in both cycle 1 and 2, indicating that participants either underestimated themselves or overestimated their peers and these outcomes are supported by our feasibility study[31] and are in line with the literature on self-assessment and peer assessment.[20,47,48] In cycle 2 differences between self-assessment and peer assessment scores diminished due to higher self-assessment scores, in particular, the scores for client communication. Awareness of clinical performance and an improved self-concept after cycle 1, may have contributed to the improvements made in cycle 2 as underpinned by the feasibility study.[31] Participants in the feasibility study were reluctant in exposing clinical performance to 'an audience' and being insecure about their own performance, they were cautious in critically appraising their peers. We assume that extended exposure to critical performance appraisal, reinforcement of desired performance by peers, and role modelling may have contributed to improved self-efficacy beliefs and increased motivation to work on personal goals as explained by cognitive learning- and self-determination theory.[4,12] Further research is necessary to underpin this assumption.

Mean self-assessment and peer assessment scores at baseline were high, showing limited room for improvement. That raises the question of how much room for improvement is left, in particular

Cycle 2 n	Min	Max	Mean	Med	SD	NA[1]	n[2]	MD[3]	Change	*p*-value	95% CI[4] Lower Bound	95% CI[4] Upper Bound	ICC[5]
314	2	5	4.14	4.17	0.59	2.53%	311	0.44	11%	.000*	0.36	0.52	0.005
333	2	5	4.36	4.42	0.46	1.32%	328	0.30	8%	.000*	0.24	0.35	0.098
310	1	5	4.14	4.17	0.58	1.55%	307	0.31	7%	.000*	0.24	0.39	0.000
328	2	5	4.45	4.55	0.40	0.82%	325	0.20	4%	.000*	0.16	0.24	0.096

for high performing physical therapists. Their motivation might lower when the ceiling effect occurs, challenging the sustainability of the system for QI purposes. Creating more room for improvement requires accurate and critical performance assessors on the one hand and high performance standards on the other. We suggest that the program should further develop in both directions: 1) continuous improvement of critical self-assessment and peer assessment skills supported by well-trained coaches, and 2) development of performance indicators for a variety of competency domains and setting higher performance standards tailored to, and informed by, high performing professionals.

Looking at the improvements made on different performance indicators, the results show that that the highest self-rated improvements were consistent with the highest peer rated improvements which strengthens the validity of the assessment outcomes.

Strengths and limitations

We developed an innovative QI program enabling participants to provide personalized feedback, tailored to the competency domains that need improvement. Program evaluation was based on a high response and rich data.

The fact that physical therapists self-selected their client records and video-recordings can be considered as a limitation because the materials might not reflect their authentic clinical practice. Nevertheless, we have chosen this option to provide participants the opportunity to get used to exposing their clinical behaviours to their peers and not to jeopardize group safety. Self-selected or not, video-recordings and client records provided powerful learning

material allowing for critical reflection on current performance. In addition, we assume that anticipating this learning experience might have triggered improvement beforehand.
Another limitation is that performance indicators were used for both educational purposes – to prospectively guide the QI process towards the program goals – and for scientific purposes – to measure retrospectively the impact on clinical performance. Knowledge of performance indicators might have biased the true impact. Moreover, this was a non-controlled study testing a short-term intervention. The robustness and the sustainability of the results are currently unclear.

Conclusions

Our study demonstrated that self- and peer assessment including conscious goal setting is effective in enhancing commitments to change and improving clinical performance of physical therapists, and despite the limitations mentioned, nation-wide implementation is justified.
The results are promising regarding self-regulation of healthcare quality and relevant to all professionals and organizations engaged in bottom-up quality assurance.
A challenge to ongoing program development is to design quality indicators that facilitate the QI process for both low- and high performing physical therapists, and address a variety of competency domains. Further research should determine the sustainability of the results and the impact on client outcomes.

Abbreviations

MM = Marjo Maas

FD = Femke Driehuis

AB = Annick Bakker-Jacobs

References

1 Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century*. Washington, DC: National Academy Press; 2001.

2 Ministerie van Volksgezondheid en Welzijn [Ministry of Health and Welfare]. BIG register [BIG registry]. https://www.bigregister.nl/. Accessed February 27, 2016.

3 Eijkenaar F. Pay-for-performance for healthcare providers. Design, performance measurement, and (unintended) effects. *Health Policy (New York)*. 2013;110(2-3):115-130.

4 Ryan R, Deci E. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol*. 2000;55(1):68-78.

5 Scholte M, Neeleman-Van Der Steen CWM, Van Der Wees PJ, Nijhuis-Van Der Sanden MWG, Braspenning J. The reasons behind the (non)use of feedback reports for quality improvement in physical therapy: A mixed-method study. *PLoS One*. 2016;11(8):1-16.

6 Littlefield L, Hewat C. *Harnessing self-regulation to support safety and quality in healthcare delivery.* 2012.

7 van der Wees PJ, Nijhuis-van der Sanden MW, Ananian JZ, Black N, Westert GP, Schneider EC. Integrating the use of patient-reported outcomes for both clinical practice and performance measurement: views of experts from 3 countries. *Milbank Q*. 2015;93(4):788-825.

8 Grol R. Quality improvement by peer review in primary care: a practical guide. Qual Health Care. 1994;3(3):147-152. http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1055218&tool=pmcentrez&rendertype=abstract.

9 Butterfield R, McCormick B, Anderson R, Ball J, White J, Eleftheriades C. Quality of NHS Care and External Pathway Peer Review.; 2012. https://www.chseo.org.uk/downloads/report3-peerreview.pdf.

10 Prochaska JO, Redding CA, Evers KE. Health behavior and health education. In: Glanz K, Rimer BK, Viswanath K, eds. *Health Behavior and Health Education: Theory, Research, and Practice*. 4th ed. Wiley & Sons; 2008:97-121.

11 Ajzen I. Nature and operation of attitudes. *Annu Rev Psychol*. 2001;52:27-58.

12 Bandura A. Self-Efficacy: *The Exercise of Control*. John Wiley & Sons, Inc.; 1997.

13 Kluger AN, Nir D. The feedforward interview. *Hum Resour Manag Rev.* 2010;20(3):235-246.

14 Pronovost PJ, Hudson DW. Improving healthcare quality through organisational peer-to-peer assessment: lessons from the nuclear power industry. *BMJ Qual Saf.* 2012;21(10):872-875.

15 Maas MJM, van Dulmen SA, Sagasser MH, *et al*. Critical features of peer assessment of clinical performance to enhance adherence to a low back pain guideline for physical therapists: a mixed methods design. *BMC Med Educ*. 2015;15(1):203.

16 Buetow SA, Roland M. Clinical governance: bridging the gap between managerial and clinical approaches to quality of care. *Qual Heal Care*. 1999;8:184-190.

17 Meerhoff GA, van Dulmen SA, Maas MJ, Heijblom K, Nijhuis-van der Sanden MW, van der Wees PJ. Development and evaluation of an implementation strategy for collecting data in a national registry and the use of patient-reported outcome measures (PROMS) in physical therapist practice: quality improvement study. *Phys Ther.* 2017;Published ahead of print.

18 Westby MD, Klemm A, Li LC, Jones CA. Emerging Role of Quality Indicators in Physical Therapist Practice and Health Service Delivery. *Phys Ther*. 2016;96(1):90-100.

19 Epstein RM. Self-Monitoring in clinical practice. *J Contin Educ Health Prof.* 2008.

20 Davis DA, Mazmanian PE, Fordis M, Harrison R Van, Thorpe KE, Perrier L. Accuracy of physician self-assessment compared with observed measures of competence. A systematic review. *JAMA*. 2006;296(9):1094-1102.

21 Harting J, Rutten GM, Rutten ST, P KS. A qualitative application of the diffusion of innovations theory to examine determinants of guideline adherence among physical therapists. *Phys Ther.* 2009;89(3):221-232.

22 Maas MJ, van der Wees PJ, Braam C, *et al.* An innovative peer assessment approach to enhance guideline adherence in physical therapy: single-masked, cluster-randomized controlled trial. *Phys Ther.* 2015;95(4):600-612.
23 van Dulmen SA, Maas MJ, Staal JB, *et al.* Effectiveness of peer assessment for implementing a Dutch physical therapy low back pain guideline: cluster randomized controlled trial. *Phys Ther.* 2014;94(10):1396-1409..
24 Maas MJM, van Dulmen SA, Sagasser MH, *et al.* Critical features of peer assessment of clinical performance to enhance adherence to a low back pain guideline for physical therapists: a mixed methods design. *BMC Med Educ.* 2015;15(1):203.
25 Ivers N, Jamtvedt G, Flottorp S, *et al.* Audit and feedback: effects on professional practice and health care outcomes (Review). Cochrane Database Syst Rev. 2012;(7):1-227. http://onlinelibrary.wiley.com/doi/10.1002/14651858.CD000259.pub2/pdf/standard. Accessed May 7, 2014.
26 Sargeant JM, Lockyer J, Mann K, *et al.* Facilitated reflective performance feedback: developing an evidence- and theory-based model that builds relationship, explores reactions and content, and coaches for performance change (R2C2). *Acad Med.* 2015;90(12):1698-1706.
27 Payne VL, Hysong SJ. Model depicting aspects of audit and feedback that impact physicians' acceptance of clinical performance feedback. *BMC Health Serv Res.* 2016;16(1).
28 Hysong SJ, Kell HJ, Petersen LA, Campbell BA, Trautner BW. Theory-based and evidence-based design of audit and feedback programmes: examples from two clinical intervention studies. *BMJ Qual Saf.* 2016;0(6):1-12.
29 Brehaut JC, Eva KW. Building theories of knowledge translation interventions: use the entire menu of constructs. *Implement Sci.* 2012;7:114.
30 Craig P, Dieppe P, Macintyre S, *et al.* Developing and evaluating complex interventions: new guidance. *BMJ.* 2008;337:a1655.
31 Maas MJ, Nijhuis-van der Sanden MW, Driehuis F, Heerkens YF, van der Vleuten CP, van der Wees PJ. Feasibility of peer assessment and clinical audit to self-regulate the quality of physiotherapy services: a mixed methods study. *BMJ Open.* 2017;7:1-10.
32 Øvretveit J. *Evaluating Improvement and Implementation for Health*. 1st ed. New York: McGraw-Hill Education; 2014.
33 Koninklijk Genootschap van Fysiotherapeuten. Centraal Kwaliteitsregister Fysiotherapie. https://www.kngf.nl/vakgebied/kwaliteit/ckr.html. Published 2016. Accessed February 27, 2016.
34 Hoens A, Li L. The knowledge broker's "fit" in the world of knowledge translation. *Physiother Canada*. 2014;66(3):223-227.
35 de Vries C, Hagenaars L, Kiers H, Schmitt M. The physical therapist – a professional profile. https://www.kngf.nl/vakgebied/vakinhoud/beroepsprofielen.html. Published 2014. Accessed October 1, 2016.
36 Frank J, Jabbour M, Tugwell P, Boyd D, Fréchette D, Labrosse J. The CanMEDS 2005 Physician Competency Framework. http://www.royalcollege.ca/portal/page/portal/rc/common/documents/canmeds/framework/the_7_canmeds_roles_e.pdf. Published 2005.
37 Compusense Business Avionics. www.compusense.nl.
38 Sargeant JM, Mann K V, van der Vleuten CP, Metsemakers JF. Reflection: a link between receiving and using assessment feedback. *Adv Health Sci Educ Theory Pract.* 2009;14(3):399-410.
39 Hysong SJ. Meta-analysis: audit and feedback features impact effectiveness on care quality. *Med Care.* 2009;47(3):356-363.
40 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):284-299.
41 Archer JC. State of the science in health professional education: effective feedback. *Med Educ.* 2010;44(1):101-108.
42 Wakefield J, Herbert CP, Maclure M, *et al.* Commitment to change statements can predict actual change in practice. *J Contin Educ Health Prof.* 2003;23(2):81-93.

43 King N, Cassel CM, Symon G. Using templates in the thematic analysis of texts. In: Cassell C, Symon G, eds. *Essential Guide to Qualitative Methods in Organizational Research*. 1st ed. London: Sage Publications; 2004:256-270.
44 Field A. *Discovering Statistics Using SPSS*. 4th ed. Sage Publications; 2013.
45 Dierckx K, Deveugele M, Roosen P, *et al*. Implementation of shared decision making in physical therapy: observed level of involvement and patient preference. *Phys Ther.* 2013;93(10):1321-1330.
46 Oostendorp RAB, Elvers H, Mikołajewska E, *et al*. Manual physical therapists' use of biopsychosocial history taking in the management of patients with back or neck pain in clinical practice. *Sci World J*. 2015.
47 Violato C, Lockyer J. Self and peer assessment of pediatricians, psychiatrists and medicine specialists: implications for self-directed learning. *Adv Health Sci Educ Theory Pract.* 2006;11(3):235-244.
48 Speyer R, Pilz W, van der Kruis J, Brunings JW. Reliability and validity of student peer assessment in medical education: a systematic review. *Med Teach*. 2011;33(11):572-585.
49 Tobergte DR, Curtis S. Revised standards for quality improvement reporting excellence (SQUIRE 2.0). *J Chem Inf Model*. 2013;53(9):1689-1699

Appendix 1 — Performance indicators

Appendix 1a — Self-assessment and peer assessment form client communication

Instruction
1–5: 1 = much improvement needed; 5 = no improvement needed
When improvement is needed, please provide written feedback and tips for improvement.
N = not relevant/not applicable

	Performance indicators and corresponding quality domains	1	2	3	4	5	N	Feedback and tips
1[1]	Is the help request clarified?	O	O	O	O	O	O	
2[1]	Are the findings of the intake and clinical examination clearly communicated in understandable, client-friendly language?	O	O	O	O	O	O	
3[1,2,3]	Are the patient reported outcomes and performance outcomes used to develop a treatment plan in dialogue with the client?	O	O	O	O	O	O	
4[1,2,3]	Are the outcome expectancies of therapist and client aligned?	O	O	O	O	O	O	
5[1,2,3]	Are the outcome expectancies formulated SMART (specific, measurable, acceptable, realistic, time contingent)?	O	O	O	O	O	O	
6[1]	Are the interventions clearly communicated in dialogue with the client?	O	O	O	O	O	O	

[1] quality domain 'client-centeredness'
[2] quality domain 'effectiveness' including evidence based practice.
[3] quality domain 'transparency'

Additional remarks

Appendix 1b — Self-assessment and peer assessment form record keeping

Instruction
1–5: 1 = much improvement needed; 5 = no improvement needed
When improvement is needed, please provide written feedback and tips for improvement.
N = not relevant/not applicable

Performance indicators and corresponding quality domains			1	2	3	4	5	N	Feedback and tips
1[3]	Readability	Is the record written in plain language and is reporting concise?	O	O	O	O	O	O	
2[3]	Completeness	Does the record adhere to the KNGF guideline for record keeping 2016?	O	O	O	O	O	O	
3[3]	Transparency	Is the process of clinical reasoning and decision making transparent?	O	O	O	O	O	O	
4[1,2,3]	Consistency	Are the different steps in the process of diagnosis, treatment, and evaluation consistent with each other (are there no contradictory steps)?	O	O	O	O	O	O	
5[1,2,3]	Client reported outcome measures (questionnaires)	Is the use of client reported outcome measures (if relevant) adequate?	O	O	O	O	O	O	
6[1,2,3]	Performance measures (clinical tests)	Is de use of performance measures (if relevant) adequate?	O	O	O	O	O	O	

[1] quality domain 'client-centeredness'
[2] quality domain 'effectiveness' including evidence based practice
[3] quality domain 'transparency'

Additional remarks

Appendix 2 — Quality improvement program

Program content	Planned activities	Intended effects
Introduction meeting	Information on the program aims, content, procedures, expected investment in time and effort. Discussion of perceived barriers to participation.	→ Improved motivation to participate. Adherence to the program guidelines.
Cycle 1 Self- and peer assessment of clinical performance. Individual level.	Uploading client record and video-recording of client communication.	
	Online self-assessment of clinical performance. Receiving peer feedback.	→ Critical self-reflection on personal performance.
	Online peer assessment of clinical performance and providing peer feedback.	→ Critical reflection on peer performance.
	Face-to-face discussion.	→ Alignment of performance standards. Adjustment of personal views. Awareness of performance and learning needs.
	Designing and uploading personal improvement goals.	→ Considering behaviour change. Deciding on personal goals and action plans. Working on personal goals.
Cycle 2 Self- and peer assessment of clinical performance. Individual level.	Uploading client record and video-recording of client communication.	
	Online self-assessment of clinical performance.	→ Critical reflection on personal performance.
	Online peer assessment of clinical performance and providing peer feedback.	→ Critical reflection on peer performance.
	Face-to-face discussion. Evaluating personal goals.	→ Alignment of performance standards with peers. Adjustment of personal views. Awareness of performance improvement.
		Improved clinical performance

Appendix 3 — Effects on performance improvement for each indicator

Table 4.1 — Differences in self-assessment scores client communication between cycle 1 and 2

	Mean Difference	**SD**	***P*-value**	**95% CI of Difference** Lower Bound	Upper Bound
Is the help request clarified?	0.46	1.16	.000*	0.33	0.60
Are the findings of the intake and clinical examination clearly communicated in understandable, client-friendly language?	0.24	0.93	.000*	0.13	0.35
Are the patient reported outcomes and performance outcomes used to develop a treatment plan formulated in dialogue with the client?	**0.65**	**1.44**	**.000***	**0.48**	**0.82**
Are the outcome expectancies of therapist and client aligned?	0.46	1.18	.000*	0.33	0.60
Are the outcome expectancies formulated SMART?[1]	0.58	1.25	.000*	0.44	0.73
Are the interventions clearly communicated in dialogue with the client?	0.30	1.09	.000*	0.17	0.43

[1] Specific, Measurable, Acceptable, Realistic, time contingent

* Significant at a .01 level

NB: Area of greatest improvement is printed bold

Table 4.2 — Differences in peer assessment scores client communication between cycle 1 and 2

	Mean Difference	**SD**	***P*-value**	**95% CI of Difference**	
				Lower Bound	Upper Bound
Is the help request clarified?	0.32	0.78	.000*	0.23	0.40
Are the findings of the intake and clinical examination clearly communicated in understandable client-friendly language?	0.15	0.64	.000*	0.08	0.21
Are the patient reported outcomes and performance outcomes used to develop a treatment plan formulated in dialogue with the client?	**0.51**	**0.98**	**.000***	**0.40**	**0.61**
Are the outcome expectancies of therapist and client aligned?	0.31	0.68	.000*	0.23	0.38
Are the outcome expectancies formulated SMART?[1]	0.33	0.74	.000*	0.25	0.41
Are the interventions clearly communicated in dialogue with the client?	0.19	0.65	.000*	0.12	0.26

[1] Specific, Measurable, Acceptable, Realistic, Time contingent
* Significant at a .01 level

Table 4.3 — Differences in self-assessment scores record keeping between cycle 1 and 2

	Mean Difference	**SD**	***P*-value**	**95% CI of Difference** Lower Bound	Upper Bound
Is the record written in plain language and is reporting concise?	0,28	0,74	.000*	0.19	0.28
Does the record adhere to the guideline for record keeping 2016?	0.23	0.89	.000*	0.13	0.33
Is the process of clinical reasoning and decision making transparent?	0.40	0.89	.000*	0.30	0.50
Are the different steps in the process of diagnosis, treatment and evaluation consistent with each other (are there no contradictory steps)?	0.21	0.86	.000*	0.11	0.30
Is the use of Patient Reported Outcome measures adequate?	0.35	1.13	.000*	0.22	0.49
Is the use of performance measures (clinical tests) adequate?	**0.47**	**1.32**	**.000***	**0.30**	**0.65**

* Significant at a .01 level
NB: Area of greatest improvement is printed bold

Table 4.4 — Differences of peer assessment scores record keeping between cycle 1 and 2

	Mean Difference	SD	*P*-value	95% CI of Difference	
				Lower Bound	Upper Bound
Is the record written in plain language and is reporting concise?	0.19	0.42	.000*	0.14	0.24
Does the record adhere to the guideline for record keeping 2016?	0.16	0.49	.000*	0.11	0.21
Is the process of clinical reasoning and decision making transparent?	0.19	0.51	.000*	0.13	0.24
Are the different steps in the process of diagnosis, treatment and evaluation consistent with each other (are there no contradictory steps)?	0.14	0.51	.000*	0.08	0.19
Is the use of Patient Reported Outcome measures adequate?	0.25	0.66	.000*	0.18	0.32
Is the use of performance measures (clinical tests) adequate?	**0.26**	**0.71**	**.000***	**0.18**	**0.33**

[1] Specific, Measurable, Acceptable, Realistic, Time contingent
* Significant at a .01 level
NB: Area of greatest improvement is printed bold

Chapter 8
The utility of an online script concordance test to enhance clinical reasoning in physical therapy education and professional practice

Marjo Maas
Ria Nijhuis – van der Sanden
Geert Rutten
Yvonne Heerkens
Philip van der Wees
Cees van der Vleuten

Submitted for publication

Abstract

Background

Clinical reasoning is considered as a critical competency for the quality of physical therapist (PT) care. The script concordance test (SCT) aims to assess clinical reasoning in the context of uncertainty.

