Contributions from Science Education Research 4

Jens Dolin
Robert Evans *Editors*

Transforming Assessment

Through an Interplay Between Practice, Research and Policy

Springer

Contributions from Science Education Research

Volume 4

More information about this series at http://www.springer.com/series/11180

Jens Dolin • Robert Evans
Editors

Transforming Assessment

Through an Interplay Between Practice, Research and Policy

Springer

Editors

Jens Dolin
Department of Science Education
University of Copenhagen
Copenhagen, Denmark

Robert Evans
Department of Science Education
University of Copenhagen
Copenhagen, Denmark

ISSN 2213-3623 ISSN 2213-3631 (electronic)
Contributions from Science Education Research
ISBN 978-3-319-63247-6 ISBN 978-3-319-63248-3 (eBook)
DOI 10.1007/978-3-319-63248-3

Library of Congress Control Number: 2017953193

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Acknowledgements

For those of us who worked on the ASSIST-ME EU research project for 4 years, this book captures some of the nuances of our efforts as well as our shareable results. It is a final product of the exceptional cooperation between many researchers and authors, both those who have contributed to the 11 chapters and those who participated in the research. We are particularly grateful to the chapter editors who not only managed their own chapters but had substantial input into the entire book including useful reviews of chapters and participating in editorial meetings. They all (except Pasi) also figure as first authors of the chapters they have been editing:

Christine Harrison, Florence Le Hebel, Monika Holmeier, Jan Alexis Nielsen, Pasi Nieminen, Mathias Ropohl and Silke Rönnebeck.

We are privileged to have worked with these talented colleagues from the science education research community.

We also extend our thanks to the 22 scholars who reviewed these chapters and made valuable suggestions for their improvement.

Contents

Part III General Reflections

About the Editors and Contributors

Editors

Jens Dolin is emeritus professor of science education at the Department of Science Education, Faculty of Science, University of Copenhagen. He has established and been head of the department for 8 years. His research areas are teaching and learning science (with focus on the assessment and development of competences) and organisational change (reform processes, curriculum development and teacher education and professional development). He has participated in and been leading a number of Danish and international science education research projects (about inquiry-based science education and assessment of science competences) and is member of a number of Danish and international boards and organisations. He was coordinator of the ASSIST-ME project.

Robert Evans is an associate professor in the Department of Science Education at the University of Copenhagen. He regularly teaches courses which focus on science teaching and learning at both secondary school and university levels. His recent research interests include self-efficacy as a capacity belief and how beliefs influence teaching and learning success. He has a long-term interest in promoting inquiry-based science teaching and in using formative assessment to facilitate inquiry teaching for both students and faculty by providing frequent feedback on progress for both groups. He is professionally active as an ESERA board member with the conference portfolio. He frequently conducts workshops at ESERA summer schools for PhD students about research design and analysis.

First Authors: Chapter Editors

Christine Harrison is reader in science education at King's College London. She taught science in UK schools for 13 years and was then recruited by King's College London, to run the science pre-service teaching course. With Paul Black and Dylan Wiliam, she co-directed the KMOFAP study which looked at assessment for learning in science, English and mathematics classrooms. Christine Harrison is known, both nationally and internationally, for the ways she can relate research to practice, made possible through her extensive experience of teacher education from both a collaborative action research and teaching perspective.

Florence Le Hebel is an associate professor in geosciences and science education at the University of Lyon. She is affiliated with ICAR and LLE (Ecole Normale Supérieure de Lyon). Her main research interest focuses on students' scientific literacy and understanding processes. She has participated in international research projects and FP7 projects like S-TEAM and ASSIST-ME projects.

Monika Holmeier obtained her PhD in Educational Science from the University of Zurich in 2012. She is currently working as a senior researcher at the Centre for Science and Technology Education of the University of Applied Sciences and Arts Northwestern Switzerland. Her research focuses on school governance, quality of instruction and assessment. Besides giving seminars on qualitative and quantitative research methods, she leads several research and evaluation projects. She is the author of the book *Grading in Statewide Exit Exams* (2013) and has published relevant articles in national and international academic books and journals.

Jan Alexis Nielsen is an associate professor at the Department of Science Education, University of Copenhagen. His research in science education has focused on understanding students' acquisition of complex competences – i.e. competences that are transferable to a wide variety of situations – and how such competences can be assessed. For example, he has studied students' argumentation, innovation competences and inquiry competences.

Mathias Ropohl, PhD is assistant professor of chemistry education at IPN. His main research interests are the formative and summative assessment of students' competences in chemistry at lower and upper secondary level as well as the development of in-service teacher candidates' and in-service teachers' professional competences. Recently, he focuses on teachers' use of traditional and digital media in order to support student learning. He has experience in EU projects from the project ASSIST-ME. Methodologically, he aims at conducting intervention studies that help to understand the effects of teaching and learning tools on teachers' and students' learning.

Silke Rönnebeck, PhD has been working as a research scientist at the University of Kiel and the Leibniz Institute for Science and Mathematics Education (IPN) since 2004. Her research interests are located at the intersection of assessment, science learning and teacher professional development and include the assessment of scientific competencies in international large-scale assessments (PISA), formative assessment, inquiry-based learning and the research-based development of professional development activities for science teachers.

Contributors

Paul Black is professor emeritus of education at King's College London. He has made many contributions to research into learning and assessment. He has served on advisory groups of the US National Research Council and has received a lifetime service award from the US National Association for Research in Science Teaching. His work on formative assessment with Dylan Wiliam and others has had widespread impact.

Jesper Bruun, PhD associate professor at the Department of Science Education, University of Copenhagen, works with integrating network theory with science and physics education research. In this work, he uses both qualitative and quantitative methods.

Rose Clesham, PhD is an assessment expert, designing and evaluating high- and low-stakes assessments up to degree level, both nationally and internationally. Her work includes research into quality-, reliability-, validity- and standards-related matters, ensuring that the evaluation and design of assessments are robust and evidence based.

Costas P. Constantinou is a Professor of Science and Education at the University of Cyprus where he directs the Learning in Science Group. He holds bachelor's, master's and PhD degrees in Physics from the University of Cambridge in the UK. He has written extensively on inquiry-oriented science teaching and learning and evidence-informed and research-based curriculum design, especially in the context of energy education and modelling-based learning. He is currently serving as president of the European Science Education Research Association (ESERA).