Objectives

To explore the utility of the SCT as a tool to enhance clinical reasoning in the musculoskeletal domain of undergraduate PT education and professional practice, by assessing its reliability, validity, and acceptability.

Design

Cross-sectional validation study

Participants

The SCT was administered to 741 PT students and 562 professionals.

Methods

We developed a computer-based SCT aimed at reducing unwanted variation in PT care. It contained 18 clinical scripts each followed by 3 test-items (n=54), some illustrated by video-recordings. Completion time was limited to 100 minutes.
Internal consistency was tested with Cronbach alpha. We pre-defined 7 expertise levels: 4 student and 3 professional levels. Informed by dual processing theory, construct validity was assessed by testing the hypothesis that higher expertise levels would produce higher SCT-scores in less response time. We tested in-between level differences for SCT-cores and response time with UNIANOVA linear models. Acceptability was explored with a 6-item questionnaire which could be scored on a 5-pnt Likert scale.

Results

Cronbach alpha was 0.69. Mean SCT-scores differed significantly between students and professionals: mean difference=6.08; $p<.001$. Higher expertise was related to higher SCT-scores but in-between differences were not always significant. Unlike our hypotheses, students used less response time than professionals: mean difference=0.55 minutes, $p<.001$. Mean acceptability scores varied (students: 2.88-3.80; professionals: 3.00-4.00).

Limitations

During testing, students were supervised and professionals were not.

Conclusions

The SCT is a promising tool to enhance clinical reasoning. Its quality can improve by increasing the number and variety of scripts.

Introduction

Assessment of clinical competence can be used for different purposes. When assessment is used for summative purposes, its results are used by universities or professional organizations to make decisions regarding academic progress, certification or accreditation. When assessment is used for formative purposes, the outcomes are used to support continuous learning and quality improvement. For example, the results can be used by individuals to identify gaps in actual performance and to inform the process of developing new knowledge and skills.[1] Organizations can use the results to evaluate their organizational goals and to benchmark their output.[2]

Clinical reasoning is generally considered as a critical competency of a physical therapists (PTs) for the quality and safety of PT care,[3] especially since PTs in the Netherlands – and in many other countries – are directly accessible without referral of a physician.[4,5] Clinical practice guidelines provide the best available evidence on clinical problems to support the clinical reasoning process.[6] However, guidelines are not available for all clinical problems or the context of the clinical problem is not appropriate to apply them. The Sicily statement on evidence-based practice (EBP) argues for the integration of the necessary knowledge, skills and attitudes into the curricula.[7] However, research showed that there are several barriers to applying EBP in clinical practice that can be attributed to patient factors, care givers factors, and encounter factors.[6] Students may face differences of opinion between teachers and clinical instructors, and those differences might be acceptable or not.

When assessing clinical reasoning competency, the pursuit of full consensus on the best clinical decision does not adequately reflect the heterogeneity of approaches to solving a clinical problem in PT practice. Creating an answering key that contains different correct answers which together indicate rather a trend than a single best answer, might better reflect the ambiguous nature of clinical problem solving. The Script Concordance Test (SCT) – originally designed for medical education – aims to assess clinical reasoning in the context of uncertainty allowing for variation in best answers.[8]

An SCT question consists of a short clinical case (script) followed by pieces of additional information (scenario) relating to diagnosis or

treatment, each followed by a test item. On a 5-point Likert scale participants indicate the effect of the additional information on the plausibility of the diagnostic hypothesis or the appropriateness of the proposed action. The participant's response to each question is compared with the answers of an expert panel. Credit is assigned to each response based on how many of the experts on the panel choose that response. A maximum score of 100% is given for the modal response. Other responses are given partial credit, depending on the proportion of experts choosing them. Responses not selected by experts receive zero points.[9] This so called 'aggregate scoring method' was found to have superior psychometric properties according to a study of Goos *et al*.[10] who compared different scoring methods on the same SCT including the single best answering method. A systematic review of Lubarsky *et al*.[13] on the SCT showed that research generally supports the use of the SCT to assess clinical reasoning in the context of uncertainty. However, validity evidence of SCT scores varies and requires verification in different contexts and for particular SCT designs. The SCT generally produces good content validity because of clear construction guidelines as well as high internal reliability.[11,12] In contrast to the review of Lubarsky, a review of Lineberry *et al*.[13] shows that the reports on the validity and reliability of the SCT are often too optimistic, for example, measurement errors caused by insufficient expertise within the expert panel are not sufficiently taken into account. Moreover, Lurie *et al*.[16] argues – supported by studies on clinical reasoning and decision making in the PT domain[14,15] – that experts (specialists) solve problems differently than novices as explained by dual processing theory. Dual processing theory distinguishes two cognitive reasoning systems: 1) the 'automated (implicit) processing mode' and 2) the 'controlled (explicit) processing mode'. System 1 is fast, intuitive, operates with little effort, requires little control or attention and is associated with 'pattern recognition' as clinical reasoning strategy.[17] System 2 is slow, step-by-step, requiring full attentional control, and is associated with 'hypothetico-deductive reasoning'[18] or 'rule-based forward reasoning' strategies.[5] Although research showed that novices and experts use both systems, experts can rely on the intuitive recognition of clinical patterns (scripts) based on their training and experience in solving domain-specific problems.[19]
The literature on assessment of clinical competence showed convincingly that there is no single best measure to assesses clinical reasoning[17] as it is an idiosyncratic, context specific and highly complex skill.[1,9,20-23] Irrespective of the summative or formative aim of an assessment instrument, its outcomes should provide valid,

reliable, and useful feedback for continuous improvement and that remains a challenge.[24] We explored the utility of the SCT as a feedback tool, aiming to enhance clinical reasoning expertise in the context of uncertainty. A pilot study in 2015 demonstrated the utility of the SCT in undergraduate PT education.[34]
We developed a new SCT tailored to both undergraduate and post-graduate PTs embedded in a quality improvement pilot study which included two additional feedback meetings organized by the Royal Dutch Society for Physical Therapy (KNGF). The feedback meetings addressed the anonymized results of the SCT and were open to students and professionals. Variation between professionals and between students was used as input for group discussions about which variation is acceptable and which is not. For example, variation was assumed acceptable when patient needs and preferences conflicted with guideline recommendations. Unacceptable variation concerned gaps in up-to-date knowledge and skills, threats to patient safety,[25] or unnecessary costs.[26]
Test utility was informed by the quality indicators for competency assessment described by van der Vleuten & Schuwirth.[27] Our research questions addressed: 1) test reliability, 2) test construct validity, 3) test acceptability for learning and improvement purposes.
Our validity argument was based on dual processing theory; we hypothesized that a) professionals would produce higher scores than students in b) less response time (including reading time) as they might rely on advanced – more efficient – clinical reasoning strategies, and that c) specialists in the musculoskeletal domain would outperform non-specialists in this respect.[5,16]

Methods

Design, context and participants

This was a cross-sectional validation study. The SCT was formally scheduled and administered to 1-4th year students of two universities (UNI-1 and UNI-2). They were provided with a personal password linked to their unique student ID. Participation was voluntary for both universities, but for UNI-1 participation was awarded with 2 education credits for full test completion. Professionals were approached by the KNGF using a digital newsletter. Full completion of the SCT was awarded with 8 credits for the KNGF quality register. Students of UNI-1 and UNI-2 completed the test in March 2016 under supervision. They were instructed to leave the test room after completing the test. Professionals completed the test independently (without supervision) between February 2016 and April 2016. Total

completion time was set on 100 minutes for both students and professionals.

Development of the SCT instrument and the answering key

The SCT was developed by a panel of 6 bachelor and master level PT teachers from 2 universities in the Netherlands: MM (PT, educational scientist, researcher), JW (PT, manual therapist, educational scientist) HN (PT, movement scientist), and FS, FA, DB (PT, manual therapist). Because we couldn't rely on an existing SCT in the PT domain, we adopted the guidelines for construction[12,28] and optimizing test items[29] developed for the medical domain and adapted the test to clinical reasoning in the context of primary PT practice. We designed a test content matrix to guide the development process, containing a variety of pathophysiological conditions, body regions, and difficulty levels informed by body of knowledge and skills of the national professional profile (appendix 1, table 1a and 1b).[30] The test was web-based, allowing us to use audio-visual materials to strengthen script designs and providing immediate feedback on the results including references to online available literature such as clinical practice guidelines. Each question could be commented in an open field if desired. Although guessing for the best answer was unavoidable, participants were asked to skip a script when they were unfamiliar with the clinical problem. Looking up for information was allowed, but not encouraged in view of the testing time. In total 18 scripts – each including 3 scenario's followed by a test item (n=54) – were developed. Based on our pilot study and informed by an additional pre-test, the standard on minimal response time was set on 1 minute per script (including 3 test-items, excluding scripts accompanied by video-recordings. The answering key was developed by a panel of 22 PTs composed on the basis of either broad general clinical experience or specific experience in musculoskeletal problems in primary PT care. The 'distance-from-mode' strategy as recommended by Gagnon *et al*.[31] and supported by Goos *et al*.[10] was used to optimize the answering key. Accordingly, we removed 5% of the panel responses (n=67) according to this procedure. Appendix 1 shows the details on the test development, the scoring algorithm, the answering key, and its optimizing procedure.

Outcome measures

Test reliability in terms of the internal consistency of 18 SCT scripts and 54 items was explored with item analysis and tested with Cronbach alpha. To explore the construct validity of the SCT we calculated differences in mean total scores on the SCT and mean

response time for each competency level. Based on our hypothesis that higher SCT scores would relate to higher levels of expertise, we pre-defined 4 bachelor levels (1st – 4th year) and 3 professional levels labeled by specialization in the musculoskeletal domain (5=specialized in different domain; 6=generalist, not specialized in specific domain, 7=specialized in musculoskeletal domain). We assumed that PTs specialized in a different domain, would be less competent in solving domain specific problems than generalist who are ´specialized´ in solving a variety of problems.
The test acceptability was tested with a short 6-item questionnaire addressing three issues: 1) perceived difficulty of the SCT, 2) impact of the SCT on improvement activities, and 3) appropriateness of the SCT to enhance clinical reasoning in the musculoskeletal domain. Questions could be scored on a 5-point Likert scale; higher scores indicated higher acceptability.

Data sampling

We sampled online SCT scores, questionnaire scores, and relevant characteristics of students such as bachelor entry-level (pre-university education), bachelor level, gender and age. For professionals we sampled the specialization domain, setting, years of experience, gender and age (bachelor entry-level was considered not relevant for professionals). The software company Infoland[42] collected scores and logged data for each completed script. Data were exported in Microsoft Excel 2013 and imported in IBM SPSS 24. Skipped answers for each script (missing values) were replaced with 'zero' scores.

Data analysis

Item analyses was conducted to assess whether removing items or scripts from the test would improve its internal consistency.
We tested with linear regression if the student variables 'gender', 'bachelor entry level', and professional variables 'gender', 'setting', and 'experience' were significantly related to the outcome variable 'mean total SCT score' to allow for controlling for these potential confounders in case of significant distribution differences between competency levels. Distribution differences were tested with Chi-square tests. Before testing differences in mean total SCT scores, outliers were removed.
For each competency level we calculated mean scores per script (n=18), and mean total scores. The same procedure was applied for calculating response time. Based on our pilot test, participants who spent less than 1 minute on each script – that is reading the script including additional information and answering 3 items – were

considered as not seriously committed and left out of the analyses, irrespective of their competency level. Scores on the questionnaire on test acceptability, were described for each question and mean and median scores were calculated for both students and professionals. Subsequently we described and tested differences in mean total SCT scores between competency levels with UNIANOVA general linear models including Bonferroni post-hoc analysis using the 'mean total score' as dependent variable and 'competency level' as independent variable including covariates if present. The same procedure was conducted for describing and testing differences in mean total response time. Differences in the acceptability of the test between students and professionals were tested with ANOVA analysis of variance.

Results

The assessment was completed by 1303 participants: 741 students (597 of UNI-1 and 144 of UNI-2) and 562 professionals. Participants covered all undergraduate and postgraduate competency levels. Table 1 shows the details on participants' characteristics.

Test reliability

The internal consistency of 18 scripts was: $\alpha = 0.69$, and of 54 items was: $\alpha = 0.77$. Removing scripts or items would not have enhanced the internal consistency, so all scripts and items were included.

Test construct validity

We identified 18 outliers for mean total SCT scores distinguished by competency level (11 students en 7 professionals), and 15 non serious responders because of extreme low response time per script (11 students and 4 professionals); 6 students were identified as both outliers regarding extreme low scores and low response time. Three professionals were identified because of extreme unexplained high total completion time (more than 100 minutes).

Differences in SCT scores related to competency levels

Students skipped more questions (7.7%) than professionals (4.1%). Linear regression of the outcome variable 'mean total score' to identify confounders, showed that higher 'bachelor entry level' was significantly related to higher 'mean total score' (beta: 0.11; p=.002), and that the distribution between bachelor levels differed significantly, so we controlled for this confounder when testing in-between level differences for students.

Table 1 — Participant's characteristics

		Students N=741	Professionals N=562
Gender (Woman / %)		433 / 58.4	355 / 63.2
Age (Mean / SD)		20.97 / 2.18	43.38 / 11.83
Professional experience (Mean / SD)			19.42 / 11.72
Bachelor level			
1	1st year student (n / %)	173 / 23.3	
2	2nd year student (n / %)	182 / 24.6	
3	3rd year student (n / %)	213 / 28.7	
4	4th year student (n / %)	173 / 23.3	
Professional level			
	Specialization domain		
5	Pelvic conditions (n / %)		16 / 2.8
	Geriatric conditions (n / %)		20 / 3.6
	Oncologic conditions (n / %)		12 / 2.1
	Orofacial conditions (n / %)		5 / 0.9
	Psycho-somatic conditions (n / %)		23 / 4.1
	Other conditions (n / %)		56 / 10.0
6	General conditions (n / %)		285 / 50.7
7	Sports therapy (n / %)		25 / 4.4
	Student Master MS (n / %)		22 / 3.9
	Manual therapy (n / %)		129 / 23.0
Setting			
Setting Primary care (n / %)			453 / 81.0
Setting Hospital or Rehabilitation center (n / %)			82 / 14.7
Missing (n / %)			24 / 4.4

Mean SCT scores were significantly lower in students than professionals as expected: students: 63.94, SD=9.84; professionals: 70.00, mean difference=6.08; SD=8.56; $p \leq .001$, 95%CI=5.04-7.09). Table 2 shows that in line with our hypothesis, higher bachelor levels resulted in higher mean SCT scores, although differences were not significant between level 2-3, and 3-4. Similarly, higher professional levels (5-7) were associated with higher SCT scores, although differences between level 5-6 were not significant. Scores of specialists were significantly higher than all other competency levels. Figure 1 illustrates the differences between competency levels by presenting means and confidence intervals.

Table 2 — Differences in mean SCT scores between competency levels tested with UNIANOVA linear models and Bonferroni post hoc analyses

	Level[1]	Estimated Mean SCT score	Level[1]	Estimated Mean SCT score	Estimated Mean Difference	*p*-value	95% Confidence Interval	
							Lower Bound	Upper Bound
Students	1	56.22	2	64.68	-8.44	.001**	-10.91	-5.96
			3	66.68	-10.50	.001**	-12.89	-8.12
			4	67.26	-11.13	.001**	-13.64	-8.62
	2	64.68	3	66.68	-2.07	.123	-4.42	0.28
			4	67.26	-2.69	.025*	-5.17	-0.21
	3	66.68	4	67.26	-0.62	1.000	-3.02	1.77
Professionals	5	67.62	6	96.61	-1.99	.115	-4.29	0.31
			7	72.27	-4.64	.001**	-7.17	-2.11
	6	69.61	7	72.27	-2.66	.004**	0.66	4.65

* Significant at a 0.05 level, ** Significant at a 0.01 level, [1] 1=1st year student; 2=2nd year student, 3=3rd year student; 4=4th year student; 5=professional specialized in different domain; 6=professional not specialized.

Figure 1 — Mean total SCT scores for each competency level and their 95% confidence intervals

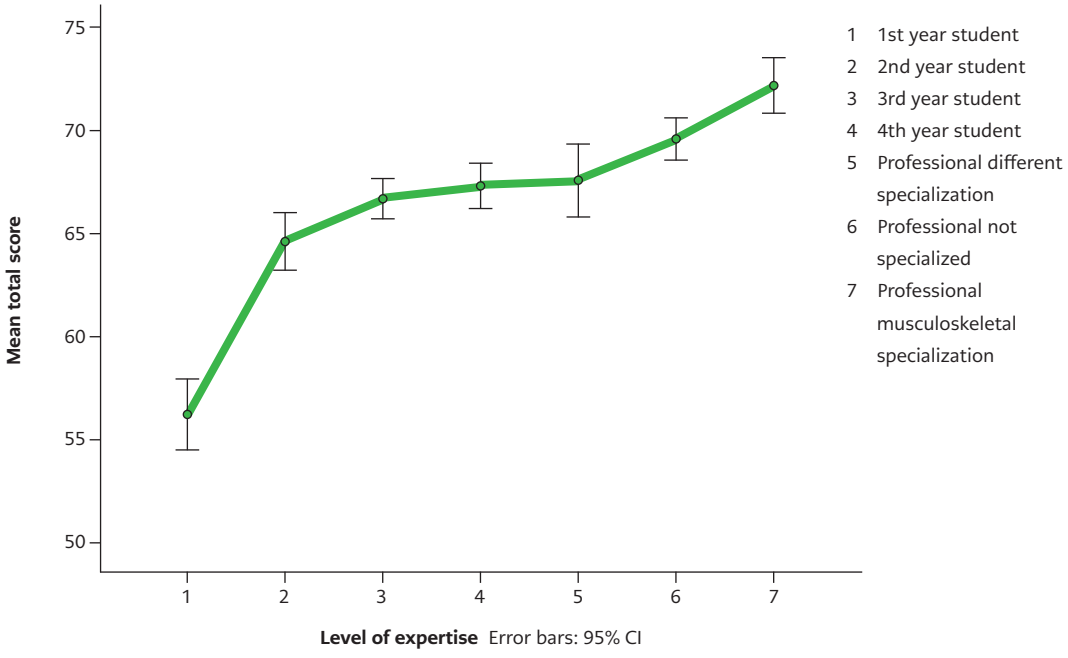


Table 3 — Differences in mean response time per script for each competency level tested with UNIANOVA linear models and Bonferroni post hoc analysis

	Level[1]	Mean response time	Level[1]	Mean response time	Mean Difference	Std. Error	*p*-value	95% Confidence Interval	
								Lower Bound	Upper Bound
Students	1	2.17	2	2.19	-0.02	0.06	1.000	-0.19	0.13
			3	2.14	0.03	0.06	1.000	-0.12	0.18
			4	1.97	0.20	0,06	.005**	0.04	0.37
	2	2.19	3	2.14	0.05	0.06	1.000	-0.09	0.20
			4	1.97	0.23	0,08	.049**	0.07	0.39
	3	2.14	4	1.97	0.17	0.06	.018**	-0.05	0.40
Professionals	5	2.82	6	2.67	0.15	0.09	.344	-0.08	0.39
			7	2.57	0.25	0.10	.055	-0.00	0.51
	6	2.67	7	2.57	0.09	0.08	.716	-0,10	0.30

** The mean difference is significant at the 0.01 level, [1] 1=1st year student; 2=2nd year student, 3=3rd year student; 4=4th year student; 5=professional specialized in different domain; 6=professional not specialized; 7=professionals specialized in musculoskeletal domain.