Catarina F. Correia worked for nearly a decade as a researcher and teaching fellow in chemistry. In 2012, she changed her research focus to science education. Two years later, she joined King's College London as a research associate at the School of Education, Communication and Society. Her research interests are in classroom assessment, inquiry-based pedagogy, curriculum development and teacher education.

David Cross is associate professor in science education at the University of Montpellier. He carries out research on science teachers' pedagogical content knowledge (PCK) at primary and secondary level from video data in order to understand the link between teachers' knowledge and practice.

Justin Dillon is professor of science and environmental education at the University of Bristol where he was head of the Graduate School of Education from 2014 to 2017. He teaches on initial teacher education courses and master's courses with a focus on teacher development and education management. Justin co-edits the *International Journal of Science Education* and was president of the European Science Education Research Association from 2007 to 2011. His research interests focus on the convergence of science and environmental education.

Michel Grangeat is emeritus professor of educational science at the University of Grenoble Alpes in France. His research focuses on teacher knowledge development in collective settings, mostly in science. As part of the European expert group on science education, he contributed to the report "Science Education for Responsible Citizenship" recently published by the European Commission. He has supervised many research projects at the Educational Sciences Laboratory of Grenoble. Before his commitment in higher education, he was a primary school teacher and a teacher educator for 20 years.

Regula Grob holds a master's degree in Earth science and worked as a geography teacher at upper secondary school for 4 years. She is now a PhD candidate in the Centre for Science and Technology Education at the University of Applied Sciences and Arts Northwestern Switzerland. Her doctorate focuses on formative assessment in inquiry-based science education.

Wynne Harlen has been involved in teaching, research and curriculum and assessment reform in science throughout her long career. She has held several posts, including head of education at Liverpool University and director of the Scottish Council for Research in Education. She was a founding member of the influential Assessment Reform Group and chair of the science expert group for PISA during its first 6 years and continues to play a significant role in the work of the IAP (Global Network of Science Academies) Science Education Programme.

Alena Hošpesová, PhD is associated professor at the Faculty of Education, University of South Bohemia in České Budějovice. In pedagogical work, she focuses on teacher training for primary school and pre-primary education, especially on didactics of mathematics. Her main research interest is focused on (mathematics) teachers' education and professional development.

Markus Hähkiöniemi, PhD is university lecturer in the Department of Teacher Education, University of Jyväskylä, Finland. He is interested in argumentation, classroom interaction and inquiry-based teaching in mathematics.

Sofie Birch Jensen is a PhD student at the School of Education, Communication and Society, King's College London. In her research, which focuses on formative assessment in the English upper secondary physics classroom, she is interested in how different assessment practices shape students' identity negotiations, particularly in relation to gender.

Doris Jorde is professor in science education at the University of Oslo and currently leader of the Norwegian Centre for Professional Learning in Teacher Education. Her research interests include classroom studies and science resource development. She is past president of ESERA and has participated in multiple EU projects in science in society.

Peter Labudde since 2008, is head of the Research Centre for Science and Technology Education, a group of 20 researchers, at the University of Applied Sciences and Arts Northwestern Switzerland. He has a PhD in applied physics, a MAS teacher diploma for upper secondary school and a habilitation in science education.

Laurent Lima is associate professor at the University of Grenoble Alpes since 2010 and head of the Department of Educational Sciences since 2016. He is an expert for the French Observatory of Student Life and for the Department of Evaluation, Prospective and Planning of the French Ministry of Education. His main research themes are efficacy of teaching approaches and reading learning processes in primary and higher education.

Michalis Livitzis, PhD in Science Education and holds a bachelor's degree in Physics from the University of Patras in 2005 and a master's degree in Micro-optoelectronics from the University of Crete in 2009. He is currently working as special teaching staff at the Department of Education, University of Cyprus. Michalis' research interests fall within the area of professional development of teachers regarding applying formative assessment methods when teaching science.

Pascale Montpied is in charge of research at the National Center for Scientific Research (CNRS). Her current research focuses on assessment of students' scientific literacy and students' motivations to learn on issues involving sciences, scientists and society. She has participated in international research projects and FP7 projects like S-TEAM and ASSIST-ME projects.

Marianne Moulin was involved in the ASSIST-ME project during 2 years after her PhD in mathematics education. She is now a mathematics educator for school teachers and focuses her research on language and verbal interactions in teaching and learning.

Nadia Nakhili, PhD is lecturer in education science at the University of Grenoble Alpes (UGA). Her research focuses on inequality of opportunity in educational systems. She uses both sociological perspective and comparative analysis of educational systems to understand the relevance of micro (teaching practices), meso (school practices and school mix effect) and macro (educational policies) factors that could enhance learning outcomes and reduce social inequalities in education.

Sanne Schnell Nielsen is a PhD student at the Department of Science Education, University of Copenhagen, and associate professor at University College UCC, Denmark. Her research focuses on science teachers' use of formative assessment with respect to competence and goal-targeted curriculums for compulsory school.

Pasi Nieminen, PhD is a postdoctoral researcher at the University of Jyväskylä, Finland. He has studied argumentation, assessment, inquiry-based teaching and multiple representations in physics.

Christopher Olley is currently a visiting lecturer at King's College London and mathematics education consultant. He was previously director of the pre-service teacher education course in mathematics at King's and has worked as department head and teacher in schools and colleges in London and East Africa. His research work has principally been on teachers' engagement with their subject knowledge, the use of technology in mathematics education and mathematical modelling in schools.

Nikos Papadouris is a postdoctoral research associate with the Learning in Science Group at the University of Cyprus. He has a PhD in science education from the University of Cyprus. His research interests concentrate on the design of teaching/learning materials that integrate epistemic awareness, reasoning skills and conceptual understanding.

Jan Petr is assistant professor at the Department of Biology, Faculty of Education, University of South Bohemia in České Budějovice. He is oriented on two fields of interest: (1) ecology and faunistic research of freshwater insects (dragonflies, Odonata) and (2) didactics of science education at preschool, primary and secondary levels. He has great experience in the field of cooperation with in-service teachers.

Elie Rached, PhD in science education, is a temporary associate professor at the University of Nice Sophia Antipolis, where he teaches Evaluation – Competencies Approaches in the graduate school for pre-service and in-service primary- and secondary-level teachers. He is a member of the Laboratoire d'Innovation et Numérique pour l'Education (LINE) of the University of Nice, where he conducts research on argumentation and decision making regarding socio-scientific issues. Also, he's an associate researcher at Laboratoire des Sciences de l'Education, the University of Grenoble Alpes.