Figure 2 — Mean response time per script for each competency level and their confidence intervals

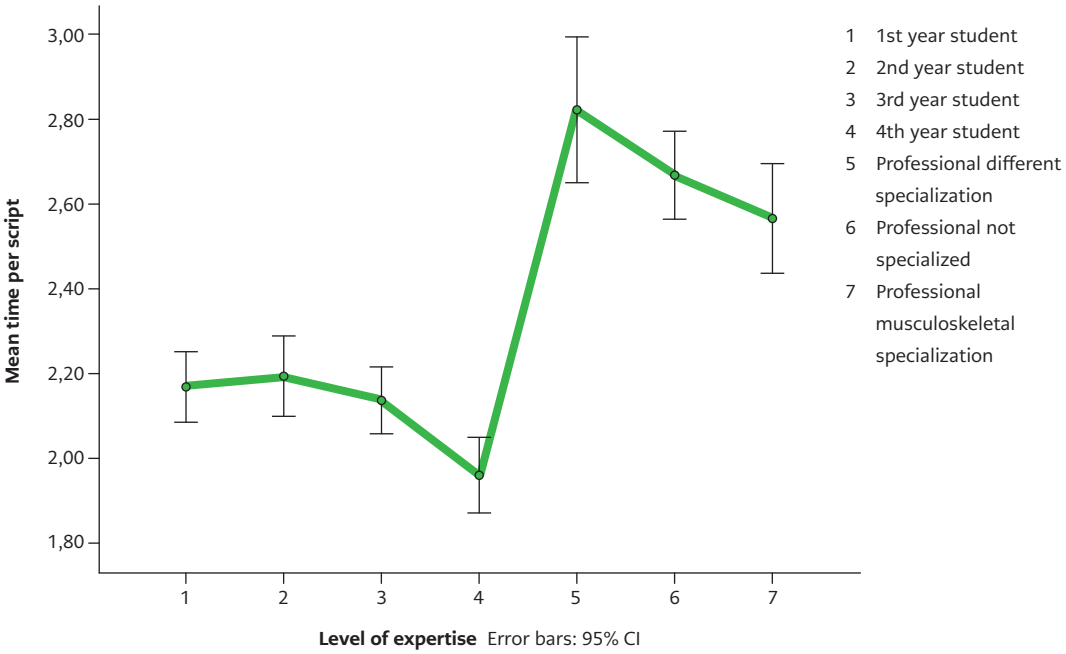


Table 4 — Acceptability scores of the script concordance test

Question	Students				
	Mean	Med	Min	Max	SD
How difficult did you perceive this assessment?[1]	3.42	4	2	5	.92
The assessment feedback provides insight into my strengths and weaknesses.[2]	3.31	4	1	5	1.12
This assessment stimulates me to search for additional information on the clinical problems described.[2]	3.66	4	1	5	.95
I will look for the literature referred to in this assessment.[2]	2.88	3	1	5	1.05
I think this test is appropriate to enhance the level of clinical reasoning of undergraduate physical therapists.[2]	3.81	4	1	5	.89
I think this test is appropriate to enhance the level of clinical reasoning of post graduate physical therapists.[2]	3.80	4	1	5	.94

[1] 1 = very easy, 2 = easy, 3 = not easy, not difficult, 4 = difficult, 5 = very difficult

[2] 1 = completely disagree, 2 = disagree, 3 = do not disagree and do not agree, 4 = agree, 5 = completely agree

Differences in response time related to competency levels

Participants completed the test within time limits: (mean completion time in minutes for students: 37.85, SD=10.10; professionals 47.94, SD=15.37). By linear regression of the outcome variable 'mean response time' we did not identify confounders in the student – or professional group. In contrast to our hypothesis, the mean response time in minutes per script was significantly higher for professionals than for students (students: 2.12, SD=0.57; professionals: 2.67, SD=0.85; mean difference 0.55, p=.000, 95%CI=0.62-0.47). Table 3 shows the differences in response time for each competency level. Figure 2 illustrates these differences by presenting means and confidence intervals.

In line with our hypothesis, 4th year students (level 4) used significantly less response time per script than lower level students, and specialists used less time than non-specialists, although these differences were not significant.

Professionals Mean	Med	Min	Max	SD	Mean Difference	p-value	95% Confidence Interval	
							Lower Bound	Upper Bound
3.00	3	1	5	.95	.42	.000**	.31	.52
3.52	4	1	5	1.00	-.21	.001**	-.31	-.08
3.74	4	1	5	.88	-.08	.153	-.18	.028
3.26	4	1	5	1.02	-.38	.000**	-.51	-.27
3.94	4	1	5	.82	-.13	.009**	-.23	-.03
4.00	4	1	5	.83	-.20	.000**	-.30	-.10

Test acceptability

The questionnaire results show that participant's perceptions varied widely for all acceptability domains, although they were generally positive. Mean acceptability scores varied: students 2.88-3.80 (SD=0.89-1.12); professionals 3.00-4.00 (SD=0.82-1.00).
As displayed in table 4, students perceived the test significantly more difficult than professionals, but the variation within the professionals group was substantially larger. Professionals were significantly more positive about the feedback function of the SCT, eager to seek for additional information on patient problems, likely to look at the literature referred to, and positive about the appropriateness of the test to enhance clinical reasoning competency.

Discussion

This study explored the utility of the SCT as a tool to enhance continuous improvement in undergraduate education and professional practice by describing and testing its reliability, validity and accept-

ability for quality improvement purposes. The results showed that the internal consistency of the SCT was acceptable, that test results increased with higher competency levels – although professionals used significantly more response time –, and that the test was generally perceived appropriate regarding its aim.

Although higher reliability coefficients (≥0.8) are reported in the literature,[11] we considered the internal consistency of the SCT acceptable given the heterogeneity of the 18 scripts varying in pathophysiological conditions, body regions, and difficulty levels (appendix 1 table 1a) and considering the fact that the SCT results were used for improvement feedback, not for high-stake decision making. However, based on research on the reliability and validity of clinical competency assessment, we assume that a larger sample of scripts might increase both the validity and reliability of the test, but that would require a longer attention span and might cause unwanted cognitive load.[23,27,32,33] Future research should determine if it does.

In line with our hypothesis, higher competency levels were associated with higher SCT scores and specialists in the musculoskeletal domain outperformed non-specialists. However, a dip in progress was observed at bachelor level 3 and 4 (figure 1) and this finding is supported by the results of our pilot-study.[34] Limited progress can be explained by factors related to the test, the student and his learning context. First, regarding the test, we consider the possibility that the difficulty levels of test scripts as presented in appendix 1, do not adequately distinguish between levels 2-3 and 3-4. Second, during their internships 3rd and 4th year students might lack triggers for the development of domain-specific knowledge in the musculoskeletal domain, because they are engaged in other content domains and contextual settings that differ from the learning context.[5,35] Previous studies have shown that students in the transition from preclinical learning to clinical practice, do not show improvement of basic knowledge – a critical feature of successful problem solving -, despite an increase of clinical experiences. This phenomenon is known as the 'intermediate dip'.[36,37]

Looking at the time-on-task spent by students and professionals, all professionals – irrespective of their specialization domain – spent significantly more time per script unlike our hypothesis. Based on these data, we can only try to find an explanation for this apparent contradiction informed by the literature. We suggest that the assumed association between automatic (implicit) system-1 processing and low time-on-task attributed to expert reasoning, was challenged by a combination of moderating factors related

to the SCT stimulus format, assessment context, script difficulty, and participant characteristics, that together might explain why professionals (including experts) used more instead of less time to complete the test. Literature showed that automatic processing is hindered when the processing procedure differs from routine processing.[38] The way information is presented in an SCT script and the sequence of presenting additional information, differs substantially from the way clinical information is sampled and analyzed in daily PT practice. In addition, the context of a computer-based assessment differs from the regular clinical encounter which might have triggered analytical processing (mode 2). Furthermore, Govaerts *et al*.[39] investigated how experienced and non-experienced raters select and use observational data to arrive at judgments and decisions about trainees' performance in the clinical workplace. Results showed that experts were faster in simple cases but needed more time in complex cases than novices, paying more attention to situation-specific cues in the assessment context.[39] Studies of Golhammer *et al*.[40] on a computerizes complex problem solving test supports this assumption. In line with this reasoning, and supported by Lurie *et al*.[16] professionals might have identified more cues than students, representing unwanted sources of ambiguity requiring more response time.[16,39] That specialists used less time-on-task than non-specialist can be attributed to their domain-specific problem solving skills. Literature also showed that test effort in a reasoning test – the extent to which a test taker cares about the result – is positively related to time-on-task and accuracy.[38,41] We assume that professionals were more intrinsically motivated to complete the test successfully than students. Moreover, students might have guessed more frequently which is generally associated with extremely short response time and less accuracy.[40] Future research is necessary to examine these assumptions.

Regarding the questionnaire results, professionals perceived the test significantly less difficult than students which can be expected, although the difference was small. That confirms the assumption that decision making in the context of the SCT was beyond professionals' comfort zone. Although professionals were significantly more positive about the acceptability of the SCT for quality improvement purposes than students, the results show that the SCT needs improvement regarding its feedback and feed forward function. Based on written questionnaire comments, we facilitated access to feedback and literature support within the test period, but that couldn't make up for the information loss of early enrolling partici-

pants, merely students. We assume that by improving the feedback function, other acceptability domains might improve simultaneously. It should be noted that the SCT was embedded in a quality improvement program including face-to-face feedback meetings. Perceptions on the value of SCT feedback might improve when the results are discussed.

Strengths and limitations

Although the SCT is not new, this is the first application within the musculoskeletal PT domain to our knowledge and the first study addressing both task response time and outcome. We developed a powerful assessment tool that was seen appropriate by participants to advance the level of clinical reasoning in undergraduate and postgraduate education.

A limitation of this study is that students were supervised and professionals were not. Although professionals were instructed to complete the test without interruption, we could not prevent that they telephoned or did otherwise while completing the test. Therefore the differences in time-on-task between students and professionals should be interpreted with caution.

Conclusion

The SCT is a promising tool to provide feedback for quality improvement purposes for physical therapy students and professionals. However, the content validity can be improved by increasing variety in the difficulty levels of scripts, and acceptability can be improved by facilitating its feedback and reference function.

Future research should explore 'how' students and professionals reason and 'what' drives them to successfully complete the test. Our results are interesting for all stakeholders in health professions education and clinical practice. By uncovering differences in clinical reasoning, the SCT might contribute to reducing unwanted variation in clinical decision-making in daily practice and bridging the transfer gap between education and workplace settings.

Abbreviations

MM = Marjo Maas

JW = Joost van Wijchen

HN = Henk Nieuwenhuijzen

FS = Fred Smedes

FA = Florian Abu Bakar

DB = Donald van der Burg

References

1 Epstein RM. Assessment in medical education. *N Engl J Med.* 2007;356(4):387-396.

2 Sessa VI, London M. *Continuous Learning in Organizations. Individual, Group, and Orananizational Perspectives.* 1st ed. Mahwah, New Jersey: Lawrence Erlbaum; 2006.

3 Association APT, ed. *Guide to Physical Therapist Practice.* 3rd ed.; 2011.

4 Ajjawi R, Higgs J. Learning to reason: a journey of professional socialisation. *Adv Health Sci Educ Theory Pract.* 2008;13(2):133-150..

5 Higgs J, Jones MA, Loftus S, Christensen N. *Clinical Reasoning in the Health Professions*. Third edit. Philadelphia: Elsevier Health Sciences; 2008.

6 Grol RP, Wensing M, Eccles MP, Davis DA, eds. *Improving Patient Care: The Implementation of Change in Health Care.* 2nd ed. Chichester, West Sussex: John Wiley & Sons, Inc.; 2013.

7 Dawes M, Summerskill W, Glasziou P, *et al*. Sicily statement on evidence-based practice. *BMC Med Educ.* 2005;5(1):1. doi:10.1186/1472-6920-5-1.

8 Charlin B, Gagnon R, Pelletier J, *et al*. Assessment of clinical reasoning in the context of uncertainty: the effect of variability within the reference panel. *Med Educ.* 2006;40(9):848-854.

9 Charlin B, van der Vleuten CPM. Standardized assessment of reasoning in contexts of uncertainty: the script concordance approach. *Eval Health Prof.* 2004;27(3):304-319.

10 Goos M, Schubach F, Seifert G, Boeker M. Validation of undergraduate medical student script concordance test (SCT) scores on the clinical assessment of the acute abdomen. *BMC Surg*. 2016;16(1):57.

11 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. *Med Educ.* 2011;45(4):329-338.

12 Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak.* 2008;8:18.

13 Lineberry M, Kreiter CD, Bordage G. Threats to validity in the use and interpretation of script concordance test scores. *Med Educ.* 2013;47(12):1175-1183.

14 Wainwright SF, Shepard KF, Harman LB, Stephens J. Factors that influence the clinical decsion making of novice and experienced physical therapists. *Phys Ther.* 2011;91(1).

15 Wainwright SF, Shepard KF, Harman LB, Stephens J. Novice and experienced physical therapist clinicians: a comparison of how reflection is used to inform the clinical decision-making process. *Phys Ther.* 2010;90(1):75-88.

16 Lurie SJ. Towards greater clarity in the role of ambiguity in clinical reasoning. *Med Educ.* 2011;45(4):326-328.

17 Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ.* 2005;39(1):98-106.

18 Engelbert R, Wittink H. *Klinische Redeneren Volgens de HOAC II [Clinical Reasoning according to the HOAC II*]. 1st ed. Houten: Houten: Bohn Stafleu van Loghum; 2010.

19 Pelaccia T, Tardif J, Triby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Med Educ Online.* 2011;16(1):1-9.

20 Eva KW. What every teacher needs to know about clinical reasoning. *Med Educ.* 2005;39(1):98-106.

21 Durning S, Artino AR, Pangaro L, van der Vleuten CPM, Schuwirth L. Context and clinical reasoning: Understanding the perspective of the expert's voice. *Med Educ.* 2011;45(9):927-938.

22 Pelaccia T, Tardif J, Triby E, Charlin B. An analysis of clinical reasoning through a recent and comprehensive approach: The dual-process theory. *Med Educ Online.* 2011;16(1):1-9.

23 Norman G. Working memory and mental workload. *Adv Health Sci Educ Theory Pract.* 2013;18(2):163-165.

24 Ramaekers S, Kremer W, Pilot A, van Beukelen P, van Keulen H. Assessment of competence in clinical reasoning and decision-making under uncertainty: the script concordance test method. *Assess Eval High Educ.* 2010;35(6):661-673.

25 Instute of Medicine. *Improving Diagnosis in Health Care.* (Balogh EP, Miller BT, Ball JR, eds.). National Academies Press; 2015.
26 Porter ME, Pabo EA, Lee TH. A stretegic vision to improve value by organizing around patients'needs. *Health Aff.* 2013;32(3):516-525.
27 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: From methods to programmes. *Med Educ.* 2005;39(3):309-317.
28 Dory V, Gagnon R, Vanpee D, Charlin B. How to construct and implement script concordance tests: Insights from a systematic review. *Med Educ.* 2012;46(6):552-563.
29 Gagnon R, Charlin B, Lambert C, Carrière B, Van Der Vleuten CPM. Script concordance testing: More cases or more questions? *Adv Heal Sci Educ.* 2009;14(3):367-375
30 de Vries C, Hagenaars L, Kiers H, Schmitt M. The physical therapist – a professional profile. https://www.kngf.nl/vakgebied/vakinhoud/beroepsprofielen.html. Published 2014. Accessed March 1, 2017.
31 Gagnon R, Lubarsky S, Lambert C, Charlin B. Optimization of answer keys for script concordance testing: should we exclude deviant panelists, deviant responses, or neither? *Adv Heal Sci Educ.* 2011;16(5):601-608.
32 Case SM, Swanson DB. Constructing written test questions for the basic and clinical sciences. *Director.* 2002;27(21):112.
33 Wass V, Van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. *Lancet.* 2001;357:945-949.
34 Maas MJM, Schilt-Mol T, van der Vleuten CPM. De online script concordance test om voortgang in klinisch redeneren te bevorderen van fysiotherapeuten in opleiding en in de beroepspraktijk [The online script concordance test to enhance progress in clinical reasoning in physiotherapy education and professional practice]. *Examens.* 2016;(2):12-18.
35 Durning SJ. Exploring the Influence of contextual factors of the clinical encounter on clinical reasoning success (unraveling context specificity). *Acad Med.* 2010;85(5):894.
36 Charlin B, Tardif J, Boshuizen HP. Scripts and medical diagnostic knowledge: theory and applications for clinical reasoning instruction and research. *Acad Med.* 2000;75(2):182-190.
37 Boshuizen HP, Bromme R, Gruber H, eds. *Professional Learning: Gaps and Transitions on the Way from Novice Tot Expert.* Dordrecht: Kluwer academic publishers; 2004.
38 Goldhammer F, Naumann J, Stelter A, Toth K, Rölke H, Klieme E. The time on task effect in reading and problem solving is moderated by task difficulty and skill: insights from a computer based large-scale assement. *J Educ Psychol.* 2014;106(3):608-626.
39 Govaerts MJ, Schuwirth LWT, Van der Vleuten CPM, Muijtjens AM. Workplace-based assessment: effects of rater expertise. *Adv Health Sci Educ Theory Pract.* 2011;16(2):151-165..
40 Goldhammer F, Naumann J, Greiff S. More is not Always Better: The Relation between Item Response and Item Response Time in Raven's Matrices. *J Intell.* 2015;3:21-40.
41 Silm G, Must O, Täht K. Test-taking effort as a predictor of performance in low-stakes tests. *Trames.* 2013;17(4):433-448.
42 https://www.infoland.nl. Accessed March 1, 2017.

Appendix — Development of the test content and the answering key

Studies on the reliability of the SCT indicate that an SCT covering a medical subdomain needs about 50–60 test items (nested within scripts) to achieve acceptable reliability ($\alpha \geq 0.8$).[36] The development panel constructed a test content matrix containing 18 clinical problems that adequately covered the musculoskeletal conditions in primary PT care coded by difficulty level. The difficulty level of the test-scripts was estimated based on the existence of a clinical guideline on the problem, its prevalence in primary care, and its complexity in signs and symptoms: 1=low difficulty, 2=medium difficulty, 3=high difficulty (see table 1a for an overview). We used a subset of 12 scripts from our pilot study and approached PT teachers / clinicians of UNI-1 and UNI-2 to provide 6 additional scripts including audio visual materials (if available) according to the test matrix. In total 18 scripts were developed each followed by 3 test-items (see table 1b for a test-item example).

Eighteen experts in the musculoskeletal domain validated the scripts and provided comments before developing the final answering key. Scripts were improved when necessary by the development panel. In total 22 panelists completed the test. The panel scores were described and the answering key was optimized by using the guidelines of Gagnon *et al.*[31] Studies by Gagnon *et al.* (2011) showed that when the panel size is sufficiently large (≥ 15), measurement error resulting from deviant panelists is negligible. Because these findings are based on medical education research, and not automatically generalizable to allied health education, we decided to use the 'distance-from-mode' strategy as recommended by Gagnon[31] and supported by Goos *et al.*[10] We removed answers more than one anchor distant from mode and answers provided by only 1 of the 22 panelists. Table 1a shows the number of panel answers removed.

Table 1a — Test matrix including number of panel answers removed

	Script content	Estimated difficulty level	Type of performance assessed	Video	Guideline	N of panel responses removed
1	Knee conditions	1	Diagnosis	Yes	Yes[1]	4
2	Pelvic – hip conditions	1	Diagnosis	No	Yes[1]	1
3	Foot – ankle conditions	1	Diagnosis	No	Yes[1]	6
4	Shoulder conditions	1	Diagnosis	Yes	Yes[2]	2
5	Neck conditions	1	Diagnosis	No	No	1
5	Hand – wrist conditions	2	Diagnosis and treatment	Yes	No	5
7	Low back conditions	2	Diagnosis and treatment	No	Yes[1]	3
8	Low back conditions	2	Diagnosis	No	Yes[1]	5
9	Shoulder conditions	2	Diagnosis	No	Yes[2]	1
10	Low back conditions	2	Treatment	No	Yes[1]	1
12	Shoulder conditions	2	Diagnosis	yes	Yes[2]	2*
13	Elbow conditions	2	Diagnosis	No	No	6
14	Whiplash	2	Treatment	Yes	Yes[1]	2
11	Foot – ankle conditions	3	Treatment	No	No	9
15	Neck – shoulder conditions	3	Diagnosis	Yes	No	6
16	Low back conditions	3	Diagnosis	No	No	4
17	Knee conditions	3	Treatment	No	Yes[1]	3*
18	Low back conditions	3	Diagnosis	No	Yes[1]	6
Total panel item responses removed / %			67 / 5%			

[1] Clinical practice guideline, [2] Evidence statement, * Limited item optimalization because of content-specific considerations

Table — 1b Script Concordance vignette level 1 including 1 scenario and test item

Personal information

Jannie Hermsen: 25 years. student occupational therapy.

Hobbies / sports: crafts. playing volleyball

Basic Information

Jannie suffers since 2 months from increasing pain in her left knee. She can't remember the onset moment. At rest, the symptoms are nagging, especially when sitting in a deep chair. The symptoms are bothersome while playing volley and she is afraid to fall on her left knee. Furthermore, all activities where she needs to squat are painful and she noticed that she has difficulty to come up from a squatting position.

Scenario A

Bea has the following hypothesis: a patellofemoral pain syndrome

Additional information

Watch the video clip of the inspection of the knee.

This information makes the hypothesis:

A	Unlikely	O
B	Less likely	O
C	No more or no less likely	O
D	More likely	O
E	Very likely	O

Table 1c — Scoring algorithm script concordance test

Alternative	Panel Score	Removed	Transformed Score
A	1	1	0 %
B	0		0 %
C	2		2/14 x 100 = 14%
D	14		100 %
E	5		5/14 x 100 = 36%

Chapter 9
General discussion

In this thesis we explored feedback interventions, developed and conducted by physical therapists aiming to support continuous improvement of professional performance. All feedback interventions described in this thesis have in common that the physical therapist is both provider and receiver of performance feedback. The interventions took part of a more comprehensive quality improvement plan (Masterplan Quality in Movement, MKIB), initiated and supported by the Royal Dutch Society for Physical Therapy (KNGF). The MKIB involves the development of a quality system that aims to self-regulate the quality of physical therapy services, and includes assessment of clinical performance (self- and peer assessment) and organizational performance (clinical audit). The outcomes of these studies served as input for continuous improvement of the quality system design and its implementation strategy. In this final chapter we will return to the main research questions:

1 How do physical therapists perceive interventions, based on performance feedback, aiming to advance the quality of physical therapy care?
2 What is the impact of interventions, based on performance feedback, on learning and professional behavior change?

Accordingly, we will critically reflect on the process of program development and implementation to inform the design of a sustainable quality improvement system in physical therapy. We will end with a set of recommendations for practitioners, program developers, and policy makers. Finally, an overall conclusion is provided.

Findings related to the main questions

Four performance feedback interventions with peer assessment, two interventions with self-assessment, peer assessment and practice visitation (clinical audit), and one intervention with the script concordance test are evaluated.

Perceptions of the feedback interventions

In this paragraph we will discuss the general perceptions of the performance feedback interventions involving peers as feedback providers and receivers. Taking the evaluation results together we can conclude that all feedback interventions using role-play or

video-recordings of real-life encounters were perceived as useful quality improvement strategies targeting the core-business of physical therapists. Nevertheless, we consciously shifted from using role-play (chapters 2-5) to video-recordings (chapters 6-7) and introduced self-assessment as an explicit program element in chapters 6-7, whereas self-assessment in chapters 2-5 was considered as an implicit response to peer feedback. We will discuss this issue later in the section on program development and implementation.

Regarding the performance in the physical therapist role, participants encountered difficulty in exposing their professional performance to the critical review of their peers. They perceived performance stress, irrespective the performance format, role-play (chapters 2-5) or video-recordings of real-life client encounters (chapters 6-7).[1] The majority of participants succeeded in coping with these stress-triggers and recognized the added value of 'performance exposure' as receiving individualized performance feedback is scarce in both undergraduate education and professional practice.[2,3] Providing feedback in the assessor or auditor role was perceived constructive, however difficult. Participants needed to become familiar with the quality concepts 'client-centeredness', and 'effectiveness' to fully understand the performance indicators and needed training in providing constructive feedback.[4] They were willing to provide narrative feedback, but reluctant in rating their peers, resulting in lenient marking (i.e. too high).[5] These perceptions are extensively reported by the literature on peer assessment.[6-11]

Views on the usability of role-play or video-recordings as learning materials differed. From the performer role perspective, role-play and video-recordings did not always reflect their day-to-day practice as these behaviors were distorted by the 'audience effect'. They viewed the role-play or video-recording as a 'testimony of their professional competence' and felt uncomfortable when they could not fully identify with it. From the assessor role perspective, these materials provided unique learning experiences, irrespective of the audience effect.[12] By 'watching what their peers do and by hearing what they say' they can compare the observed behavior with their own behavior. Learning activities such as case discussion, for example, do not offer this opportunity.

Surprisingly, the results of students (chapter 2) and professionals (chapter 4) on the ranking procedure of learning tasks according to their perceived learning value, point at the superior value of performing the physical therapist role; considered even more powerful than observing the professional behavior of a peer. This apparent contradiction shows that exposing professional behavior for peer

review, and coping with anxiety triggers, is a necessary sacrifice that peers need to make to allow for meaningful group learning experiences.

Impact on learning and professional behavior change

This section describes the outcomes of the various interventions evaluated in this thesis in terms of learning and professional behavior change. A distinction will be made between tested outcomes and self-reported outcomes.

Tested outcomes

Beginning with the tested outcomes, interventions with peer assessment targeting the effectiveness of physical therapy show that peer assessment leads to an increase of knowledge and evidence-based reasoning, and is more effective than case based discussion as implementation strategy for clinical practice guidelines (chapters 3 and 5). Moreover, the peer assessment strategy is more effective in raising awareness of professional performance and attaining personal goals (chapter 5). A study by Meerhoff *et al*.[13] using peer assessment to implement patient reported outcome measurement supports these findings. Interventions with both self- and peer assessment enhance commitments to professional behavior change and improve clinical performance as shown in chapters 6-7. Regarding the two trials described in chapters 3 and 5 comparing, the effectiveness of peer assessment with case-based discussion, the outcome measures used were scores on clinical vignettes (cases). Although the intervention group outperformed the control group in both studies, the second study might better reflect the true outcomes. First, the number of communities of practice (CoPs) and participants in the first trial described in chapter 3 (n=90; CoPs = 10) was substantially smaller than in the second trial described in chapter 5 (n=149; CoPs=20). The larger sample may have reduced bias in the results. Second, in the latter trial, participants in the intervention – and control group were provided with the model answers to all the clinical cases discussed in the program before the final test, to compensate for unwanted differences within and between groups due to the influence of the coach. Thus, knowledge of the existing evidence regarding diagnosis and treatment of the cases discussed was aligned between the two groups. In the final test new cases were presented, requiring a transfer of knowledge and evidence-based reasoning to new clinical problems and that

may better reflect the true intervention effect.[14] Third, the response format on clinical vignettes in the second trial differed from the first. In the first trial, single best answers were used as outcome measure, whereas in the second trial a script concordance model was used allowing for variation in best answers. The latter method used 'aggregate scoring' (the match between a participant's response with a group of experts). Research showed superior psychometric properties of the 'aggregate scoring method' compared to other scoring methods including the single best answer method.[15] Furthermore, in both studies the pre- and post-test results show considerable variation in outcomes among physical therapists, showing that there is still much room for improvement for low performers regarding evidence-based reasoning. This finding is confirmed by other studies on guideline adherence among physical therapists.[16-19] Given the fact that both interventions were short-term, prolonged engagement might be more effective for low performing professionals.

In the studies described in chapters 6 and 7, quality indicators were used to support the feedback process. The results show that quality indicators were effective in steering the quality improvement process towards the intended competency domains: client-centered communication and record keeping. For example, 54 % of the improvement goals focused on client-centered communication including goal setting and shared decision making. That raises the question if awareness of performance gaps in client-centered communication would have been identified without performance indicators. In the latter case, participants might have focused on other aspects of the observed performance, missing the quality aims of the program. We assume that performance indicators may have a crucial role in developing awareness of performance standards and are powerful tools to trigger professional behavior change into the desired direction. We therefore argue in favor of continuous indicator development and validation for varying competency domains. Whether 'scoring' performance indicators is necessary to raise quality awareness remains to be seen; using indicators as a narrative feedback support might do as well.

Self-reported outcomes

By combining the results of the studies addressing self-reported outcomes on learning and behavior change (chapters 2,4, and 6), some similarities can be identified. In table 1 the main results are presented.

Table 1 — Self-reported impact on learning and behavior change

Learning processes triggered by peer assessment	
Implicit learning	Coping with performance stress
	Mirroring observed performance
	Modeling professional roles
Explicit learning	Reasoning aloud
	Critical performance appraisal
	Discussion of performance standards and quality indicators
Learning outcomes	
Attitude change	Improved attitudes towards evidence-based practice
	Improved attitudes towards client-centered care
Knowledge and skills	Knowledge of performance standards
	Knowledge of new clinical reasoning perspectives and strategies
Behavior change	Awareness of strengths and weaknesses in professional performance
	Improved self-efficacy beliefs

Reflecting on the self-reported outcomes on learning and behavior change, some critical features of peer assessment can be identified. The first reported trigger for learning refers to 'showing what you do', challenging the performer to critically reflect on his personal performance and allowing the observers to see what usually happens behind closed doors. Since the discovery of mirror neurons in the brain, research showed that this form of learning is intuitive, causes little cognitive burden, and is very effective.[20,21] In addition, exposure of professional behavior was perceived as challenging, but resulted in increased self-efficacy beliefs. The literature shows that self-efficacy beliefs are conditional to the intrinsic motivation to learn[22,23] and to behavior change.[24,25] The second apparent trigger for learning refers to 'showing what you think' known as reasoning aloud, challenging the speaker to explicit reasoning that has become implicit by experience, providing the listener of access to mental models and reasoning strategies that are created in a usually inaccessible brain.[26,27]

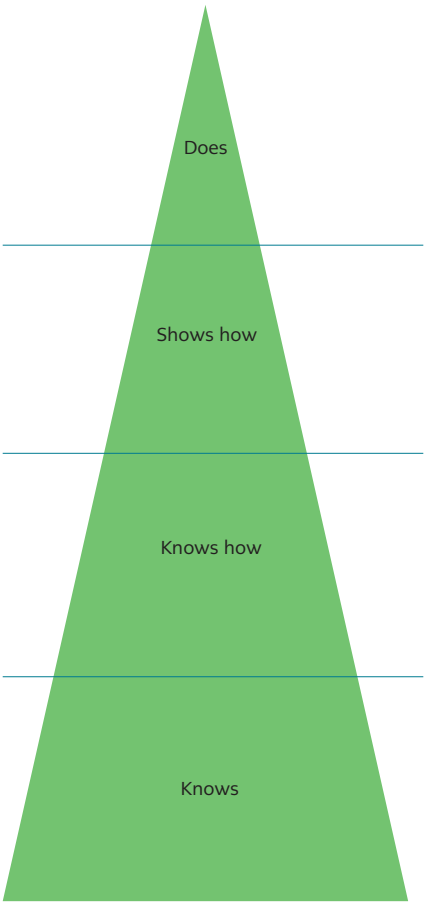
What peer assessment distinguishes from other performance assessment methods, is the combination of cognitive, emotional, and social involvement in learning. Cognitive involvement refers to solving clinical problems that apply to their daily practice.[28–30] Social involvement refers to mirroring and modeling observed peer behaviors,[21,24] sharing knowledge and reasoning perspectives, and

providing each other tips for improvement.[31] Emotional involvement – possibly the most powerful feature - refers to physical therapists exposing professional behaviors to provide their peers access to the confidential area of their clinical practice, their personal modes of reasoning, styles of communication, views on illness and health which touches their professional beliefs, identity, and mission allowing for deep learning and reflection.[32,33] Research showed that emotion has a powerful impact on memory and the transfer of learning.[34-36]

Critical reflection on program development and implementation

When physical therapists intend to self-regulate the quality of their services, they need valid performance assessment information and usable feedback for self-directed quality improvement. The results of a systematic review of Overheem *et al.*,[37] supported by the literature on clinical competency assessment,[38-40] shows that there is no single best method to assess clinical performance, and that each instrument has its advantages and disadvantages. To explain the steps taken in program development and implementation, we will use the competency assessment framework of Miller *et al*.[41]
Every performance assessment design contains a 'stimulus format' and a 'response format'. The stimulus format refers to the assessment task such as completing written clinical vignettes, simulations of clinical encounters in a role-play, authentic clinical records, or real-life clinical encounters. The response format refers to the way the answer or judgment is captured such as multiple choice, checklist, global rating forms, verbal and/or written feedback. Miller's framework classifies competency assessment methods and is helpful in explaining our choices for the various assessment methods.[41] It is presented in Figure 1, showing four competency layers and the assessment methods used in this thesis that correspond to these layers. The 'knows' level stands for the assessment of factual knowledge which is not explicitly addressed in this thesis, though implicitly in assessing knowledge of clinical practice guidelines. The 'knows how' level refers to the appliance of knowledge requiring higher order cognitive skills such as clinical reasoning and decision making skills in a standardized stimulus and response format. When it comes to the 'shows how' level, professional performance is assessed in a controlled context by observation, such as simulations or role-play. Moving to the final 'does' level, professional performance is assessed in the working context by observing artefacts of profes-

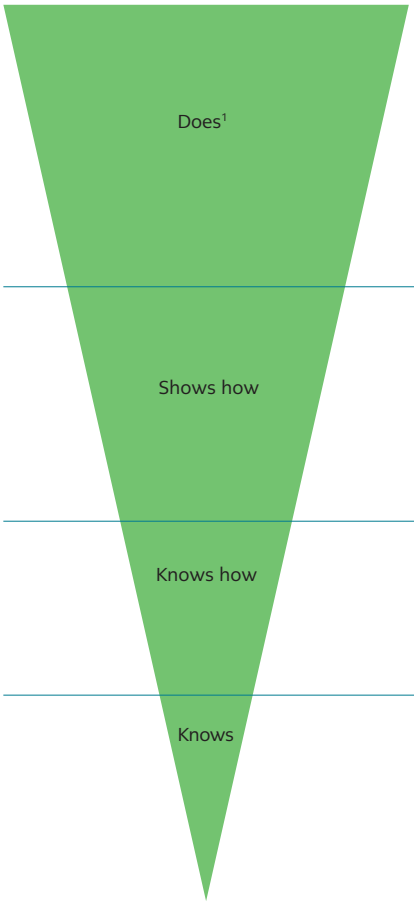
Figure 1 — The assessment of authentic professional behaviors



	Assessment format	Behavior assessed
Does	*Stimulus format:* non-standardized. Observation of self-selected video recordings of real-life client encounters and client records. Clinical audit *Response format:* global rating form. Quantitative and qualitative personalized feedback.	Clinical reasoning Record keeping Client communication Practice organization and management (chapter 6-7)
Shows how	*Stimulus format:* semi-standardized. Role-play based on pre-defined written scripts. *Response format:* global rating form. Quantitative and qualitative personalized feedback.	Clinical reasoning Clinical examination and treatment (chapter 2-5).
Knows how	*Stimulus format:* standardized. Computer-based test based on pre-defined written clinical scripts. *Response format:* script concordance model. Standardized feedback	Clinical reasoning (chapter 8)
Knows		Factual knowledge

sional performance (e.g. client records or video-recordings) or direct observation. The first three layers of Miller's pyramid are about standardized (level 1-2) or semi-standardized (level 3) assessment, the fourth layer is on non-standardized assessment. We continuously improved the feedback interventions, informed by participants' perceptions, the intended areas for improvement, and supported by the literature on quality improvement interventions.[42-44] Figure 1 shows that the chapters in this thesis stepwise climb the Miller's pyramid approaching as much as possible the assessment of authentic professional behaviors. However, assessment of authentic behaviors

Figure 2 — The educational arangements to strenghten the peer feedback process



	Feedback delivery	Feedback acceptance	Feedback use
Does[1]	Training of knowledge brokers to facilitate program implementation. Training of group coaches to support the feedback process.	Involving personal views (self-assessment) – and peer views (peer assessment) in performance feedback. Enhancing dialogue between feedback provider (peer assessor) and receiver (self-assessor).	The feedback recipient summarizes the feedback obtained, indicates which feedback was perceived the most impressive and how this feedback is translated into improvement actions.
Shows how	Performance indicators to enhance knowledge of performance standards and to guide feedback delivery towards its intended goals. Guidelines for providing constructive feedback.	Involving multiple peer views in performance feedback (peer assessment). Discussing feedback in peer groups. Guidelines for receiving and processing feedback.	Formulating and prioritizing improvement goals.
Knows how	Reference to online knowledge resources. Standardized quantitative feedback.	Involving multiple expert views in performance feedback. Discussion of the results for interested participants.	
Knows			

[1] Arrangements applying to the 'shows how' level also apply to the 'does' level but not vice-versa.

comes with new challenges regarding the feedback intervention design and its implementation.

Figure 2 shows that educational arrangements on the 'shows how' level to strengthen the peer feedback process, were limited to performance indicators and feedback rules. For the 'does' level, new arrangements were needed to strengthen the feedback process (chapters 6-7). Online self-assessment was introduced as an integral program element to create more transparency in differences between personal – and peer views, serving as input for face-to-face dialogues between feedback provider and feedback recipient.

This program adjustment allowed to tailor feedback to recipient's stages of change from the feedback provider perspective and weigh and prioritize feedback from the receiver perspective.[33,45] We also learned that the role of the coach needed empowerment. When participants provide video-recordings of real client encounters and authentic client records, they become extremely vulnerable to a lack of perceived group safety and are strongly emotionally involved with the feedback process. They tend to avoid arguing aloud for the choices they have made, anxious to make mistakes, resulting in less inspiring group sessions and limited group learning. The coach's role was strengthened by providing a training program that addressed: 1) building trust among group members and fostering a safe learning environment in which learning from errors or misconceptions is the primary aim, 2) facilitating reasoning aloud by posing triggering questions if necessary, 3) building group cohesion to facilitate shared responsibility for the group process and the learning outcomes.

In short, interventions using peer assessment of authentic professional behaviors touch the heart of physical therapy, but the price of extra effort and time needs to be paid to foster desired outcomes.

Besides extra time and effort, peer assessment on the 'does' level has more disadvantages. First, the assessment materials brought in by peers provide a poor case-mix. Given the notion that clinical reasoning is both context- and content-specific,[14,46,47] implying that adequate reasoning in one case does not always predict adequate reasoning in another case, the transfer of new knowledge and reasoning strategies can be hampered by too much difference between cases discussed in the peer group and cases encountered in clinical practice, known as 'the transfer gap.'[48-50] Thus, it remains to be seen if peer groups succeed in applying new insights to the context of their own practice. Second, the peer group determines the performance level. The coach cannot be expected to represent the golden standard, and if so, the group would lean back immediately, drop their assessor glasses, and pass them back to the coach, freeing themselves of being responsible for the learning outcomes. As a consequence, poorly performing peer groups, supported by poorly educated coaches, may strengthen each other in maintaining low performing standards, causing unwanted variety in physical therapy care.[51,52] Third, the learning value of the materials depends on the ability, the courage, and the willingness of peers to present cases that bring about dilemmas in decision making. For example – the case of Eric, discussed in the introduction of this thesis – where

no intervention seemed to work – would appeal to their problem solving skills. When peers remain in their comfort zone regarding the cases they bring in, when they don't expose their uncertainties and true learning needs, the group may get bored, losing intrinsic motivation to participate. Positive appraisal of participants who provide challenging cases, might encourage others to do likewise. Comparing these considerations to the script concordance test (SCT) – a standardized written test – it needs no explanation that the SCT is lacking social and emotional involvement with the assessment task. However, a number of advantages of the SCT can be identified. First, although the SCT development is time consuming, the implementation is simple. Testing and receiving feedback is realized in an e-learning environment. Of course, additional feedback sessions take time, but do not necessarily apply to all test takers. Second, the SCT allows for broad sampling of clinical cases varying in content and complexity, facilitating the transfer of learning to clinical practice.[53,54] Third, the SCT answering key is developed by a panel of experts that together can be considered to represent the golden standard, allowing participants to benchmark their results to both the expert panel and their peers. In the following paragraph we will address the implications of these considerations.