Lukáš Rokos is an Assistant Professor at the Department of Biology, Faculty of Education, University of South Bohemia in Ceske Budejovice, Czech Republic. He is focused on the formative assessment with accent on peer assessment within the inquiry biology lessons, on implementation of inquiry-based education (IBE) into human physiology lessons and on IBE efficacy with comparison to traditional teaching methods.

Natasha Serret is senior lecturer at Nottingham Trent University, UK, where she leads primary science on several teacher education courses. Prior to this, she worked as a senior researcher at King's College London for 10 years, working on thinking skills in science (Cognitive Acceleration Through Science Education) and teacher assessment (King's Oxfordshire Summative Assessment Project). She was a post-doc researcher for the King's College ASSIST-ME research team.

Kay Stables is emeritus professor of design education at Goldsmiths, University of London. A founding member of the Technology Education Research Unit, she has directed and contributed to research projects with focus on design, creativity and sustainable development, digital tools in assessment (the E-scape Project) and designerly well-being. Current research includes the development of design capability in digital environments through the use of on-screen avatars.

Iva Stuchlíková is Professor of Educational Psychology at the Department of Pedagogy and Psychology, Faculty of Education, University of South Bohemia in České Budějovice. Her main research is concentrated on motivation and emotions, science education and teachers' professional development. Her recent research projects are related to inquiry-based science education, assessment and emotions in schools.

Andrée Tiberghien is emeritus "directeur de recherche au CNRS" in science education with more than 30 years of experience in research including collaboration with science teachers and production of teaching resources. Currently she studies science classroom practices. She has participated in international research projects and FP7 projects like S-TEAM, Mind the Gap and ASSIST-ME. She was a member of the science expert group of PISA 2006, 2009 and 2015.

Sofie Tidemand is a PhD student at the Department of Science Education, University of Copenhagen. Her research focuses on assessment of generic competences as well as aspects related to socio-scientific issues. She is investigating science teachers' work with and challenges related to formative assessment and learning goal-oriented teaching.

Olia Tsivitanidou holds a PhD in science education and is currently working as a postdoctoral researcher at the University of Cyprus. Her research interests include (web-based) reciprocal peer-assessment processes in (computer) supported learning environments, inquiry-oriented science learning and teaching and science communication.

Jouni Viiri is professor of mathematics and science education at the University of Jyväskylä, Finland. He has taught physics in different educational levels. His research interests include the relation of classroom interactions to students' learning, use of technology in education and educational research, e.g. use of eye tracking in multiple representation studies.

Iva Žlábková, PhD is Assistant Professor of the Department of Pedagogy and Psychology, Faculty of Education, University of South Bohemia in České Budějovice. In pedagogical work, she focuses on teacher training for primary and lower secondary schools, especially on general didactics. Her main research interest is focused on assessment, teachers' education and professional development.

Introduction

Over the last decade, inquiry-based learning has been encouraged in schools across Europe, but one of the factors inhibiting the successful implementation of this has been the lack of focus on assessment issues. All educational systems have witnessed a growing gap between the demands for new student competences like inquiry, innovation, creativity and communication – and the ability of the assessment methods and the assessment systems to capture these new educational goals.

This book is about a project that has taken up this challenge. "Assess Inquiry in Science, Technology and Mathematics Education" (ASSIST-ME) is a research project that has investigated assessment methods aimed at supporting and improving inquiry-based approaches in European science, technology and mathematics (STM) education. The ASSIST-ME project responded to a call from the European Union's research and innovation funding programme for 2007–2013, the so-called FP7 programme. The call recognised that merely developing new assessment items was not sufficient; it also had to be a political priority to reform the educational system to give room for formative processes in a system with strong emphasis on summative assessments. Thus, the FP7 call contained an explicit demand to enter into the political world. Consequently, the project also had a societal impact in using the research to give policy makers and other stakeholders guidelines for ensuring that assessment enhances learning in STM education. Even if the call did not include E for engineering, the project included E as a natural part of the technology subjects and in order to be in alignment with the mainstream focus on STEM educations.

As a research project, the work was driven by formulated research questions. These were (with some of the sub-questions):

- What are the main challenges related to the uptake of formative assessment in the daily teaching-learning practices in science, technology and mathematics within primary and secondary schools in different European educational systems?

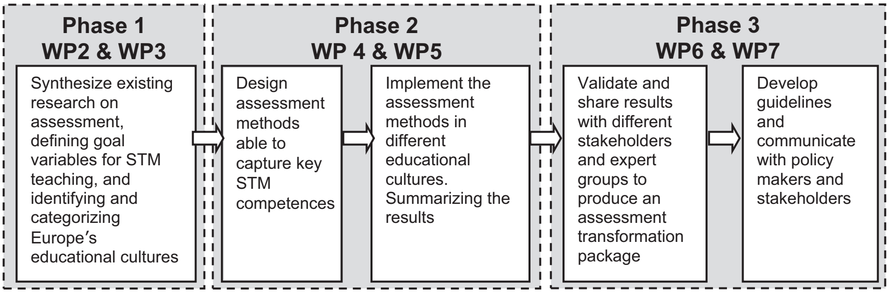
  What systemic support measures and what tools do teachers need in order to integrate formative assessment of student learning in their classroom practices?

- What changes are needed in summative assessment practices to be consistent with the learning goals of IBE within STM?
- How can formative and summative assessment methods be used interactively to promote learning in inquiry-based STM?
- How can research-based strategies for the use of formative/summative assessment be adapted to various European educational traditions to ensure their effective use and avoid undesirable consequences?
- How can relevant stakeholders be invited to take co-ownership of the research results, and how can a partnership between researchers, policy makers and teachers be established in order to secure relevant actions following meaningful and effective implementation guidelines?

The project lasted from January 2013 to December 2016 and had 10 partners from 8 countries:

University of Copenhagen, Department of Science Education (Coordinator).
University of Kiel, Leibniz Institute for Science and Mathematics Education.
University of Cyprus, Department of Educational Sciences, Learning in Science Group.
Fachhochschule Nordwestschweiz, Pädagogische Hochschule, Centre for Science and Technology Education.
Centre National de la Recherche Scientifique, Lyon, ICAR, ENS Lyon.
King’s College London, Department of Education and Professional Studies.
University of Jyväskylä, Department of Teacher Education.
University of Grenoble Alpes, Laboratoire des Sciences de l’Éducation.
University of South Bohemia, Department of Pedagogy and Psychology.
Pearson Education International.