Implications for policy makers and program designers

To date, the integrated quality system of peer assessment and clinical audit is part of the quality policy of professional organizations of physical therapists and a step-wise implementation is realized. The quality system is expected to be fully implemented in the short term as peer assessment and clinical audit is advocated by the Ministry of Health.[55,56] We have trained a considerable number of knowledge brokers for implementing the quality system, train-the-trainers to educate peer group coaches, and clinical auditors to conduct the audits. Although the implementation of the system opens doors to the self-regulation of the quality of physical therapy services, the process of implementation requires careful monitoring and the critical success features should be consciously addressed, in particular the 'alignment of expectancies on program aims, outcomes, and consequences'. When participants become afraid of the consequences of 'telling what they think' and 'showing what they do', self-regulation by peer assessment and clinical audit will be a dead end. In addition, when external authorities delegate the responsibility for developing and maintaining performance standards

to the professionals – which may be considered as a privilege – that does not automatically imply that professionals are capable of doing so. Professionals need time to develop the necessary knowledge, skills, and attitudes to be trusted with self-regulation. Moreover, the system needs monitoring to allow for continuous improvement of the critical system features and to optimize the intended outcomes. This implies that external authority-based incentives to speed up the implementation process, might discourage physical therapists to fully commit.[22,23] Conversely, the development of communities of practice and professional networks, supporting individuals in being accountable for their services as a group, might both strengthen the self-regulation system and facilitate the implementation.[23,57,58]

Based on the findings of the studies in this thesis, and informed by the involvement of the research team in the implementation of peer assessment and clinical audit in professional practice, we present the following recommendations for sustainable self-regulation:

- Integrate peer assessment in the curricula of undergraduate physical therapy education – in preparing students to take the assessor perspective on clinical performance, and to be accountable as a professional and as a group of professionals.[59]
- Continue program development tailored to changing learning needs and preferences.
- Continue the development of quality indicators – that includes the client perspective – to help physical therapists to identify blind spots in the quality of their performance and to anticipate on future quality demands of clients and other stakeholders in healthcare.
- Continue the ongoing selection and training of peer assessment coaches to support the feedback process and raise performance standards where needed.[56]
- Strengthen the development of communities of practice and professional networks.
- Monitor the process and outcomes of peer assessment and clinical audit to allow for continuous improvement of the quality system and to account for the outcomes.
- Complement the online assessment system for peer assessment and clinical audit[60] with an e-learning environment containing new performance assessment designs, allowing individual professionals or groups of professionals, to self-assess their professional competence in varying content domains and varying professional behaviors, and to compare their performance to a benchmark. These feedback interventions may contribute to reducing unwanted variation in physical therapy care.

– Develop a quality register based on a portfolio that provides evidence of continuous improvement based on a variety of performance assessment methods including peer assessment.

Implications for physical therapists

The outcomes of the studies involved in this thesis show that physical therapists are motivated to self-regulate the quality of their services, and that peer assessment and clinical audit are effective in identifying areas for improvement. Meanwhile professionals need to take the consequences. The outcomes of the various feedback interventions show room for improvement regarding clinical reasoning, performance- and outcome measurement, client-communication, and record keeping. Particularly in the area of client communication is much to learn as one of the participants strikingly stated: *"…my record keeping was all right, but regarding patient communication… I explained a lot, but I didn't check to see if my message was understood. That's an improvement I need to make. I try now to ask my patient: 'What did you learn about what I explained just now?' Moreover, I pay more attention to their personal goals. I can have a plan, but that plan might not be in line with their expectations… I might be too dominant in this respect because I think that I know what they need, but I should not think for them."*
Based on the program outcomes and informed by the involvement of the research team in the training of coaches and auditors, the following areas for improvement are recommended:

Effectiveness of peer assessment

The learning value of peer assessment and the sustainability of the outcomes can be improved by supporting feedback recipients in clarifying their specific learning needs and feedback providers to tailor their feedback accordingly. In addition, participants should be encouraged to present dilemma's in clinical reasoning and decision making to challenge the problem solving skills of their peers.

Qualitiy of physical therapy care

Table 1 shows that peer assessment has a positive impact on attitudes towards clinical practice guidelines. We – as a research team – observed that cognitions and beliefs related to evidence-based reasoning need adjustment. The misperception that evidence-based reasoning is solely based on evidence as information resource is still a prevailing view, whereas evidence based practice needs a

holistic instead of a reductionist approach to problem solving by integrating different sources of knowledge related to the client perspective, the professional perspective, and the scientific perspective.[61-63]
Regarding client-centered care there is substantial room for the improvement of communication skills, in particular aligning mutual expectancies, involving clients in goal setting and defining outcomes in terms of what is meaningful to the client, and empowering clients in self-managing their health problems. The 'instructing' therapist and the 'nodding' client is still common practice.

Strengths and Limitations

The studies in this thesis addressed the development and evaluation of quality improvement interventions. We used existing theory for intervention design, and we involved end-users to tailor the programs to the application context and to facilitate program implementation.[43,44] This resulted in a sound design for the quality improvement interventions and commitment of participating physical therapists. Both qualitative and quantitative methods were applied, to evaluate perceptions of the intervention and the impact on quality improvement as advocated by the literature.[42,64] The integration of qualitative and quantitative methods allowed for evaluating the effects and explore the underlying mechanisms to explain the effects.[42]
A limitation is that all the interventions evaluated were short-term and the results on the long-term remain unclear. Moreover – although we have demonstrated effects on professional behavior change – we have not demonstrated effects on client experiences or -outcomes.
The feedback interventions described in chapters 3 and 5 were preceded by a standardized test based on completing clinical vignettes. The feedback interventions described in chapters 6-7 were preceded by an unstandardized test based on scoring performance indicators. In both cases, it remains unclear what the proportion of the pre-test effect is on the overall intervention effect.
Another limitation concerns the use of performance indicators for client communication and record keeping (chapters 6-7). The performance indicators were informed by the literature and qualitatively validated by different stakeholders in program development. This procedure is defensible when the outcomes are used for quality improvement purposes. When the outcomes are used

for summative decisions, more rigorous validation procedures are desired.
Finally, we need to address the influence of the peer assessment coach on the results. Although coaching is necessary for untrained peer assessment groups, we do not know how the group coach affected the intervention results and how groups would perform without the presence of the coach.

Recommendations for future research

Future research should address the sustainability of the effects of the feedback interventions on professional and organizational performance. In addition, research into the effects on client outcomes should be part of the research agenda of authorities involved with the quality of physical therapy. Looking at sustainable implementation of peer assessment, it would be interesting to know whether peer groups finally succeed in self-regulating their quality improvement process including the barriers and facilitators for successful peer group functioning. Self-regulation also depends on choosing the right quality improvement interventions. Developing, implementing, and evaluating innovative feedback interventions is therefore a challenge for future research.

Conclusion

Peer assessment and clinical audit might be promising tools to support self-regulation and professional accountability.
Interventions with peer assessment are perceived to provide useful feedback for quality improvement. Exposing professional behaviors was perceived challenging and sometimes stressful but participants succeeded in coping with performance anxiety. Taking the assessor perspective and providing performance feedback was perceived difficult and required training. The interventions resulted in two major conditions allowing for learning and professional behavior change: increased awareness of performance, and self-efficacy beliefs. However, the feedback process needs coaching to maintain psychological safety and to enhance feedback delivery, feedback acceptance, and feedback use; the coaching of peer groups needs continuing training. The perceived value of peer assessment increases with the use of authentic materials derived from daily physical therapy practice such as video recordings of real-time behaviors and client records.

However, the implementation of these interventions calls for more attention to implementation barriers as well as to coaching the feedback process. Moreover, the learning value depends on the ability and the willingness of peers to present cases that bring about dilemmas in decision making to keep the peer group motivated. The outcomes of the interventions show that there is considerable room for improvement regarding evidence-based clinical reasoning, client-centered communication, and performance and outcome measurement. The use of performance indicators was effective and can be considered as an adequate strategy to respond to current and anticipated future challenges for the quality of healthcare. Feedback interventions with self- and peer assessment have – similar to all performance assessment methods – advantages and disadvantages and cannot be considered as the holy grail. A cocktail of assessment methods – including for example the script concordance test – is therefore desired to adequately self-regulate the quality of physical therapy care.

References

1 Pope NKL. The impact of stress in self- and peer assessment. *Assess Eval High Educ.* 2010;30:37-41.

2 Sargeant J, Bruce D, Campbell CM. Practicing physicians' needs for assessment and feedback as part of professional development. *J Contin Educ Health Prof.* 2013;33(1):54-62..

3 Watling C, Driessen E, van der Vleuten CPM, Vanstone M, Lingard L. Music lessons: revealing medicine's learning culture through a comparison with that of music. *Med Educ.* 2013;47(8):842-850.

4 Sluijsmans DMA, Van Merriënboer JJG, Brand-gruwel S, Bastiaens TJ. The training of peer assessment skills to promote the development of reflection skills in teacher education. *Stud Educ Eval.* 2003;29(1):23-42.

5 Topping KJ. The effectiveness of peer tutoring in further and higher education: A typology and review of the literature. *High Educ.* 1996;32:321-345.

6 Topping KJ. Methodological quandaries in studying process and outcomes in peer assessment. *Learn Instr.* 2010;20(4):339-343.

7 Dannefer EF, Henson LC, Bierer SB, *et al.* Peer assessment of professional competence. *Med Educ.* 2005;39(7):713-722.

8 Falchikov N. *Improving Assessment through Student Involvement: Practical Solutions for Aiding Learning in Higher and Further Education.* 2nd ed. New York: Routledge Falmer; 2013.

9 Speyer R, Pilz W, Van Der Kruis J, Brunings JW. Reliability and validity of student peer assessment in medical education: a systematic review. *Med Teach.* 2011;33(11):572-585.

10 Finn GM, Garner J. Twelve tips for implementing a successful peer assessment. *Med Teach.* 2011;33(6):443-446.

11 Liu N-F, Carless D. Peer feedback: the learning element of peer assessment. *Teach High Educ.* 2006;11(3):279-290.

12 Maas MJM, van Dulmen SA, Sagasser MH, *et al*. Critical features of peer assessment of clinical performance to enhance adherence to a low back pain guideline for physical therapists: a mixed methods design. *BMC Med Educ.* 2015;15(1):203.

13 Meerhoff GA, van Dulmen SA, Maas MJ, Heijblom K, Nijhuis-van der Sanden MW, van der Wees PJ. Development and evaluation of an implementation strategy for collecting data in a national registry and the use of patient-reported outcome measures (PROMS) in physical therapist practice: quality improvement study. *Phys Ther.* 2017;Published ahead of print.

14 Norman G, Bordage G, Page G, Keane D. How specific is case specificity? *Med Educ.* 2006;40(7):618-623.

15 Goos M, Schubach F, Seifert G, Boeker M. Validation of undergraduate medical student script concordance test (SCT) scores on the clinical assessment of the acute abdomen. *BMC Surg.* 2016;16(1):57.

16 Swinkels ICS, van den Ende CHM, van den Bosch W, Dekker J, Wimmers RH. Physiotherapy management of low back pain: Does practice match the Dutch guidelines? *Aust J Physiother.* 2005;51(1):35-41.

17 Rutten GM, Harting J, Bartholomew LK, Schlief A, Oostendorp R a B, de Vries NK. Evaluation of the theory-based Quality Improvement in Physical Therapy (QUIP) programme: a one-group, pre-test post-test pilot study. *BMC Health Serv Res.* 2013;13(1):194.

18 Rutten GMJ, Harting J, Rutten STJ, Bekkering GE, Kremers SPJ. Measuring physiotherapists' guideline adherence by means of clinical vignettes: a validation study. *J Eval Clin Pract.* 2006;12(5):491-500.

19 Rutten G., Degen S, Hendriks E, Braspenning J, Harting J, Oostendorp R. Adherence to Clinical Practice Guidelines for Low Back Pain in Physical Therapy: Do Patients Benefit? *Phys Ther.* 2010;90(8):1111-1121.

20 Paas F, Sweller J. An Evolutionary Upgrade of Cognitive Load Theory: Using the Human Motor System and Collaboration to Support the Learning of Complex Cognitive Tasks. *Educ Psychol Rev.* 2012;24(1):27-45.

21 Iacoboni M. Mirroring *People: The New Science of How We Connect with Others*. 2nd ed. (Farrar S and G, ed.). New York: Macmillan; 2009.
22 ten Cate OTJ, Kusurkar RA, Williams GC. How self-determination theory can assist our understanding of the teaching and learning processes in medical education. *Med Teach*. 2011;33(12):961-973.
23 Ryan R, Deci E. Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *Am Psychol*. 2000;55(1):68-78.
24 Bandura A. *Self-Efficacy: The Exercise of Control*. Vol 50. (Anonymous, ed.). Freeman; 1997.
25 Ajzen I. Nature and operation of attitudes. *Annu Rev Psychol*. 2001;52:27-58.
26 Durning S, Artino AR, Pangaro L, van der Vleuten CPM, Schuwirth L. Context and clinical reasoning: Understanding the perspective of the expert's voice. *Med Educ*. 2011;45(9):927-938.
27 Higgs J, Jones MA, Loftus S, Christensen N. *Clinical Reasoning in the Health Professions*. Third edit. Philadelphia: Elsevier Health Sciences; 2008.
28 Dolmans DHJM, De Grave W, Wolfhagen IHAP, van der Vleuten CPM. Problem-based learning: future challenges for educational practice and research. *Med Educ*. 2005;39(7):732-741.
29 Schmidt HG, Rotgans JI, Yew EHJ. The process of problem-based learning: What works and why. *Med Educ*. 2011;45(8):792-806.
30 Biggs J. What the student does : teaching for enhanced learning. *High Educ Res Dev*. 2006;18(1):57-75.
31 Van Gennip NA, Seger MS, Tillema HH. Peer assessment as a collaborative learning activity: the role of interpersonal variables and conceptions. *Learn Instr*. 2010;20(4):280-290.
32 Korthagen F, Vasalos A. Levels in reflection: Core reflection as a means to enhance professional growth. *Teach Teach Theory Pract*. 2005;11(1):47-71.
33 Sargeant JM, Lockyer J, Mann K, *et al*. Facilitated reflective performance feedback: developing an evidence- and theory-based model that builds relationship, explores reactions and content, and coaches for performance change (R2C2). *Acad Med*. 2015;90(12):1698-1706.
34 McConnell MM, Eva KW. The role of emotion in the learning and transfer of clinical skills and knowledge. *Acad Med*. 2012;87(10):1316-1322.
35 Roediger HL, Karpicke JD. Test-enhanced learning: taking memory tests improves long-term retention. *Psychol Sci*. 2006;17(3):249-255.
36 Artino AR, Holmboe ES, Durning SJ. Control-value theory: using achievement emotions to improve understanding of motivation, learning, and performance in medical education: AMEE Guide No. 64. *Med Teach*. 2012;34(3):e148-60.
37 Overheem K, Faber MJ, Onyebuchi AA, *et al*. Doctor performance assessment development in daily practise: does it help doctors or not? A systematic review. *Med Educ*. 2007;41(11):1039-1049.
38 van der Vleuten CPM, Sluijsmans DM, Joosten-ten Brinke D. Competence assessment as learner support in education. In: Mulder M, ed. *Competence-Based Vacational and Professional Education*. 1st ed. Springer International Publishing AG; 2017:607-630.
39 van der Vleuten CPM, Schuwirth LWT, Driessen EW, *et al*. A model for programmatic assessment fit for purpose. Med Teach. 2012;34(3):205-214
40 van der Vleuten CPM, Schuwirth LWT. Assessing professional competence: From methods to programmes. *Med Educ*. 2005;39(3):309-317.
41 Miller GE. The assessment of clinical skills/competence/performance. *Acad. Med*. 1990;65(9):63-67.
42 Craig P, Dieppe P, Macintyre S, *et al*. Developing and evaluating complex interventions: new guidance. *BMJ*. 2008;337:a1655.
43 Grol RP, Wensing M, Eccles MP, Davis DA, (Eds). I*mproving Patient Care: The Implementation of Change in Health Care*. 2nd ed. Chichester, West Sussex: John Wiley & Sons, Inc.; 2013.
44 Brehaut JC, Eva KW. Building theories of knowledge translation interventions: use the entire menu of constructs. *Implement Sci*. 2012;7(114):1-14.

45 Sargeant J, Eva KW, Armson H, *et al*. Features of assessment learners use to make informed self-assessments of clinical performance. *Med Educ.* 2011;45(6):636-647. doi:10.1111/j.1365-2923.2010.03888.x.
46 Norman G. Research in clinical reasoning: Past history and current trends. *Med Educ.* 2005;39(4):418-427.
47 Durning SJ. Exploring the Influence of Contextual Factors of the Clinical Encounter on Clinical Reasoning Success (Unraveling context specificity). *Acad Med.* 2010;85(5).
48 Eraut M. Transfer of knowledge between education and workplace settings. *Knowledge, Values Eudcational Policy A Crit Perspect.* 2009;65:1-17.
49 Burke L, Hutchins H. Training transfer: an integrative literature review. *Hum Resour Dev Rev*. 2007;6(3):263-296. http://hrd.sagepub.com/cgi/doi/10.1177/1534484307303035. Accessed January 26, 2014.
50 Simons P. Transfer of learning: paradoxes for learners. *Int J Educ Res.* 1999;31(7):577-589.
51 Institute of Medicine. *Crossing the Quality Chasm: A New Health System for the 21st Century.* Washington, DC: National Academy Press; 2001.
52 Porter ME. What is value in health care? *N Engl J Med.* 2010;363(26):2477-2481.
53 Lubarsky S, Charlin B, Cook DA, Chalk C, van der Vleuten CPM. Script concordance testing: a review of published validity evidence. *Med Educ.* 2011;45(4):329-338.
54 Fournier JP, Demeester A, Charlin B. Script concordance tests: guidelines for construction. *BMC Med Inform Decis Mak.* 2008;8:18.
55 Schippers E. Speech van de minister van vws, Edith Schippers, bij het Jaarcongres van het Koninklijk Nederlands Genootschap Fysiotherapie op de Dag van de Fysiotherapeut in Utrecht. 2015. https://www.rijksoverheid.nl/regering/inhoud/bewindspersonen/edith-schippers/documenten/toespraken/2015/11/06/speech-van-de-minister-van-vws-edith-schippers-bij-het-jaarcongres-van-het-koninklijk-nederlands-genootschap-fysiotherapie-op-de-dag-van-de-fysio.
56 Zorginstituut Nederland. Systeemadvies fysio- en oefentherapie; een nieuwe balans tussen de toegang tot en de betaalbaarheid van goede zorg. https://www.rijksoverheid.nl/regering/inhoud/bewindspersonen/edith-schippers/documenten/rapporten/2016/12/20/systeemadvies-fysio-en-oefentherapie.
57 le May A. Introducing communities of practice. In: le May A, ed. *Communities of Practice in Health and Social Care*. Oxford: Wiley-Blackwell; 2008:3-16.
58 Li LC, Grimshaw JM, Nielsen C, Judd M, Coyte PC, Graham ID. Use of communities of practice in business and health care sectors: a systematic review. *Implement Sci.* 2009;4:27. doi:10.1186/1748-5908-4-27.
59 Dall'Alba G. Learning professional ways of being: Ambiguities of becoming. *Educ Philos Theory.* 2009;41(1):34-45.
60 Compusense Business Avionics. www.compusense.nl.
61 Ajjawi R, Higgs J. Learning to reason: a journey of professional socialisation. *Adv Health Sci Educ Theory Pract.* 2008;13(2):133-150.
62 Dannapfel P, Peolsson A, Nilsen P. What supports physiotherapists' use of research in clinical practice? A qualitative study in Sweden. *Implement Sci.* 2013;8:31.
63 Gabbay J, May A. *Practice-Based Evidence for Healthcare: Clinical Mindlines.* London: Routledge; 2011.
64 Øvretveit J. *Evaluating Improvement and Implementation for Health.* 1st ed. New York: McGraw-Hill Education; 2014.

Chapter 10
Summary
Samenvatting
Dankwoord
PhD portfolio

Summary

Chapter 1

People who seek the help of a physical therapist deserve the best possible care[1] for their health problem. The best possible care is provided by up-to-date trained professionals who can flexibly respond to changing client needs and increasing political and social demand for value-based care. The challenges for physical therapists are related to the effectiveness, client-centeredness and transparency of the process and outcomes of their services.
To continuously work on quality improvement, physical therapists – both undergraduate and post-graduate – need criteria for the best possible care, and useful feedback on the extent to which they individually and collectively meet these criteria. The chapters in this thesis describe feedback interventions based on performance assessment, developed by – and for physical therapists to support professionals, teams, and organizations in self-regulating and accounting for the quality of their services.
This dissertation addresses the following research questions:
How do physical therapists perceive interventions, based on performance feedback, aiming to advance the quality of physical therapy care?
What is the impact of interventions, based on performance feedback, on learning and professional behavior change?

Chapter 2

Chapter 2 describes the results of a mixed-methods study evaluating the impact of peer assessment on the development of clinical performance in undergraduate physical therapy education. In peer assessment, participants (students or professionals) evaluate (assess) the performance of their colleagues and provide each other of performance feedback. In this study, participants alternately perform the role of physical therapist, assessor, and client in a role-play simulating physical therapy practice. Students alternately performed in the role of physical therapist, assessor, and patient. Oral face-to-face feedback was provided as well as written feedback and scores based on performance indicators. In this study peer assessment is conceived as a learning task containing varying learning elements that potentially impact on learning and improvement. To explore how peer assessment impacts on learning, a selected group of 14 students was asked to rank these task elements from the highest to the lowest learning value and to motivate their choice. The analyses showed that 'performing the task in the physical therapist role' was

[1] The term 'care' includes all services described in the professional profile of the physical therapist.

perceived as the most powerful learning experience despite the performance stress that some participants experienced, followed by 'receiving teacher feedback'. The third place (before receiving peer feedback) was assigned to 'observing the performance of others'. Peer assessment triggered explicit learning such as 'reasoning aloud' and 'reflection' and implicit learning such as 'coping with performance stress' and 'role-modeling'. Students reported that the peer assessment task resulted in increased self-confidence, insight in performance standards and awareness of improvement areas. Conditions for learning related to the quality of feedback and the safety of the learning environment.

Chapter 3

Chapter 3 describes the effectiveness of an intervention aimed at improving guideline adherence in professional practice. In a randomized controlled trial the effectiveness of case based peer assessment is compared to case based discussions. The interventions were designed to enhance the implementation of a clinical practice guideline on a-specific low back pain. Participants were physical therapists (n=90) organized in 10 communities of practice (10F's) randomly assigned to the intervention group (peer assessment) and control group (case discussion). All groups participated in a program of four sessions focusing on a set of written clinical cases. The peer assessment intervention design was adopted from the study described in chapter 2. Additionally, they developed and evaluated a personalized improvement plan. The feedback process was supported by a coach. Case-based discussion – the regular implementation strategy – focuses on problem elaboration supported by a number of questions; participants' roles are not defined. The primary outcome was the increase of knowledge and guideline-based reasoning measured with an online test based on four clinical vignettes followed by closed questions at baseline and follow-up. The second outcome measure was the increase of self-reflection, measured by the 'self-reflection and insight scale'. The online test with clinical vignettes was completed at baseline and follow up by 78 participants (87%). Multilevel analysis showed that the estimated progress of the intervention group was 8.4% and of the control group - 0.1% (intervention effect = 8.7%, confidence interval = 3.9-13.4; *P* = .001). We found no difference between groups in reflection and insight. We concluded that peer assessment is more effective than case-based discussion for the implementation of clinical practice guidelines.

Chapter 4

Chapter 4 describes an in-depth analysis of the results of the trial in chapter 3 to identify the critical factors of peer assessment that contributed to the effectiveness of this implementation strategy. By unpacking the program we identified three main tasks (performance in the physical therapist role, assessor role, and patient role) and eleven subtasks. With an online survey, participants (n=49) were asked to rank the eleven subtasks from the highest to the lowest learning value and to motivate their choice in an open field. Additional semi-structured interviews were conducted (n=6) to gain more insight in the questionnaire results. The ranking results were statistically analyzed, the written comments and interview data with qualitative analysis. The ranking results showed that performance in the physical therapist role was perceived as the most valuable learning experience (respectively receiving peer feedback and expert feedback) although task perceptions ranged from challenging to threatening. These perceptions were related to differences in views on the usefulness of role-play and features of the task structure. In general, participants showed a strong intellectual and emotional involvement with the task. Peer assessment stimulated self-assessment, critical reflection and the integration of different perspectives on the cases discussed. The intervention resulted in changed attitudes towards the guideline on a-specific low-back pain, increased awareness of personal performance, shared quality standards of performance, and increased self-efficacy beliefs.

Chapter 5

This chapter includes a study on the effectiveness of two interventions aiming to enhance the implementation of the multi-disciplinary guideline on upper extremity complaints (CANS). It is a trial comparable to the trial in chapter 3, but the number of participants was twice as big. Participants were physical therapists (n=149) organized in 20 communities of practice randomly assigned to the intervention (peer assessment) or control group (case-based discussion). Both interventions were comparable to those described in chapter 3. All groups participated in a program of four sessions and worked on identical clinical cases representing the patient profiles described in the guideline on CANS (complaints of the arm, neck and shoulder). The outcome measures were: 1) increase in knowledge and guideline consistent reasoning measured with an online test on clinical vignettes at baseline and follow up, 2) increase in reflection and insight, measured with the Self-reflection and Insight Scale at baseline and follow up, 3) awareness of performance,

tested via the correlation between perceived and assessed improvement, 4) the extent to which personal goals were achieved, measured with a 3-point Likert scale (1=not achieved, 2=partly achieved, 3=completely achieved). Multilevel analysis showed that both groups improved. The average improvement for the peer assessment group was 5.8%, for the case discussion group 2.0% , and that difference was significant (intervention effect: 22.52 points, 95% confidence interval = 2.38 - 42.66, *P* = .03). Both groups were significantly improved in reflection and insight, but the differences between the groups were not significant. Self-awareness – the correlation between perceived and assessed improvement – was higher in the intervention group (r = 0.36) than in the control group (r = 0.08) and that difference was significant (intervention effect : 14.73; 95% confidence interval = 2.78-26.68, *P* = .01). The intervention group was also more effective in achieving personal goals (intervention effect: 0.50, 95% confidence interval = 0.04 - 0.96, *P* = .03). We concluded that peer assessment is more effective as an implementation strategy for clinical guidelines than case discussion, and this conclusion confirms the findings of the trial in chapter 3.

Chapter 6

This chapter presents a new quality improvement program that aims to enhance the effectiveness, client-centeredness, and transparency of physical therapy care. Participants were physical therapists (n = 64) of working communities of practice organized in a network of primary care professionals. The program consisted of two feedback interventions: 1) peer assessment and 2) practice visitation (clinical audit). Both interventions are part of a more comprehensive quality system that aims to self-regulate the quality of physical therapy services and is part of the Master Plan Quality in Motion, launched by the Professional Association (KNGF) which also involves benchmarking patient reported experiences (PREMS) and – outcomes (PROMS). Peer assessment focuses on professional performance, practice visitation on organization and management. In this chapter, the feasibility of this program is explored in a pilot study with a small group of physical therapists. The program consisted of 1) online self- and peer assessment and face-to-face discussion of results in two cycles over a period of approximately 6 months, followed by 2) practice visitation. Self- and peer assessment of professional performance focused on client communication and record keeping. Participants video-recorded a conversation with their client and uploaded this video recording on a specially designed website and added the corresponding client record. Self and peer assessment

was based on predefined performance indicators that could be scored online on a 5-point Likert scale of 1 = much improvement needed to 5 = no improvement needed, completed with written feedback. During the sessions participants reflected on the feedback received in dialogue with the feedback provider and differences in opinion were discussed. The sessions were supported by a trained coach. After cycle 1, personal improvement goals were formulated that were evaluated after cycle 2.

Practice visits were conducted by two physical therapists, trained to perform this role.

The perceived usability of the program for quality improvement was evaluated qualitatively with two focus groups and 10 in-depth interviews. In addition, we evaluated the impact on quality improvement quantitatively by comparing self- and peer assessment scores on performance indicators between Cycles 1 and 2.

Content analyses of interviews allowed for identifying critical success features relevant to program development and implementation, such as 'alignment of program expectations', 'training in peer assessment skills', 'skilled group coaches' and 'user-friendly information technology'. Participants reported, among other things, 'more awareness of clinical performance', more understanding of 'evidence-based and client-centered performance' and 'motivation to continue with peer assessment and visitation within their professional network'.

The results of the quantitative analyses show that the online activities of participants were limited in cycle 1 so data on the improvements made in cycle 2 were limited. Participants were reluctant in making client information available and needed to get familiar with using information technology and using a website that was not user friendly. Differences between self-scores and peer scores on performance indicators were not significant, although self-scores were lower on the majority of performance indicators. Between Cycle 1 and 2, the scores for record keeping significantly improved (average improvement: self-assessment = 0.20 points, $P = .007$, peer assessment = 0.15, $P = .002$). That did not apply to client communication (average improvement: self-assessment = 0.10, $P = .674$; peer assessment = -0.09, $P = .386$).

This study has shown that self-assessment, peer assessment, and visitation can be effective in supporting self-regulation of healthcare quality although the generalizability of the results is limited due to the small sample size. If the program is evaluated on a larger scale, improvement of the program design and the implementation strategy is recommended.

Chapter 7

The quality program described in Chapter 6 has been adjusted based on the evaluation results and subsequently implemented on a larger scale. Chapter 7 describes the results of a study on the effectiveness of self- and peer assessment on quality improvement in order to decide on the maturity of the program for nation-wide implementation. Participants were physical therapists (n = 379) related to four networks of professionals working in primary care. Quality improvement was tested quantitatively by comparing scores on performance indicators for client communication and record keeping in Cycle 1 and Cycle 2. In addition, personal improvement goals formulated after cycle 1 were analyzed thematically. On a questionnaire administered after cycle 2, participants could indicate on a 3-point Likert scale to what extent their personal goals were achieved (1 = not achieved, 2 = partly achieved, 3 = completely achieved).

In total 364 (94%) participants were active in online self-assessment and peer assessment. However, online activities varied between cycle 1 and 2, and between client communication and record keeping. Personal goals addressed: client-centered communication (54%), record keeping (24%), performance- and outcome measurement (15%), other (7%). Personal goals were completely attained by 29% of the participants, partly by 64%, and 7% did not attain their personal goals. Self-assessment and peer assessment scores improved significantly for both client communication (self-assessment=11%; peer assessment=8%) and record keeping (self-assessment=7%; peer assessment=4%).

We concluded that the program was effective in improving the effectiveness and client-centeredness of physical therapy care, and that nation-wide implementation is justified.

Chapter 8

In chapter 8 a new performance assessment design is introduced: the Script Concordance Test (SCT). The SCT aims at improving clinical reasoning and reducing unwanted variation among professionals. The answering key allows for variation in best answers. Clinical reasoning is considered as a critical competency for solving clinical problems and for the quality of physical therapy care, especially since physical therapists are directly accessible without referral of a physician. Evidence-based clinical practice guidelines provide the best available evidence to support the problem solving process. However, guidelines are not available for all clinical problems or the context of the clinical problem is not appropriate to apply them.

Therefore decisions on the diagnosis or treatment of a clinical problem may vary among professionals. The study in this chapter explores the utility of the SCT as a tool to enhance clinical reasoning in the musculoskeletal domain for both students and professionals, by assessing its reliability, validity, and acceptability. Participants were students of two universities (n=741) and professionals working in clinical practice (n=562). An SCT question consists of a short clinical case (script) followed by three pieces of additional information (scenario) relevant to the diagnosis or treatment. Each scenario is followed by a test item. On a 5-point Likert scale participants indicate the effect of the additional information on the plausibility of the hypothesis or the appropriateness of the proposed action. The participant's response to each question is compared with the answers of an expert panel (n = 22). Credit is assigned to each response based on how many of the experts on the panel choose that response. A maximum score of 100% is given for the modal response. Other responses are given partial credit, depending on the proportion of experts choosing them. Responses not selected by experts receive zero points. A team of teachers from two universities developed an online SCT containing 18 validated scripts and 54 test items, some illustrated by video-recordings. Completion time was limited to 100 minutes. Participants were provided with immediate feedback and reference to relevant, online available, literature. The reliability in terms of internal consistency was tested with Cronbach alpha. We pre-defined 7 expertise levels: 4 bachelor levels and 3 professional levels. Informed by dual processing theory, construct validity was assessed by testing the hypothesis that higher expertise in the musculoskeletal domain would produce higher SCT-scores in less response time. We tested in-between level differences for SCT-cores and response time with UNIANOVA linear models. Test acceptability was explored with a 6-item questionnaire which could be scored on a 5-points Likert scale.

The results show that the internal consistency of 18 scripts was acceptable (Cronbach alpha = 0.69). Mean SCT-scores differed significantly between students and professionals: mean difference = 6.08; $P < .001$. Higher expertise was related to higher SCT-scores but in-between differences were not always significant. Unlike our hypotheses, students used less response time than professionals: mean difference = 0.55 minutes, $P < .001$. On average, participants perceived the SCT as an acceptable tool for quality improvement, although the online feedback functionality can be improved (students: 2.88-3.80; professionals: 3.00-4.00).

We concluded that the SCT is a promising tool to enhance clinical

reasoning. Its quality can be improved by improving the feedback functionality and increasing the number and variety of scripts.

Chapter 9

Finally, in chapter 9 we returned to the research questions and critically reflect on the process of program development and implementation to inform the de design of a sustainable quality improvement system. Seven interventions were evaluated: six interventions using peer assessment, one intervention using the script concordance test.

All programs with peer assessment were perceived as useful quality improvement strategies. However, some physical therapists encountered difficulty in exposing their professional performance to the critical review of their peers and perceived performance stress. They however succeeded in coping with these stress triggers and recognized that exposure – 'Say what you think' and 'show what you do' is necessary to receive personal feedback. Receiving personal feedback was embraced because it is scarcely provided in daily practice. Feedback from practice visitors (auditors) was also appreciated and viewed as useful input for organizational development.Observing and assessing others was perceived instructive, however difficult. Participants needed time to become familiar with using performance indicators and needed training in providing constructive feedback. All studies involving peer assessment showed that providing narrative feedback was clearly preferred over providing scores.

Regarding the outcomes of peer assessment interventions – the impact on learning and change of professional behavior change – a distinction can be made between 'tested change' and 'self-reported change'. The tested outcomes show that peer assessment is a more effective strategy to improve evidence-based clinical reasoning than case discussion. In addition, peer assessment enhances the development of a realistic self-concept of performance. Looking at the outcomes of the trials described in chapters 3 and 5, the baseline and follow-up scores on the online test with clinical vignettes vary widely, implying that there is still much room for improvement for the peer assessment group. Presumably, longer interventions are needed to reduce that variation.

When physical therapists scored themselves or each other on performance indicators, less variation was observed. In addition, baseline scores were high. However, the difference between baseline and follow-up scores for all indicators was significant. Apparently, performance indicators were effective in uncovering shortcomings in competency development and in steering improve-

ment processes. Development and validation of performance indicators for different quality domains is therefore recommended.
To describe the self-reported impact of peer assessment on learning and behavior change, a distinction can be made between learning processes and learning outcomes. Participants reported implicit learning processes, such as coping with performance stress, mirroring and role-modeling the behavior of their peers, demonstrating the added value of 'show what you do'. Explicit learning related to reasoning and reflecting aloud, arguing for the added value of 'say what you think'.
This thesis also demonstrated that the process providing peer feedback, receiving feedback, and using feedback for targeted quality improvement, requires coaching. The role of the coach to facilitate the process and to foster psychological safety is of increasing importance when participants become more vulnerable in showing themselves, such as in video recordings. In addition, coaches are important to monitor the level of clinical reasoning, or increasing this level when needed, by in-depth questioning.
In this chapter we also reflect on the process of program development and implementation including the choices made on the basis of program evaluation. Subsequently, some recommendations are provided for policy makers, such as integrating peer assessment in the curricula of universities, strengthening professional networks, ongoing development and validation of performance indicators, and the development of a new quality registry.
We also provided some recommendations for physical therapists regarding areas for professional development, in particular for client-centered communication and shared decision-making.
Currently, the program with peer assessment and visitation introduced in Chapter 6 will be nation-wide implemented. Coaches and practice visitors are trained on a large scale and new programs are developed that cover other areas of professional competence.
We can conclude that peer assessment and visitation is a promising strategy to support professionals and organizations to self-regulate and account for the quality of their services. Future research will focus on the sustainability of the impact of peer assessment and visitation on professional and organizational development. In addition, the effects on client experiences and client outcomes are currently unclear and should be part of the research agenda.

Samenvatting

Hoofdstuk 1

Mensen die de hulp van een fysiotherapeut inroepen hebben recht op de best mogelijke zorg voor hun gezondheidsprobleem. Die wordt geleverd door up-to-date opgeleide professionals die kunnen inspelen op veranderingen in de zorgvraag van cliënten en de toenemende politieke en maatschappelijke vraag naar 'zinnige en zuinige zorg[1] van goede kwaliteit'. De uitdagingen voor de fysiotherapie liggen in het bevorderen van de effectiviteit, cliëntgerichtheid en transparantie van het proces en de uitkomsten van de fysiotherapeutische zorg. Om doelgericht aan kwaliteit te werken, hebben fysiotherapeuten – in opleiding en in de beroepspraktijk – criteria nodig voor zorg van goede kwaliteit en bruikbare feedback over de mate waarin zij individueel en collectief aan deze criteria voldoen. In deze thesis worden feedbackinterventies beschreven op basis van *performance assessment* die ontwikkeld zijn dóór en vóór fysiotherapeuten om professionals, teams en organisaties te ondersteunen in de zelfregulatie en verantwoording van hun kwaliteit. Dit proefschrift bespreekt / beantwoordt de volgende onderzoeksvragen:
Hoe ervaren fysiotherapeuten interventies, gebaseerd op *performance*-feedback, om de kwaliteit van fysiotherapie te verbeteren?
Wat is de impact van interventies, gebaseerd op *performance*-feedback op het leren en het veranderen van professioneel gedrag?

Hoofdstuk 2

In dit hoofdstuk worden de resultaten beschreven van een evaluatief onderzoek naar de impact van *peer*-assessment op de ontwikkeling van klinische vaardigheden bij bachelorstudenten fysiotherapie. *Peer*-assessment is een leeractiviteit waarbij studenten of collega's (*peers*) elkaar beoordelen en feedback geven op de kwaliteit van hun handelen (*performance*). De leertaak bestaat uit een rollenspel waarin studenten afwisselend de rol van fysiotherapeut, assessor en cliënt spelen. In de rol van fysiotherapeut demonstreert de student zijn vaardigheden, ontvangt hij *peer*-feedback en schrijft hij na afloop een reflectieverslag. In de rol van assessor observeert de student de performance van zijn peer en geeft mondelinge en schriftelijke feedback op basis van performance-indicatoren (beoordelingscriteria). In deze studie wordt *peer*-assessment opgevat als een leertaak die verschillende elementen bevat die een impact kunnen hebben op het leren en verbeteren. Om een indruk te krijgen hoe *peer*-assessment het leren stimuleert is aan een geselecteerde groep van veertien studenten door middel van interviews gevraagd

[1] De term 'zorg' omvat alle diensten die in het beroepsprofiel van de fysiotherapeut omschreven zijn.

om deze elementen te rangschikken van de hoogste naar de laagste leerwaarde en vervolgens hun keuze te motiveren. Uit de analyse bleek dat studenten performance in de rol van fysiotherapeut het leerzaamst vonden, ondanks dat deze activiteit voor sommige studenten als stressvol ervaren werd. Op de tweede plaats stond het krijgen van docent-feedback en op de derde plaats (boven het krijgen van *peer*-feedback) het observeren van de performance van anderen. *Peer*-assessment stimuleerde expliciet leren zoals 'hardop klinisch redeneren' en 'reflecteren', maar ook impliciet leren door 'coping met performance stress' en 'role-modeling'. Studenten rapporteerden dat de peer-assessment taak onder andere resulteerde in meer inzicht in de criteria voor een goede performance, informatie over de vaardigheden die nog verder ontwikkeld moeten worden en een toename van het vertrouwen in eigen kunnen (*self-efficacy*). Deelnemers rapporteerden dat de impact van *peer*-assessment op het leren wordt bevorderd door kritische verbeterfeedback in een veilige leeromgeving.

Hoofdstuk 3

Dit hoofdstuk beschrijft de effectiviteit van een interventie die gericht is op het bevorderen van 'evidence-based practice' in de beroepspraktijk. Het is een gerandomiseerde, gecontroleerde studie waarin de effectiviteit van twee interventies wordt vergeleken: 'casusbespreking' en '*peer*-assessment'. De interventies zijn ontworpen om de implementatie van de richtlijn aspecifieke lage rugklachten te bevorderen. Deelnemers waren fysiotherapeuten (n=90) georganiseerd in tien intercollegiale overleggroepen fysiotherapie (IOF's) die willekeurig aan de interventiegroep (*peer*-assessment) of de controlegroep (casusdiscussie) werden toegewezen. Alle groepen namen deel aan een programma van vier sessies waarin een vooraf vastgestelde set schriftelijke casussen besproken werd. De *peer*-assessment-interventie is ontleend aan het ontwerp dat beschreven is in hoofdstuk 2. Daarnaast werd een persoonlijk verbeterplan gemaakt op basis van feedback. Het feedbackproces werd ondersteund door een coach. Bij casusbespreking – de reguliere implementatiestrategie – wordt het probleem in de groep uitgewerkt aan de hand van een aantal vragen; rollen binnen de groep zijn niet gedefinieerd. De primaire uitkomstmaat was de toename van kennis en klinisch redeneren conform de richtlijn. Het meetinstrument was een *online*-casustoets op basis van vier patiëntprofielen die als voormeting en eindmeting werd gebruikt. Secundaire uitkomstmaat was de toename van reflectie en inzicht gemeten met de 'self-reflection and insight scale'. De online-casustoets werd inge-

vuld door 78 deelnemers (87%). *Multilevel*-analyse liet zien dat de interventiegroep een geschatte vooruitgang boekte van 8,4% en de controlegroep een terugval van 0,1% (interventie-effect = 8,7%, betrouwbaarheidsinterval = 3,9-13,4; *P* = ,001). We vonden geen verschil tussen de groepen in reflectie en inzicht.
We concludeerden dat *peer*-assessment een effectievere interventie is dan casusdiscussie voor de implementatie van praktijkrichtlijnen.