This book is not a tight summary of the project but slightly arranged for reader comprehension into three sections. But as the chapters follow the same logic that formed the project, it is useful to have an overview of the different phases of ASSIST-ME:


Phase 1
WP2 & WP3
Synthesize existing research on assessment, defining goal variables for STM teaching, and identifying and categorizing Europe’s educational cultures

Phase 2
WP 4 & WP5
Design assessment methods able to capture key STM competences
Implement the assessment methods in different educational cultures. Summarizing the results

Phase 3
WP6 & WP7
Validate and share results with different stakeholders and expert groups to produce an assessment transformation package
Develop guidelines and communicate with policy makers and stakeholders

In the first phase, partners built a knowledge base for the rest of the project, and the first section of the book introduces this foundation. The concept of competence as a common goal for STM education is analysed and compared to other goals like literacy and the central European concept of bildung. Commonalities and differences of the concept across the three domains are mapped, and twenty-first-century goals such as innovation competence are defined. The project decided to focus on three subject-specific competences: *scientific inquiry* within science education, *engineering design* within engineering education and *problem solving* within mathematics education. It also worked with three cross-disciplinary, generic competences: *argumentation*, *modelling* and *innovation*. This work is described in Chap. 1.

Then, since these new competence-oriented goals can only be sustainably implemented if they are aligned with teaching and assessment methods, Chap. 2 shows how inquiry-based education supports the achievement of competencies. It defines the six chosen competences and outlines how the implementation of these competences in the classroom affects teaching and the changes in assessment necessary in order to be able to assess them.

Understanding assessment for both formative and summative purposes has been a key aspect of ASSIST-ME, and Chap. 3 puts together the achieved insights. The chapter offers a model for how the two purposes of assessment are defined and what their main characteristics are. It also discusses two different approaches to linking formative and summative assessment and reports how teachers involved in ASSIST-ME dealt with the challenges related to combining and connecting the two purposes.

The second section consists of four chapters, each of which introduces and explores one of the formative assessment methods selected for study in the ASSIST-ME project: on-the-fly assessment (Chap. 4), structured assessment dialogue (Chap. 5), peer assessment (Chap. 6) and written feedback (Chap. 7). In each chapter, the basis for use of the assessment method is explained along with how it was employed and what the project learned about the opportunities, dilemmas and constraints of use. These four assessment methods were implemented in classrooms in the partner countries by teachers who worked together with researchers in so-called local working groups (LWGs). An LWG typically consisted of around 25 teachers and 2–4 researchers from the partner institution. The teachers were mostly experienced teachers, and often they had previously collaborated with the partner institution. A research design for each assessment method made it possible to reliably collect and analyse data across partner countries related to the research questions. As a supplement to the common research questions and methods, each LWG formulated research questions relevant for the national context but within the common framework.

The work was organised as action research and took place over three semesters. The teachers were financially compensated for their meeting time. EU doesn't pay teachers' teaching, but as the meetings could be argued as research time, compensation was possible. Each LWG decided which assessment method and which competence they would work with in each of the semesters. During four meetings in each semester, teachers together with researchers prepared and monitored the implementation. These collaborations could typically involve designing templates with progression steps for a competence and criteria for each progression step. The LWG also organised the data collection, mostly done by the teachers following the predesigned research design. At each meeting, experiences were exchanged and teachers participated in focused discussions of common questions for all partners. All data was collected and centrally stored. It was analysed on a general level by one partner and shared among all partners. Detailed and specific analyses were then possible by individual partners or groups of partners.

The final section of the book extracts some general reflections and outcomes from ASSIST-ME. In Chap. 8, the educational and assessment systems of the eight European countries considered by the project are analysed to reveal similarities and differences relevant to formative and summative assessment. The purpose is to adapt the findings from the implementation processes to the national contexts and to synthesise the findings to formulate recommendations at a European level. Then, in Chap. 9, teacher perspectives among the partner countries are shared with a focus on changes in teacher self-efficacies and teacher subjective theories while implementing formative assessment. The chapter also examines the use of a specially developed Internet-based platform to aid in inquiry-based formative assessment.

The second main goal of ASSIST-ME – to inform and influence policy makers to initiate educational changes in a direction that allows teachers to a more formative use of assessment instead of the dominating summative use – is reported in Chap. 10. It was a key aspect of the project to put together the three main players in the changing process: the teachers, the researchers and the policy makers. The collaboration with policy makers was managed through the formation of national stakeholder panels (NSPs) with representatives from industry, ministry, heads' association, teachers' association, media, Parliament, trusts, etc. The chapter describes the selection process and the work of the NSPs. It also gives the recommendations for policy makers, curriculum developers, teacher trainers and other stakeholders in the different European educational systems on how to support this drive for inquiry teaching and learning.

Finally, Chap. 11 gives recommendations for future research. As a research project, ASSIST-ME produced a large number of results both within and across the eight participating partner countries using a variety of research methods. Chapter 11 organises, prioritises and summarises the principal outcomes relevant to the original research questions and describes the research methods behind the findings. The chapter identifies and outlines the current gaps in knowledge in research in assessment practice and ties the results of the ASSIST-ME project onto this outline. In this way, the chapter presents a number of concrete research vistas that are still needed in international assessment research.

All information from ASSIST-ME is accessible at the website, http://assistme.ku.dk/, where it is also possible to download a 16-page booklet giving an overview of the project.

Copenhagen, Denmark

Jens Dolin
Robert Evans

Part I
Background

Chapter 1
The Concept of Competence and Its Relevance for Science, Technology and Mathematics Education

Mathias Ropohl, Jan Alexis Nielsen, Christopher Olley, Silke Rönnebeck, and Kay Stables

Locating the Concept of Competence in Relation to Other Concepts

During the last decades, the conditions under which people live have changed rapidly. Advances in science and technology have influenced the way people live, and environmental problems have destroyed many people's livelihood in many regions of the world. Globalisation means events happening in places far away affecting people's everyday lives worldwide. Thus, people are confronted with new opportunities but also with new challenges and problems. Therefore, living in today's complex world requires adaptation to new conditions as well as lifelong learning. Consequently, educational systems have to respond to these societal, economic and ecological changes by defining new educational goals that are reflected in the concept of competence. Competences are demand and function oriented. In order to solve complex problems, people should be able to realise certain competences in a particular context and to transfer these competences to other contexts. They thereby provide one foundation for lifelong learning. However, looking at the literature, the

M. Ropohl (✉)
Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany
e-mail: ropohl@ipn.uni-kiel.de

J.A. Nielsen
Department of Science Education, University of Copenhagen, Copenhagen, Denmark

C. Olley
King's College London, London, UK

S. Rönnebeck
Kiel University, Kiel, Germany

Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

K. Stables
Goldsmiths, University of London, London, UK

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_1

concept of competence is not unambiguously defined and some confusion exists with respect to related constructs like Bildung or literacy. The reason is that the construct of Bildung is understood as a more ‘general formula for what is expected from (lifelong, not just school-based) learning processes [constituting] a precise description of the ability of subjects to act under the conditions of undecidability, indeterminacy, uncertainty and plurality’ (Klieme et al. 2004, p. 59). In comparison, competences have a stronger focus on what students should be able to do in order to, for example, compete in the labour market of a modern society:

> When scholars of educational science speak about the general goals of training within modern societies, they quarrel with finding a balance between [on the one hand] Bildung in the tradition of German philosophy, i.e. developing personality and allowing individuals to participate in human culture, and [on the other hand] qualification, i.e. establishing knowledge and skills that are relevant for vocational practice. (Klieme et al. 2008, p. 6)

In addition, the concept of competence is also closely linked to the concept of literacy as it is operationalised in international large-scale studies like the Programme for International Student Assessment (PISA). The PISA framework for scientific literacy emphasises the importance of the application of scientific knowledge and skills in the context of real-life situations. This application of knowledge and skills is influenced by students’ attitudes or dispositions towards science that can determine ‘their level of interest, sustain their engagement, and may motivate them to take action’ (OECD 2016, p. 20). In a similar way, the definition of mathematical literacy in PISA strongly emphasises the need to develop students’ ‘capacity to formulate, employ and interpret mathematics in a variety of contexts’ (OECD 2016, p. 65) that reflect the ‘range of situations in which individuals operate in the twenty-first century’ (OECD 2016, p. 73).

Rychen and Salganik (2003) even argue that the ‘convergence between the concept of literacy as defined in current assessment frameworks and DeSeCo’s [the OECD project ‘Definition and selection of competences’] concept of competence, and the difficulties associated with the definition of the term literacy, together suggest that international assessments would benefit from replacing the concept of literacy with the concept of competence’ (p. 53).

The Concept of Competence in Education

In many countries, a shift from an input- towards an output-orientation could be observed within educational systems (Waddington et al. 2007). Instead of relying solely on extensive content descriptions, educational standards define competences as learning goals for students at specific stages of their educational career (e.g. National Research Council 2012; Qualifications and Curriculum Authority 2007; Schecker and Parchmann 2007). Despite this importance, however, the definition of the construct has to some extent remained fuzzy within educational research (e.g. Blömeke et al. 2015; Koeppen et al. 2008; Weinert 2001). In his review of theory-grounded approaches to the concept of competence, Weinert (2001) found that no broadly accepted definition or unifying theory existed. Instead, a tendency could be observed 'to use terms such as skill, qualification, competence, and literacy, either imprecisely or interchangeably, in order to describe what individuals must learn, know, or be able to do in school, at the workplace, or in social situations (Rychen and Salganik 2003, p. 41).

In 1997, the Organisation for Economic Cooperation and Development (OECD) initiated the DeSeCo project with the aim of providing solid theoretical and conceptual foundations for the broad range of competences needed to face the challenges of the present and the future. They considered an explicit definition of the meaning and nature of competences crucial to enable a coherent discourse on competences from a lifelong learning perspective. In line with an earlier recommendation by Weinert (2001), the project decided on a demand- or function-oriented approach to define competences: 'A competence is the ability to successfully meet complex demands in a particular context through the mobilization of psychosocial prerequisites (including both cognitive and non-cognitive aspects)' (Rychen and Salganik 2003, p. 43). In order to avoid reducing competences to mere *ability-to* expressions, the demand-oriented approach requires the conceptualisation of competences as internal mental structures, i.e. resources embedded in the individual such as cognitive skills, intellectual abilities (e.g. critical thinking) and knowledge but also social and behavioural components such as motivation, emotion and values. Moreover, possessing a competence includes an action component. Individuals always operate in specific contexts that set the criteria for effective performance. It is thus not sufficient to possess the different resources but one must be able to mobilise and orchestrate them in a complex situation. This understanding of competence is consistent with the action competence model described by Weinert (2001) that represents a holistic and dynamic perspective by combining complex demands, psychosocial prerequisites and contexts into a complex system that people need in order to solve problems. It is necessary to convert this very general definition into more domain-specific definitions, e.g. what does it mean to be competent in science, technology or mathematics?

In educational contexts, the concept of competence usually refers to those context-specific dispositions for achievement that can be acquired through learning, in contrast to basic cognitive abilities that can only be learned and trained to a far lesser degree (Klieme et al. 2008). Understanding competences as reflecting a person's potential to meet cognitive demands in specific areas of learning and behaviour makes them amenable to external interventions such as opportunities to learn and systematic training, thus increasing the utility of the concept for teaching and learning as well as for the empirical assessment of educational outcomes (e.g. Klieme et al. 2008).

Despite the huge amount of research in the last decade however, a recent review still considers competence to be a ‘messy construct’ (Blömeke et al. 2015, p. 4). From the authors’ perspective, the main reason for this is that the concept has been ‘plagued by misleading dichotomies’ (p. 11). Interpretations of the definition of competence as ‘complex ability constructs that are context-specific […], and closely related to real life’ (Koeppen et al. 2008, p. 61) tend to focus either on the performance or the disposition aspect which leads to a dichotomy: is behaviour the focus of competence or is it the criterion against which dispositions are validated as measures of competence? Both approaches have their advantages and disadvantages. The first position describes a holistic perspective in which dispositions and performance are complexly linked and may change during the course of performance. In this understanding, it is of specific importance how precisely the different dispositions are linked and what influences their interaction. The second position takes a more analytical stance by dividing competence into multiple constituents. The major question is whether competence can be exhaustively decomposed into identifiable constituents. Sadler (2013) argues that while decomposition reduces complexity and provides highly visible learning goals, it becomes more difficult to see the whole. Moreover, decomposition can lend itself to seriously deficient implementation. Teachers might become encouraged to deliberately coach students over the pass lines for specific competences – however, this ‘does not necessarily translate into a coordinated ability to complete a complex task with proficiency’ (p. 17). According to Blömeke et al. (2015) this dichotomised discussion neglects the processes that connect dispositions and performance. They argue for regarding competence as a process, a continuum ‘from traits that underlie perception, interpretation, and decision-making skills, which in turn give rise to observed behaviour in real-world situations’ (p. 3).