Hoofdstuk 4

In hoofdstuk 4 wordt een diepteanalyse gegeven van de resultaten van het experiment in hoofdstuk 3 om de kritische factoren van *peer*-assessment te kunnen identificeren die bijgedragen hebben tot de effectiviteit van deze implementatiestrategie. Door middel van taakanalyse werd het programma uiteengerafeld in drie hoofdtaken (performance in de rol van fysiotherapeut, van assessor en van cliënt) en elf subtaken. Na afloop van de trial werd aan de deelnemers in de *peer*-assessment-groep (n=49) via een *online survey* gevraagd de elf subtaken te rangschikken van het meest naar het minst leerzaam drie hoofdtaken (*performance* in de rol van fysiotherapeut, van assessor en van cliënt) en elf subtaken en hun keuze te motiveren. Aanvullende semigestructureerde interviews (n=6) werden uitgevoerd om een dieper inzicht te verwerven in de verkregen informatie. De resultaten van de rangschikking werden statistisch geanalyseerd, de schriftelijke en mondelinge informatie met een kwalitatieve analyse. De resultaten lieten zien dat *performance* in de rol van fysiotherapeut als het meest leerzaam ervaren werd (respectievelijk het krijgen van *peer*-feedback en *expert*-feedback), hoewel de taakpercepties varieerden van uitdagend tot bedreigend. In het algemeen gaven deelnemers blijk van een sterke intellectuele en emotionele betrokkenheid bij het uitvoeren van de opdracht. *Peer*-assessment stimuleerde *self*-assessment, kritische reflectie en het integreren van verschillende perspectieven op de besproken casuïstiek. De interventie resulteerde in een veranderde houding ten opzichte van de richtlijn aspecifieke lage rugklachten, meer inzicht in de eigen prestaties, een gedeelde visie op kwaliteit en een toename van het vertrouwen in eigen kunnen (*self-efficacy*).

Hoofdstuk 5

In dit hoofdstuk wordt de effectiviteit van twee interventies beschreven die gericht zijn op de implementatie van de multidisciplinaire richtlijn 'Klachten van de nek, arm, en schouder (KANS)'. Het betreft een gerandomiseerde, gecontroleerde *trial* die vergelijkbaar is met de trial in hoofdstuk 3, maar het aantal deelnemers was dubbel zo groot.

Deelnemers waren fysiotherapeuten (n=149) georganiseerd in twintig IOF's die willekeurig aan de interventiegroep (*peer*-assessment) of de controlegroep (casusdiscussie) werden toegewezen. Ook de interventies zijn vergelijkbaar met die beschreven in hoofdstuk 3. Alle groepen namen deel aan een programma van vier sessies en werkten aan dezelfde schriftelijke casussen ontleend aan de gezondheidsprofielen die in de richtlijn KANS beschreven zijn. De uitkomstmaten waren: 1) toename van kennis en klinisch redeneren conform de richtlijn, gemeten met een online-casustoets als voor- en nameting, 2) toename van reflectie en inzicht, gemeten met de 'self-reflection and insight scale' als nul- en eindmeting, 3) bewustzijn van de eigen prestaties, gemeten met de correlatie tussen de zelf-gerapporteerde verbetering en getoetste verbetering van kennis en klinisch redeneren conform de richtlijn en 4) de mate waarin persoonlijke veranderdoelen na afloop van de interventie gerealiseerd zijn, getest met 3-punt-Likert-schaal (1=niet behaald, 2=deels behaald, 3=volledig behaald). *Multilevel*-analyse liet zien dat in beide groepen verbetering was opgetreden. De gemiddelde verbetering voor de *peer*-assessmentgroep was 5,8%, voor de discussiegroep 2,0% en dat verschil was significant (interventie-effect: 22,52 punten, 95% betrouwbaarheids interval = 2,38 – 42,66, *P* = ,03). Reflectie en inzicht waren in beide groepen verbeterd, maar de verschillen tussen de groepen waren niet significant. Bewustzijn van de eigen prestaties – de correlatie tussen ervaren en getoetste verbetering – was hoger in de interventiegroep (r = 0,36, *P* = ,002) dan in de controlegroep (r = 0,08, *P* = ,50) en dat verschil was significant (interventie-effect: 14,73; 95% betrouwbaarheidsinterval = 2,78–26,68, *P* = ,01). De interventiegroep was ook effectiever in het realiseren van persoonlijke doelen (interventie-effect: 0,50, 95% betrouwbaarheidsinterval = 0,04 – 0,96, *P* = ,03). Wij concludeerden dat *peer*-assessment een geschiktere implementatiestrategie is voor praktijkrichtlijnen dan casusdiscussie en die conclusie bevestigt de bevindingen van de *trial* in hoofdstuk 3.

Hoofdstuk 6

In hoofdstuk 6 wordt een nieuw programma gepresenteerd dat gericht is op effectiviteit, cliëntgerichtheid en transparantie van de dienstverlening door fysiotherapeuten. Deelnemers waren fysiotherapeuten (n=64) georganiseerd in thematische werkgroepen en aangesloten bij een professioneel eerstelijns netwerk. Het programma bestond uit twee feedbackinterventies: *peer-assessment* en visitatie. Beide interventies zijn onderdeel van een meer omvattend kwaliteitssysteem dat de zelfregulatie van de kwaliteit van de fysiotherapeutische dienstverlening beoogt en onderdeel is

van het Masterplan Kwaliteit in Beweging (MKIB). Dat programma is gelanceerd door het Koninklijk Genootschap voor Fysiotherapie (KNGF) en omvat ook het benchmarken van patiënt-gerapporteerde ervaringen (PREMS) en patient-gerapporteerde uitkomsten (PROMS). *Peer-assessment* is gericht op professioneel handelen, visitatie – ook wel audit genoemd – is gericht op organisatie en management. In dit hoofdstuk wordt de bruikbaarheid van het programma onderzocht in een *pilotstudy* met een kleine groep fysiotherapeuten.
Het programma bestond uit 1) een *online-self-* en *peer-assessment*, gevolgd door een face-to-face bespreking van de resultaten in twee cycli over een periode van ongeveer zes maanden, gevolgd door 2) praktijkvisitatie. *Self-* en *peer-assessment* was gericht op cliëntcommunicatie en dossiervorming. De deelnemers maakten een video-opname van een gesprek met hun cliënt, plaatsten deze opname op een daarvoor ontworpen website en voegden vervolgens het bijbehorende cliëntendossier toe. *Self-* en *peer-assessment* was gebaseerd op vooraf gedefinieerde *performance*-indicatoren die op een 5-punt-Likert-schaal gescoord konden worden van 1 = 'veel verbetering nodig' tot 5 = 'geen verbetering nodig', aangevuld met schriftelijke feedback. Tijdens de besprekingen werd op de feedback gereflecteerd in dialoog met de feedbackgever en verschillen in opvatting besproken. De besprekingen werden begeleid door een getrainde coach. Na cyclus 1 werden persoonlijke verbeterdoelen geformuleerd, die na cyclus 2 geëvalueerd werden.
Visitatie van de praktijk werd uitgevoerd door twee fysiotherapeuten, getraind als visiteur.
De bruikbaarheid van het programma voor kwaliteitsverbetering is kwalitatief onderzocht met twee focusgroepen en tien diepte-interviews. Daarnaast is de impact op kwaliteitsverbetering kwantitatief geëvalueerd door het vergelijken van *self-* en *peer-assessment*-scores op *performance*-indicatoren tussen cyclus 1 en 2.
Met behulp van een kwalitatieve analyse zijn de kritische succesfactoren geïdentificeerd die relevant zijn voor de programmaontwikkeling en implementatie, zoals 'het afstemmen van verwachtingen', 'training van *peer-assessors* in beoordelingsvaardigheden', 'bekwame groepscoaches' en 'gebruiksvriendelijke informatietechnologie'.
Deelnemers rapporteerden onder andere meer 'bewustzijn van hun handelen', meer 'inzicht in evidence-based en cliëntgericht handelen' en 'motivatie om door te gaan met *peer*-assessment en visitatie binnen hun professionele netwerk'.
De resultaten van de kwantitatieve analyse laten zien dat deelnemers beperkt online actief zijn geweest in cyclus 1, met als gevolg dat informatie over de gemaakte verbeteringen in cyclus 2 beperkt was;

deelnemers waren terughoudend in het beschikbaar stellen van cliëntinformatie en moesten wennen aan het gebruik van informatietechnologie. Bovendien werkte de website onvoldoende intuïtief. De verschillen tussen *self*- en *peer*-scores waren niet significant in cyclus 1 en 2, maar de *self-scores* waren op het merendeel van de prestatie-indicatoren wel lager. Tussen cyclus 1 en cyclus 2 bleek dat de scores voor dossiervoering significant verbeterd waren (gemiddelde verbetering: *self-assessment* = 0,20 punten, *P*=,007, *peer-assessment* = 0,15, *P*=,002). Dat gold niet voor de scores voor communicatie (gemiddelde verbetering: *self-assessment* = 0,10, *P*=,674; *peer-assessment* = - 0,09, *P*=,386).
Deze studie heeft aangetoond dat *self-assessment, peer-assessment* en visitatie effectief kunnen zijn om de zelfregulatie van de kwaliteit van de zorg te ondersteunen ofschoon de resultaten beperkt generaliseerbaar zijn door de kleine steekproef. Als het programma op grotere schaal wordt geëvalueerd, wordt verbetering van het programmaontwerp en de implementatiestrategie aanbevolen.

Hoofdstuk 7

Het kwaliteitsprogramma zoals beschreven in hoofdstuk 6, werd bijgesteld op basis van de evaluaties en vervolgens op grotere schaal geïmplementeerd. In hoofdstuk 7 worden de resultaten beschreven van een onderzoek naar de effectiviteit van het *self*- en *peer-assessment* voor kwaliteitsverbetering om een uitspraak te kunnen doen over de rijpheid van het systeem voor landelijke implementatie. Deelnemers waren fysiotherapeuten (n=379) aangesloten bij vier professionele eerstelijns netwerken. De kwaliteitsverbetering werd kwantitatief getoetst door de scores op *performance*-criteria voor communicatie en dossiervoering in cyclus 1 en cyclus 2 met elkaar te vergelijken. Daarnaast werden de persoonlijke verbeterdoelen die na cyclus 1 werden geformuleerd, thematisch geanalyseerd. Na cyclus 2 ontvingen deelnemers een vragenlijst waarbij ze op een driepunts Likert-schaal konden aangeven in hoeverre hun verbeterdoelen bereikt waren (1 = 'niet bereikt', 2 = 'gedeeltelijk bereikt', 3 = 'volledig bereikt'). In totaal zijn 351 (93%) fysiotherapeuten *online* actief geweest in cyclus 1 en 2. *Self*- en *peer-assessment*-scores zijn significant verbeterd voor zowel cliëntcommunicatie als dossiervoering. De gemiddelde verbetering van *self*- en *peer*-assessment-scores voor cliëntcommunicatie was respectievelijk 11% en 7% en voor dossiervoering 8% en 4%.
De leerbehoeften na cyclus 2 betroffen voornamelijk cliëntcommunicatie met inbegrip van gezamenlijke besluitvorming (54%), dossiervoering met inbegrip van het formuleren van meetbare doelen (24%),

het gebruik van performance-testen en patiënt-gerapporteerde uitkomsten (15%) en overige thema's (7%). Deze doelen werden door 29% volledig , 64% deels gerealiseerd en door 7% niet gerealiseerd. Wij hebben geconcludeerd dat *self-* en *peer-assessment* effectief is in het verbeteren van de doelmatigheid, cliëntgerichtheid en transparantie van het fysiotherapeutisch handelen en dat landelijke implementatie gerechtvaardigd is.

Hoofdstuk 8

In hoofdstuk 8 wordt een nieuwe vorm van *performance assessment* geïntroduceerd. Het betreft de *Script Concordance Test* (SCT). De SCT is bedoeld als een feedbacktool om het klinisch redeneren te bevorderen en ongewenste variatie tussen professionals te verminderen. De antwoordsleutel laat verschillen in het gewenste antwoord toe. Klinisch redeneren wordt beschouwd als een kritische competentie voor het oplossen van gezondheidsproblemen en voor de kwaliteit van het fysiotherapeutisch handelen. Evidence-based praktijkrichtlijnen kunnen het proces van klinisch redeneren ondersteunen. Richtlijnen zijn echter niet beschikbaar voor alle gezondheidsproblemen of de context van het probleem is niet geschikt om ze toe te passen. Daarom kunnen professionals van mening verschillen over de diagnose of behandeling van een gezondheidsprobleem. De studie in dit hoofdstuk onderzoekt de bruikbaarheid van de SCT als een middel om klinische redeneren te bevorderen in het musculoskeletale domein, zowel voor studenten als professionals. De bruikbaarheid wordt onderzocht door de betrouwbaarheid, de validiteit en de ervaren geschiktheid te beoordelen. Deelnemers waren studenten van twee hogescholen (n = 741) en professionals werkzaam in de beroepspraktijk (n = 562). Een SCT-vraag bestaat uit een korte casus (*script*) gevolgd door drie stukjes aanvullende informatie (*scenario*) relevant voor de diagnose of behandeling gevolgd door een test-item. De deelnemer geeft op een 5-pts Likert schaal aan in hoeverre de aanvullende informatie invloed heeft op de werkhypothese, het onderzoeks- of het behandelvoorstel. Het antwoord van de deelnemer op elk item wordt vergeleken met de antwoorden van een expertpanel (n = 22). Punten worden toegewezen op basis van de mate waarin het gekozen antwoord overeenkomt met het antwoord dat gekozen is door experts. Het maximale aantal punten wordt gegeven voor het modale (meest frequent gekozen) antwoord van het expertpanel (100%). Minder frequent gekozen antwoorden worden beloond naar verhouding. Antwoorden die niet door experts gekozen zijn, krijgen nul punten. Een docententeam afkomstig van twee hogescholen heeft de SCT

ontwikkeld. Deze bestond uit achttien gevalideerde scripts en 54 test-items, sommige geïllustreerd met video-opnames. De testtijd was beperkt tot 100 minuten. Deelnemers kregen na afloop direct feedback en verwijzingen naar relevante, *online* beschikbare literatuur. De betrouwbaarheid in termen van interne consistentie, werd getest met Cronbach alpha. Er zijn zeven niveaus van expertise gedefinieerd: vier bachelorniveaus en drie professionele niveaus. De constructvaliditeit werd onderzocht op basis van 'dual processing theory' door de hypothese te toetsen dat meer expertise op het musculoskeletale domein tot hogere scores zou leiden in minder responstijd. Verschillen tussen de niveaus van expertise in gemiddelde score en responstijd werden getest met UNIANOVA lineaire modellen. De ervaren geschiktheid van de toets werd onderzocht met een korte vragenlijst (zes vragen), die op een vijfpunts Likert-schaal kon worden gescoord.
Uit de resultaten blijkt dat de betrouwbaarheid (op basis van achttien scripts) acceptabel was (Cronbach alpha = 0,69). De gemiddelde scores verschilden significant tussen studenten en professionals: gemiddeld verschil = 6,08; p <,001. Meer expertise was gerelateerd aan hogere SCT-scores, maar de verschillen waren niet altijd significant. In tegenstelling tot onze hypothese gebruikten studenten minder responstijd dan professionals: gemiddelde verschil is 0,55 minuten, p <,001. Deelnemers waren gemiddeld tevreden over de geschiktheid van de toets als *feedbacktool*, maar er was ruimte voor verbetering van de *online*-feedbackfunctie (studenten: 2,88-3,80; professionals: 3,00-4,00).
We hebben geconcludeerd dat de SCT een veelbelovend instrument is om klinische redeneren te bevorderen. De kwaliteit kan worden verbeterd door de feedbackfunctie te optimaliseren en het aantal en de variatie in scripts te verhogen.

Hoofdstuk 9

Ter afsluiting komen we in hoofdstuk 9 terug op de onderzoeksvragen en reflecteren we op het proces van programmaontwikkeling en -implementatie om aanbevelingen te kunnen doen met betrekking tot de ontwikkeling van een duurzaam kwaliteitssysteem.
Zeven interventies werden geëvalueerd: zes interventies met *peer-assessment*, een interventie met de *script concordance*-test.
Alle programma's met *peer-assessment* werden gezien als nuttige interventies om kwaliteit te verbeteren. Echter, sommige fysiotherapeuten hadden moeite om zichzelf bloot te stellen aan het kritisch oog van hun collega's (*exposure*) en voelden prestatiedruk. Zij slaagden er vervolgens wel in om daarmee om te gaan

en erkenden dat *exposure* – 'vertellen wat je denkt' en 'laten zien wat je doet' – noodzakelijk is om persoonlijke feedback te krijgen. Persoonlijke feedback werd omarmd, omdat die in de dagelijkse praktijk zelden wordt gegeven. Ook bij visitatie werd feedback van collega's gewaardeerd en als bruikbare input gezien voor organisatieverbetering.
Het observeren en beoordelen van anderen werd als leerzaam, maar moeilijk ervaren. Deelnemers hadden tijd nodig om vertrouwd te raken met het gebruik van performance-indicatoren en training in het geven van constructieve feedback. Bij alle studies had het geven en ontvangen van kwalitatieve (woordelijke) feedback de voorkeur boven indicatorscores.
Met betrekking tot de uitkomsten van interventies met *peer-assessment* – de impact op leren en verandering van professioneel gedrag – kan een onderscheid gemaakt worden tussen 'getoetste' verandering met de *online*-casustoets en 'zelf-gerapporteerde verandering'. De getoetste uitkomsten laten zien dat *peer-assessment* een effectievere strategie is om *evidence-based* klinisch redeneren te bevorderen dan casusdiscussie. Bovendien helpt *peer-assessment* een realistisch zelfbeeld te ontwikkelen. Als we de uitkomsten van de experimenten uit hoofdstuk 3 en 5 bekijken, dan valt op dat de scores op de voor- en nameting sterk variëren hetgeen impliceert dat er ook voor de *peer-assessment*-groep nog veel ruimte voor verbetering is. Vermoedelijk zijn er langere interventies nodig om die variatie te reduceren.
Als fysiotherapeuten zichzelf of elkaar scores gaven op *performance*-indicatoren, was er minder variatie te zien. Bovendien waren de scores bij de voormeting hoog. Desondanks was het verschil tussen de voor- en nameting voor alle indicatoren significant. *Performance*-indicatoren – in dit geval voor communicatie en dossiervoering – zijn blijkbaar effectief geweest in het blootleggen van tekort-komingen in competentieontwikkeling en in het sturen van verbeterprocessen. Doorontwikkeling en validering van *performance*-indicatoren voor verschillende kwaliteitsdomeinen wordt daarom aanbevolen.
Om de zelf-gerapporteerde impact van *peer-assessment* op leren en gedragsverandering, te beschrijven, moet een onderscheid gemaakt worden tussen leerprocessen en leeruitkomsten. Deelnemers rapporteerden impliciete leerprocessen, zoals het omgaan met *performance stress (coping),* het spiegelen en modelleren van het gedrag van *peers (role-modeling*) hetgeen pleit voor de meerwaarde van 'laten zien wat je doet'. Expliciet leren had betrekking op hardop redeneren (*reasoning aloud*) en reflecteren wat pleit voor de meerwaarde van 'vertellen wat je denkt'.