Based on the prominence given to the concept of competence in international comparisons of educational outcomes and national standards, the assessment of competences has become a major focus in educational research. The valid assessment of competences is regarded as essential for the enhancement of educational processes and the development of educational systems. Assessments developed to measure competences have to meet specific requirements that differ from traditional knowledge tests. It has to be insured, for example, that the sampling of real-life situations is representative of the universe of tasks (Blömeke et al. 2015). The dichotomy between holistic and analytical views of competence, however, is also reflected in the assessment of competences (Blömeke et al. 2015).

The analytic view of competence assessment focuses on measuring different latent traits (cognitive, conative, affective and motivational) with different instruments (Blömeke et al. 2015). This requires the development of sophisticated models of the structure and levels of these constructs that precisely define them in specific domains. According to Klieme et al. (2008), the development of cognitive competence models faces two major challenges. The first challenge is related to the contextualised character of competences which means that individual- and situation-specific components have to be simultaneously considered. This leads to the distinction between two types of theoretical models to describe competences that are ideally complementary: models of competence levels, defining the specific situational demands that can be mastered by individuals with certain levels of competences, and models of competence structures – dealing with the relations between performances in different contexts and seeking to identify common underlying dimensions. The second challenge is related to the question of how competences develop. Only few models have addressed this developmental aspect and their conceptualisations differ. Whereas some models regard competence development as a continuous progression from the lowest to the highest competence level, others conceptualise it as a noncontinuous process characterised by qualitative leaps. Related to the question of how competences develop is the question of to what extent such developmental models can represent cognitive processes (Leuders and Sodian 2013). Once theoretical competence models have been developed, they need to be linked to the results of empirical assessment by psychometric models. Here again the contextualised and complex nature of the competence construct defines certain requirements. The models need to incorporate all relevant characteristics of the individuals whose competences are to be evaluated while at the same time they have to take domain-specific situational demands into account. The holistic perspective on competence assessment focuses instead on assessing real-life performance without accounting for the contribution of specific dispositional resources (e.g. Shavelson 2010). However, this approach also has specific challenges. Performance tasks are often time-consuming and introduce considerable amounts of measurement error due to their complexity. Nevertheless, recent examples show that the approach is not impossible (e.g. Theyßen et al. 2014). Especially in large-scale assessments, simulation-based test instruments have been shown to provide potential in this context (OECD 2016; Theyßen et al. 2016).

In their review, Blömecke and colleagues (2015) argue that moving the field forward is not a question of choice between the analytical or holistic approaches but rather of finding ways to productively combine them (cf. Grugeon-Allys et al. 2016), thus moving ‘beyond dichotomies’ (p. 9). As stated above, in educational practice and research, competences usually relate to specific content areas (Koeppen et al. 2008), so-called domains. Typical domain-specific competences in primary and secondary education include scientific competence, technological competence and mathematical competence. These domain-specific competences are described in the next three sections. In addition, transversal competences often referred to as key competences are becoming more and more important for the participation of individuals in society and in the workplace. This development is picked up in the fourth section on innovation competence.

Competences in Science Education

In science education many educational standards and curricula are based on the concept of competence or related to it. The educational standards in many countries define specific competences as learning outcomes at a certain educational level (e.g. National Research Council 2012; Qualifications and Curriculum Authority 2007). However, the exact definition of these learning outcomes might differ from country to country (cf. Waddington et al. 2007).

From the perspective of science education, competent students have to solve specific types of problems and have to deal with certain kinds of concrete situations relevant for science. More specifically, students have to detach science-specific cognitive skills and knowledge from one situation and apply it to scientific problems in another situation, e.g. a social setting (Kauertz et al., 2012). A typical classroom situation where students are faced with scientific phenomena and problems that are explored or investigated through inquiry is described in the following box.

Starting question given by the teacher:

Usually when I am taking my effervescent Vitamin C drink I wait until it [has] stopped fizzing before I drink it. Some mornings I am running late for school. Can I speed up this reaction?

Equipment students receive:

Vitamin C effervescent tablets, boiling tubes, test tubes, 250 ml beakers, 100 ml measuring cylinders, stop watches, […].

Problems students have to solve:

1. Which variable should be changed?
2. How to measure the effect of the change?

(Black and Harrisson 2016, p. 25; see also http://results.sails-project.eu/units/reaction-rates).

A more general understanding of being competent in science was described by the OECD in the context of international large-scale studies in terms of scientific literacy. Within PISA, scientific literacy is defined as 'the ability to engage with science-related issues, and with the ideas of science, as a reflective citizen. A scientifically literate person is willing to engage in reasoned discourse about science and technology which requires the competencies to: 1. Explain phenomena scientifically […], (2) Evaluate and design scientific enquiry […], and (3) Interpret data and evidence scientifically' (OECD 2016, p. 20).

Students need to apply these competences in different situations reflecting personal, local, national and global contexts which in turn require them to draw on their scientific knowledge and apply it in the context of life situations (OECD 2016).

Relating this back to science teaching and learning, effective learning strategies should include 'using a wide range of contexts, inductive rather than deductive processes, problem-based learning contexts in which problems are integrated rather than broken into discrete, artificial elements, and encouragement of self-directed learning and self-reflection on learning styles' (Rychen and Salganik 2003, p. 58).

Although some disagreement with respect to the definition of scientific literacy has been found (Bybee 1997; DeBoer 2000; Roberts 2007), the different definitions show some common ground, namely, 'that scientific literacy usually implied a broad and functional understanding of science for general education purposes and not preparation for specific scientific and technical careers' (DeBoer 2000, p. 594). In this sense, scientific literacy is closely related to science for *all* and can be seen as a precondition for participating in a society that is shaped by science and technology (DeBoer 2000). The objective of these two concepts is the constitution of 'a common core of learning in science for all young people, regardless of their social circumstances and career aspirations' (American Association for the Advancement of Science 1989, 1990).

But, what are these types of problems and concrete situations that students face in science education? To answer this question one has to look at teaching and learning situations as well as at assessment situations. For both types of situations, examples have been published. For an overview, it is worthwhile to have a look at the PISA framework that defines relevant contexts for its assessment items: 'health and disease, natural resources, environmental quality, hazards, and the frontiers of science and technology' (OECD 2016, p. 24). These contexts can also serve as a frame for teaching and learning in school (Gilbert 2006). Science subjects like chemistry have to emphasise the relevance of the taught content and make this relevance explicit to the students (Bulte et al. 2006). It is assumed that this might help students understand the contribution of science to their future lives. The approach of contextualising teaching and learning has been picked up by several EU-funded development and research projects like ASSIST-ME (assistme.ku.dk/practical-examples/swiss-examples/), Establish (www.establish-fp7.eu), and mascil (www.mascil-project.eu).

Within the ASSIST-ME project, the different partners were asked in a nonrepresentative survey to define the concept of competence from the perspective of their respective countries. From the answers it becomes obvious that in all countries competences are seen as something each individual possesses and that have to be applied to specific situations in order to solve problems. Prerequisites for being competent in a specific domain are cognitive constructs like knowledge and skills as well as affective constructs like interest or motivation. In the following, some key quotations from the survey are shown in order to reflect the understanding of competence in some of the countries:

- 'The ability to perform a task that requires knowledge and skills is an outcome of a competence'. (Cyprus)
- 'Competence is such a development of person's knowledge, skills, attitudes, values and self-regulation in particular domain (e.g. problem solving, modelling, argumentation, empirical investigation), that make the person able to cope with

relevant challenges and demands, effectively solve the tasks, and creatively adjust oneself to domain relevant situations either in individual or the social perspective'. (Czech Republic)

- 'A person is scientifically competent when she/he has the ability and commitment to act, alone and together with others, in a way that takes advantage of scientific curiosity, knowledge, skills, strategies and meta-knowledge to create meaning and autonomy and exert codetermination in relevant life contexts'. (Denmark)
- 'In the national standards [of Switzerland] a competence includes both a skill (German: Handlungsaspekt, French: aspects de compétences) and a content (German: Themenbereiche, French: domaines thématiques). That is, a competence is determined by both a skill and a content'. (Switzerland)
- 'It can be defined as a combination of basic knowledge relevant to live in our society, capacities to involve them in diverse situations but also lifelong necessary attitudes as to be open to others, the appetite for search for truth, the respect of oneself and others, the curiosity and the creativity'. (France)

Competences in science education are manifold and often no concurrent or well-accepted definition of competences exists (cf. Rönnebeck et al. 2016). That's why the issue of measuring competences is still one of the driving questions of research and practice in science education (e.g. Gitomer and Duschl 1998; Harlen 2013). There is an ongoing debate on the validity of standardised and teachers' assessments (e.g. Black et al., 2010), the influence of assessment practices on daily teaching (Binkley et al. 2012; Cizek 2001) and the advantages and disadvantages of specific item formats (e.g. Haladyna et al. 2002). In view of the assessment of scientific competences, especially validity is of high importance: does the test measure the competence, e.g. planning an investigation that it aims to measure? Here it becomes obvious why a precise definition of the assessed construct is a prerequisite for each assessment. It is much easier to develop a test that is valid with regard to the competence of interest if a precise definition exists and if the behaviour students should show when planning an investigation is well defined.

Competences in Technology Education

The discussion in the introduction to this chapter on the relationship between the terms *literacy*, *Bildung* and *competence* is an interesting one in the context of technology education as it resonates with debates within technology education where the linguistic concepts and their definitions are not universally shared or understood. Different definitions of terminology across national contexts mean that there is no universal agreement about the terms or their use within technology education curricula. To further complicate matters, a fourth term is also in common use, that of *capability*. To understand the different perspectives, it is useful to start with the debate between the terms literacy and capability, as these provide two distinct approaches.

Technological literacy has been at the core of developments in technology education in many parts of the world, arguably spearheaded by the International Technology and Engineering Association (International Technology Education Association 2007). Technological literacy is defined as:

> … the ability to use, manage, assess and understand technology. A technologically literate person understands, in increasingly sophisticated ways that evolve over time, what technology is, how it is created and how it shapes society, and in turn is shaped by society. [...] A technologically literate person will be comfortable with and objective about technology, neither scared of it nor infatuated with it. (International Technology Education Association 2007, pp. 9–10)

An alternative position critiques the literacy definition as denying the importance of action: a capability perspective. Led by UK developments that argue for the need to develop agency through taking action, capability is about creating intentional change and improvement, through intervention in response to a need, want or opportunity. This concept has been at the core of the English National Curriculum for Design and Technology since its creation in 1990, the essence of which is captured in the following extracts from a national curriculum statement:

> Design and technology prepares pupils to participate in tomorrow's rapidly changing technologies. They learn to think and intervene creatively to improve quality of life. The subject calls for pupils to become autonomous and creative problem solvers, as individuals and members of a team. [...] Through design and technology, all pupils can become discriminating and informed users of products, and become innovators. (Departmentment for Education and Employment/Qualifications and Curriculum Authority 1999, p. 15)

Critical terms here are *intervene* and *become innovators*. Capability is an active, rather than passive mode, having resonance with Sen's (1992) capabilities perspective that focuses on how a person functions – their beliefs and their actions. Critical in this is a sense of agency and, from a technology education perspective, capability that is manifest in a learner who can see a technological challenge that needs addressing and who has the confidence and competence to successfully intervene to create improvement.

The definition of competence, as outlined at the beginning of this chapter, relates to literacy as well as capability perspectives as both place strong emphasis on the procedural aspects of technology education: the processes of designing. But the term competence is not routinely applied across different national contexts. The increased global emphasis on core or key competences and their link to twenty-first-century skills has been of mixed value for technology education. As something that has been seen as important for high-stakes assessment in core subjects, there is evidence that technology education, not being seen as a core subject, is hence deprioritised in the curriculum (International Technology Education Association 2007). There are indications that implementing key competences has allowed technology education to be identified with more industry-focused competences such as problem-solving, rather than recognising its broader contribution to more generic competences (Ritz and Reed 2006; Williams 2006). Some national curricula have identified competences as ways of assessing standards and levels of attainment and have done this consistently, New Zealand's technology education curriculum being an example of this. Some curricula have used the term in the past but have then dropped it, for example, Sweden, where competences were highlighted in 2000 but are not mentioned in 2010, *ability* having taken their place (Skolverket 2009, 2012). But many national curricula for technology education make no mention of competences at all.