Deze thesis heeft ook laten zien dat het proces van peer-feedback geven, feedback ontvangen, en feedback gebruiken om doelgericht kwaliteit te verbeteren, coaching behoeft. De rol van de coach om het proces te faciliteren en de groepsveiligheid te bewaken is van toenemend belang naarmate deelnemers meer van zichzelf laten zien, zoals in video-opnames. Bovendien zijn coaches belangrijk om het niveau van klinisch redeneren te bewaken, zo niet te verhogen door verdiepende vragen te stellen.
In dit hoofdstuk wordt vervolgens gereflecteerd op het proces van programma-ontwikkeling en implementatie inclusief de keuzes die gaandeweg gemaakt zijn op basis van programma-evaluatie. Vervolgens worden aanbevelingen gedaan voor beleidsmakers, zoals het integreren van *peer-assessment* in de curricula van hogescholen, het versterken van professionele netwerken, selectie en training van coaches, doorontwikkeling en validering van *performance*-indicatoren, en de ontwikkeling van een nieuw kwaliteitsregister. Ook worden aanbevelingen gedaan voor fysiotherapeuten met betrekking tot de ontwikkeling van verschillende competentiegebieden, maar in het bijzonder voor cliëntgerichte communicatie en gemeenschappelijke besluitvorming. Op dit moment wordt het programma met *peer-assessment* en visitatie dat geïntroduceerd is in hoofdstuk 6, landelijk geïmplementeerd. Coaches en visiteurs worden op grote schaal opgeleid en nieuwe programma's worden ontwikkeld die andere competentiegebieden bestrijken.
We kunnen concluderen dat *peer-assessment* en visitatie een veelbelovende strategie is om zelfregulatie en verantwoording van kwaliteit – dóór en vóór professionals – te ondersteunen. Toekomstig onderzoek zal zich moeten richten op de duurzaamheid van de impact van *peer-assessment* en visitatie op professionele- en organisatieontwikkeling. Bovendien zijn de effecten op cliëntervaringen en cliëntuitkomsten tot nu toe niet duidelijk en zullen een plaats moeten krijgen op de onderzoeksagenda.

Dankwoord

Mijn zoon Jurriaan moest de auto wassen. Daar begon het mee. Met forse tegenzin nam hij de tuinslang in zijn hand, zette de kraan aan en ontgrendelde het gloednieuwe spuitstuk. Hij was zichtbaar verrast door de krachtige waterstraal die het ding produceerde en spoot geamuseerd in het rond. Dat Ria Nijhuis op deze zonnige namiddag een ommetje maakte door onze straat en een flinke plens water te pakken kreeg, was achteraf een gelukje bij een ongelukje. Ria en ik raakten door dit voorval noodgedwongen aan de praat en twee maanden later begon ik als wetenschappelijk medewerker bij IQ healthcare. Die functie zou zeven jaar later resulteren in een promotie. Jurriaan, mijn zoon, dank voor je perfecte timing om aan de spontane loop van gebeurtenissen een subtiele wending te geven.
Graag wil ik alle andere mensen die me geholpen hebben om te kunnen promoveren ook bedanken, ook al worden ze in dit dankwoord niet specifiek genoemd. Die mensen weten hoe chaotisch ik ben en hoezeer ik aangewezen ben op mijn sociale vaardigheden om alles achteraf weer goed te praten.
Ria Nijhuis, mijn promotor, dank voor de kansen die je mij gegeven hebt en voor je loyaliteit in tijden dat ik die hard nodig had. Ik heb genoten van je authentieke, informele stijl van communiceren, je creatieve brainwaves waarop ik heerlijk kon surfen, je brede kennis van de literatuur en je kritische houding ten opzichte van de zuiverheid van de gebruikte methoden en de interpretatie van de resultaten. Hoewel deze combinatie van eigenschappen niet zo voor de hand ligt, vallen ze bij Ria heel mooi samen. Daarnaast heb ik bij Ria – als bij geen ander – de passie voor ons vak gevoeld en een soulmate gevonden in de bedoeling ervan. Omdat mijn studies zich bewogen hebben op het snijvlak van onderwijs en beroepspraktijk, was het perspectief van mijn tweede promotor, Cees van de Vleuten, onmisbaar. Cees, dank voor je verdiepende vragen, je helpende hand in de richting van de theorie, je stevige en beknopte taalgebruik, je pijlsnelle, ongewatteerde feedback en je bescheidenheid. Dat laatste vond ik heel indrukwekkend.
Vanaf de eerste dag bij IQ Healthcare, werkte ik als kersverse wetenschapper samen met de meer ervaren Philip van der Wees die later mijn co-promotor werd. We hebben door het hele land gereisd om onze implementatieprogramma's uit te voeren en dat werd een uitdaging toen ik ziek werd. Philip, jouw optimistische en sympathieke gezelschap maakte onze opdracht in die tijd licht en vrolijk. Ik bewonder je enorme werklust, je helicopterview als het gaat om de ontwikkelingen binnen de wetenschap en vooral je

reflectieve houding – “doen we het wel goed, brengt dit ons voor-uit?” Ik hoop oprecht dat we onze rijkdom aan ervaringen breder kunnen gaan inzetten. Jij hebt je grenzen inmiddels al verlegd, naar Washington nota bene. Yvonne Heerkens was mijn tweede promotor die het proces begeleid heeft als lector verbonden aan de HAN. Yvonne, dank voor jouw zorgvuldige reviews, je opmerkzaamheid voor details, je beschikbaarheid en je altijd warme belangstelling. Ook mijn co-auteurs wil ik bedanken voor hun bijdrage aan het tot stand komen van dit proefschrift.
Ik heb vroeger niet kunnen vermoeden dat ik nog ooit zou promoveren. Het is geweldig dat ik die kans heb gekregen van de HAN die mij een promotiebeurs heeft toegekend en ik dank Theo Joosten en Menno Pistorius voor het in mij gestelde vertrouwen in de goede afloop. Daarnaast wil ik het KNGF als belangrijkste subsidiegever voor mijn promotieonderzoek bedanken en de ondersteuning die ik daarbij gekregen heb, in het bijzonder van Annemarie Trompert en Having Perdon. Gerhard Zielhuis, Anneke Kramer en Lia Fluit wil ik bedanken voor het bestuderen en goedkeuren van het manuscript.
Met veel plezier kijk ik terug naar de samenwerking met mijn onderzoeksgroep, Simone van Dulmen, Guus Meerhoff, Femke Driehuis, Juliette Cruijsberg, Annick Bakker en Janine Lieffers. Samen met Philip en Ria opereerden we onder de naam A-team, die doet vermoeden dat wij zeer slagvaardig waren, maar de ‘A’ stond voor: Alles-altijd-op-het-nippertje-klaar. Zonder de stevige en toch charmante regie van Simone was dat laatste nooit gelukt. We hebben als team heel wat kilometers gemaakt, weerstanden getrotseerd, uitbundig gelachen en dapper de tegenvallers geslikt.
Voor alle tegenvallers die moeilijk te slikken waren kon ik terecht bij mijn kamergenoten bij IQ Healthcare, Mirella, Tim en Caroline en bij mijn kamergenoten bij de HAN: Ine, Piet, Anneke, Ria, Jaap en Wietske. Alle collega’s in het veld die onbezoldigd geholpen hebben bij de ontwikkeling en validering van verschillende toetsen – een tijd-rovende en ingewikkelde klus – wil ik bedanken, in het bijzonder Martin Opheij, Michel ten Bokum, Phia Dekker, Patrick Koekenbier en Peter Eemers. En ook mijn collega-docenten van de HAN en SAXION. Marcel van Brunschot van Vakbekwaamheid in Zicht, jouw tomeloze inzet voor de digitalisering van de script concordance test en de psychometrische analyse van de resultaten was onmisbaar. Jij stond altijd voor me klaar met creatieve oplossingen. En dat geldt ook voor de medewerkers van Compusense die online peer-assessment van dossiers en video’s mogelijk gemaakt hebben.
Bij de uitvoering van de verschillende pilots heb ik dankbaar gebruik gemaakt van de medewerking van de fysiotherapeuten aangesloten

bij de verschillende netwerken en de inspanningen van hun 'knowledge-brokers'. Ik wil in het bijzonder Menno Bouman, Frits van Trigt, Ron van Heerde, JanDiet Berendsen en Mathieu de Krieger bedanken.
Aan de wieg van de ontwikkeling en implementatie van peer-assessment stonden mijn creatieve collega's van de HAN: Els Lamers, Henk van Enck, Henk Nieuwenhuijzen en Volcmar Visser. Els was nog student toen van haar hand de eerste iconen van 'peren' verschenen, een surrealistische improvisatie op peer-assessment – zoiets als 'Ceci n'est pas une poire' – en die hebben we gehandhaafd. Dankzij de inspanningen van vriend en fotograaf Luuk Huiskes zijn aan deze peren meerdere dimensies toegevoegd voor dit proefschrift.
Mijn oude vrienden Thom Mertens en Inge Adelmeijer wil ik bedanken voor het reviseren van het manuscript en Robbert Zweegman voor de vormgeving van dit boek.
De 'frisse meisjesclub' (Margriet, Lisette en Jeanne) ontleent haar naam aan het vermogen om 'verfrissende' vragen te stellen. Dankjewel meisjes voor het zoeken naar mijn rode draad.
De laatste jaren heb ik heel wat tijd achter mijn computer besteed, ook in het weekend. Dat was voor mijn kinderen, Liselore en Jurriaan, niet altijd fijn. Ik hoop dat ik mijn tekortkomingen in deze periode goed heb kunnen maken tijdens onze – soms behóórlijk avontuurlijke – wereldreizen. Als ik daarop terug kijk, dan zie ik twee kinderen die tegen een stootje kunnen, met weinig tevreden zijn en op een ludieke, soms onnavolgbare manier, aan elkaar gehecht en gewaagd zijn. Marc, al zeventien jaar mijn trouwe ex-partner en vader van mijn kinderen, heeft daaraan zijn unieke bijdrage geleverd. Dankjewel Marc voor het geduldig ophalen van alle steekjes die ik gaandeweg liet vallen, voor je onvoorwaardelijke steun en je barokke tomatensoep op maandag.
Cees, mijn vader, heeft 95 jaar moeten worden om zijn eerste kind te zien promoveren. Dankjewel lieve vader voor je geduld. Helaas hebben we tijdens dit promotietraject afscheid moeten nemen van mama, wat jammer dat je deze gebeurtenis niet met haar kunt delen. Ik draag dit proefschrift aan haar op. Ik zou het ook best kunnen opdragen aan mijn ongepolijste juweel van een zus Nicky die met zoveel welgemeende interesse mijn verhalen heeft aangehoord, maar je moet ergens een grens trekken. Immers, de rol van mijn broers Edwin en Ruud en mijn zus Rian was ook heel constructief, zo niet constructiever, en laat ik de warmte van de 'kouwe kant' ook niet veronachtzamen.
Aan de zijlijn stonden altijd mijn lieve, trouwe, originele vrienden Vic en Patty, met wie ik hand in hand en door schade en schande 'wijs' geworden ben, vrolijk meedeinend op de spontane loop van gebeurtenissen.

Over de auteur

Marjo Maas werd in 1958 geboren in Den Bosch. Zij sloot in 1976 het VWO af aan het Van der Puttlyceum in Eindhoven. Daarna studeerde ze aan de Opleiding Fysiotherapie in Nijmegen waar ze haar diploma haalde in 1980 en aansluitend werk vond als fysiotherapeut. Na een tweejarige verkenning van het beroep van beeldend kunstenaar aan de Academie voor Beeldende Kunsten in Arnhem, kwam ze terug op haar eerste beroepskeuze. Ze pakte de draad op bij fysiotherapiepraktijk van Doorn-van Walterop. Deze praktijk zou ze later overnemen met drie collega's in maatschapsvorm onder de naam Fysiotherapie de Goffert. Ze specialiseerde zich in de behandeling van beroepsgerelateerde klachten bij musici en volgde daarvoor een aanvullende opleiding in Duitsland. In 1986 rondde zij de eerstegraads lerarenopleiding HBO-gezondheidszorg af aan de Rijksuniversiteit Limburg. Aansluitend kreeg zij een aanstelling als docent fysiotherapie bij de Hogeschool van Arnhem en Nijmegen. Als docent hield zij zich bezig met de implementatie van het PGO-onderwijs en de ontwikkeling van het vaardigheidsonderwijs (Skillslab) in samenwerking met het Transferpunt Vaardigheidsonderwijs Maastricht. Daarnaast legde ze zich toe op het verbeteren van de toetsing. De behoefte aan meer wetenschappelijke oriëntatie op het gezondheidszorgonderwijs resulteerde in de *Master of Health Professions Education* aan de Universiteit van Maastricht waar ze in 2009 cum laude afstudeerde. In 2010 kreeg zij een aanstelling als wetenschappelijk onderzoeker aan de Radbouduniversiteit afdeling IQ Healthcare. Als onderzoeker hield zij zich onder andere bezig met de uitvoering van het programma Kwaliteit in Beweging van de beroepsvereniging KNGF. Deze functie werd in 2012 gecombineerd met een promotietraject wat in 2017 werd afgesloten. In de tussentijd heeft Marjo haar taak als lid van de maatschap fysiotherapie de Goffert neergelegd, maar is op een laag pitje actief gebleven als fysiotherapeut. Momenteel werkt ze als docent bij de HAN Opleiding fysiotherapie, als onderwijskundig medewerker en als trainer basis- en seniorkwalificatie examinering (BKE en SKE) bij het Instituut Paramedische Studies. Tevens is ze als programmaontwikkelaar, trainer en wetenschappelijk onderzoeker verbonden aan Radboudumc IQ healthcare. Daarnaast werkt ze freelance als auditor bij de beoordeling van opleidingen en als adviseur bij vraagstukken rondom toetsing en professionalisering. Marjo heeft twee kinderen, Liselore (1993) en Jurriaan (1996) en woont in Nijmegen.

Lijst van publicaties

Wetenschappelijke publicaties (*opgenomen in dit proefschrift)

1 *Maas M, Sluijsmans D, van der Wees P, Heerkens Y, Nijhuis-van der Sanden M, van der Vleuten C. Why peer assessment helps to improve clinical performance in undergraduate physical therapy education: a mixed methods design. *BMC Med Educ*. 2014;14(1):117.

2 *van Dulmen S, Maas M, Staal B, Rutten G, Kiers H, Nijhuis-van der Sanden M., van der Wees P. Effectiveness of peer-assessment for implementing a Dutch physical therapy low back pain guideline: a cluster randomized controlled trial. *Phys Ther.* 2014;94(10):1396–1409.

3 *Maas M, van Dulmen S, Sagasser M, Heerkens Y, van der Vlleuten C, Nijhuis-van der Sanden M, van der Wees P. Critical features of peer assessment of clinical performance to enhance adherence to a low back pain guideline for physical therapists: a mixed methods design. *BMC Med Educ.* 2015;15(1):203.

4 *Maas M, van der Wees P, Braam C, Koetsenruijter J, Heerkens Y, van der Vleuten C, Nijhuis-van der Sanden M. An Innovative peer assessment approach to enhance guideline adherence in physical therapy: A single-masked, cluster-randomized controlled trial. *Phys Ther.* 2015;95(4):600–612.

5 Staal B, van Haren I, Maas M, Kiers H, Nijhuis-van der Sanden M, de Graaf-Peters V. Serious gaming voor het vergroten van de adherentie van fysiotherapeuten en manueel therapeuten aan de richtlijn lage rugpijn: Een gerandomiseerde gecontroleerde studie. *Tijdschrift voor Gezondheidswetenschappen* 2016;7.

6 Meerhoff G, van Dulmen S, Maas M, Heijblom K, Nijhuis-van der Sanden M, van der Wees P. Development and evaluation of an implementation strategy for collecting data in a national registry and the use of patient reported outcome measures (PROMs) in physical therapist practices: quality improvement study. *Phys Ther*, 2017; 97,(8):837-851.

7 *Maas M, Nijhuis-van der Sanden M, Driehuis F, Heerkens Y, van der Vleuten C, van der Wees P. The feasibility of peer assessment and clinical audit to self-regulate the quality of physiotherapy services. A mixed methods study. *BMJ-open* 2017;7:1-10.

8 *Maas M, Driehuis F, Meerhoff G, Heerkens Y, van der Vleuten C, Nijhuis-van der Sanden M, van der Wees P. The impact of self-assessment and peer assessment on clinical performance of physical therapists in primary care: a cohort study. Accepted for publication. *Physiotherapy Canada*.

9 Otterman N , Maas M, Schiemanck S, van der Wees P, Kwakkel G. Can we identify specialized physical therapists in stroke rehabilitation? Submitted for publication.

10 *Maas M, Nijhuis-van der Sanden M, Rutten G, Heerkens Y. van der Wees P, van der Vleuten C. The utility of a script concordance test to self-direct continuous learning in physical therapy education and professional practice. Submitted for publication.

Andere publicaties

11 van Enck H, Biermans P, Maas M. *De Fysiotherapeutische Anamnese*. (Transferpunt Vaardigheidsonderwijs Rijksuniversiteit Limburg ed.) Maastricht: Unigraphic; 1993.

12 van Enck H, Maas M, Biermans P. *De Fysiotherapeutische Behandeling*. (Transferpunt vaardigheidsonderwijs Rijksuniversiteit Limburg ed.) Maastricht; Unigraphic 1994.

13 Ophey M, Maas M, de Beer J. Onderzoek naar de inhoudsvaliditeit van het performanceassessment in de hoofdfase van de bacheloropleiding fysiotherapie. *Tijdschrift voor Medisch Onderwijs*, 2006;25:2, 88-95.

14 van Dulmen S, van der wees P, Meerhoff G, Maas M, Nijhuis-van der Sanden M. *Eindrapport uitkomsten van zorg binnen de coöperatie Fysiocare. Pilot binnen het onderzoeksprogramma Kwaliteit in beweging*. Nijmegen, IQ healthcare, 2015

15 van Dulmen S, Meerhoff G, Maas M, van der Wees, Nijhuis-van der Sanden M. *Eindrapport uitkomsten van zorg binnen de coöperatie Fysio-Omni. Pilot binnen onderzoeksprogramma Kwaliteit in beweging.* Nijmegen, IQ healthcare 2015.

16 van Dulmen S, Maas M, van der Wees P, Nijhuis-van der Sanden M. *Eindrapport uitkomsten van zorg binnen de coöperatie Fysio-Ijsselland. Pilot binnen het onderzoeksprogramma Kwaliteit in beweging.* Nijmegen, IQ Healthcare 2015.

17 Maas M, van 't Schilt T, van der Vleuten C. *De Online Script Concordance Test om voortgang in klinisch redeneren te bevorderen van fysiotherapeuten in opleiding en in de beroepspraktijk.* Examens 2016;2.

18 Maas M. Script Concordance Test: Klinisch redeneren in onderwijs en beroepspraktijk. *Fysiopraxis* 2016;2.

Institute for Health Sciences
Radboudumc

PhD portfolio

Name	Marjo Maas (V)
PhD period	September 2012 – September 2017
Department	Radboudumc IQ healthcare
Promotors	Prof. Dr. MWG Nijhuis-van der Sanden
	Prof. Dr. CPM van der Vleuten
Co-promotors	Dr. PJ van der Wees
	Dr. YF Heerkens

Activities		Year	ECT
Training activities			
Qualitative Research Introduction Course CaRe		2012	1,0
Qualitative Research Interview training CaRe		2013	1,0
Qualitative Analysis with Atlas-ti Evers Research		2014	2,0
Academic Writing Radboud into languages		2014	3,0
Biometrics PAO Heyendaal		2015	3,0
Implementation Science IQ healthcare		2016	2,0
BROK certificate for Good Clinical Practice		2017	1,5
Teaching activities			
Writing learning materials for therapists, coaches and auditors		2015 - 2017	3,0
Training peer group coaches KNGF		2015 - 2017	3,0
Training coach trainers KNGF		2015 - 2017	2,0
Training auditors KNGF		2015 - 2017	2,0
Training SCT test development HAN AMC HZuyd		2014 - 2017	1,5
Reviewing activities			
Review of scientific publications for serveral journals		2015 - 2017	1,0
Symposia and Congresses			
Jaarcongres KNGF Amersfoort	Oral presentation	2012	0,2
HGZO congres Amsterdam	Oral presentation	2013	0,2
WCF Wetenschapsdag Amersfoort	Oral presentation	2014	0,2
IQ Healthcare Congres Nijmegen	Workshop	2014	0,5
World Congress Physical Therapy Singapore	Poster presentation	2015	0,5
KNGF jaarcongres Amersfoort	Workshop	2015	0,5
THIM Congres Utrecht	Oral presentation	2015	0,2
SROF dag Utrecht	Oral presentation	2015	0,2
SURF tender presentations Utrecht	Oral presentation	2016	0,2
WCF Wetenschapsdag Utrecht	Oral presentation	2016	0,2
Jaarcongres KNGF Amersfoort	Poster presentation	2016	0,2
EBME European Congress Egmond aan Zee	Oral presentation	2017	0,5
NCPA symposium Nijmegen	Workshop	2017	0,5

ISBN 978 94 6295 785 5