In the spirit of the key features of competence outlined in the introduction to this chapter, the consistent aspect of technology education that relates to competence is found in the procedural nature of the curriculum. There is a common pattern to the ways that technology education programmes are structured in different national contexts through three dimensions: the knowledge and skills base, the societal context and the processes of designing that enable technological developments. The essence of the subject lies in the ways in which these three dimensions interact and interplay through a common learning and teaching approach of project-based learning. Historically, the major focus in the forerunners of technology education was the development of craft skills, but a major change occurred in the late 1960s/early 1970s with a shift in focus from curriculum content to process. The shift occurred through innovations in assessment as early models of design processes were introduced as assessment criteria (Kimbell 1997). The initial developments were in the UK, but the approach has spread incrementally across the globe, and now having processes of designing at the core of learning, teaching and assessment is ubiquitous, and these processes lie at the heart of competence in technology education. However, competence does not reside in process alone; competence is developed and evidenced in the ways in which processes draw on knowledge, skills and understandings and how these are used within societal contexts.

The context is an important dimension in technology education, where the context of a task, or project, relates to societal needs (Kimbell and Stables 2007; Stables 2013). Design and technology contexts are situations in which challenges are embedded and provide the background to the people, places and purposes at the heart of the challenge. Good design and technological challenges, from a learning and teaching perspective, are rich in issues, competing priorities and conflicting values. One aspect of competence, therefore, is shown in ways that learners engage with a context, for example, by researching the types of stakeholders involved and using their research as a resource for understanding the needs in their challenge and in evaluating their approaches to address these needs. Knowledge and skill are critical, not for their own sake, but for the ways in which they are drawn on, learnt and developed based on the needs in the task. Thus, another dimension of competence is evidenced through a learner being able to draw appropriately on his or her existing knowledge and skills, recognise what new knowledge and skills are needed and know how to acquire them. Layton (1993) referred to this as seeing knowledge as a quarry to be exploited, not a cathedral to be worshipped. But the reality is that the quarry itself is difficult to define. Some knowledge can be easily identified as core to technology education, such as the properties of materials. But if a learner is designing a learning aid, for example, for a child with cerebral palsy, then knowledge of cerebral palsy is needed. Basically, any knowledge can become technological knowledge if it is needed for the task in hand. In relation to the content of a technology curriculum, the subject has been called 'a restive, itinerant, non-discipline' (Kimbell and Perry 2001), something that can present significant challenges for teachers.

The interlacing of the societal context of technological activities, the processes of designing and the knowledge and skills base that underpins competence in technology education has resulted in an inevitable shift away from atomistic approaches to teaching and assessment. While the extent of this differs, curriculum descriptors that integrate these dimensions are evident across national curricula documentation, as illustrated in the following extracts:

> Students create designed solutions for each of the prescribed technologies contexts based on an evaluation of needs or opportunities. They develop criteria for success, including sustainability considerations, and use these to judge the suitability of their ideas and designed solutions and processes. They create and adapt design ideas, make considered decisions and communicate to different audiences using appropriate technical terms and a range of technologies and graphical representation techniques. Students apply project management skills to document and use project plans to manage production processes. They independently and safely produce effective designed solutions for the intended purpose. (Australian Curriculum, Assessment and Reporting Authority 2016, p. 2)
>
> Students […] understand that all design and technological practice takes place within contexts which inform outcomes […] use different design strategies, such as collaboration, user-centred design and systems thinking, to generate initial ideas and avoid design fixation […] design and develop at least one prototype that responds to needs and/or wants and is fit for purpose, demonstrating functionality, aesthetics, marketability and consideration of innovation […] make informed and reasoned decisions, respond to feedback about their own prototypes (and existing products and systems) to identify the potential for further development and suggest how modifications could be made. (Department for Education 2015, p. 7)
>
> Students will: Critically analyse their own and others' outcomes and their determination of fitness for purpose in order to inform the development of ideas for feasible outcomes. Undertake a critical evaluation that is informed by ongoing experimentation and functional modelling, stakeholder feedback, trialling in the physical and social environments, and an understanding of the issue as it relates to the wider context. Use the information gained to select, justify, and develop an outcome. Evaluate this outcome's fitness for purpose against the brief. Justify the evaluation using feedback from stakeholders and demonstrating a critical understanding of the issue that takes account of all contextual dimensions. (Ministry of Education 2010, p. 77)

Supporting learners to achieve competence in technology education is complex. Research has created understandings of a range of issues and opportunities to support learning, teaching and assessment of procedural competence, including importance of iterative approaches to design processes (Kimbell et al. 1991; Kimbell and Stables 2007), approaches to assessment for learning (McLaren 2012; Moreland et al. 2008; Moreland 2009), the impact of education paradigms (Mioduser 2015), maintaining authenticity in learning and assessment (Turnbull 2002; Snape and Fox-Turnbull 2013; Stables 2013), the use and challenges of assessment portfolios including e-portfolios (Doppelt 2009; Kimbell et al. 1991, 2009; Seery et al. 2012; Stables and Kimbell 2000; Williams 2013) and the use of judgement, holistic and comparative, in assessment (Kimbell et al. 1991; Kimbell 2012). Common to all the above research is a desire to provide insights and solutions to enable teachers to take on the significant challenges of teaching and learning to develop a breadth of procedural competences through technology education, some of which will be explained in the following chapter of the book.

Competences in Mathematics Education

In mathematics education, competence describes mastery of the cognitive requirements for successful performance in the content area of school mathematics: 'To master mathematics means to possess mathematical competence' (Niss 2004, p. 6). Kilpatrick (2014) describes the structuring of mathematical competence as frameworks which 'in mathematics education fall primarily into Weinert's specialized-cognitive-competencies category' (p. 85) in contrast to general cognitive competences, characterised by psychometric models of intelligence. The specialised nature of mathematics education is described as having potentially two components: specific mental processes sometimes coupled with collections of content over which these processes will be deployed. This is a characteristic of mathematics curriculum specifications. For example, the PISA study sets out a framework consisting of content categories (quantity, uncertainty and data, change and relationships, space and shape) together with process categories describing the problem-solving process and a set of seven 'fundamental mathematical capabilities' (OECD 2014, pp. 38–39). Basing new mathematics curricula on a notion of mathematical competence, the Danish KOM project set out two groups of competences describing 'the ability to ask and answer questions in and with mathematics' and 'the ability to deal with and manage mathematical language and tools' (Niss 2004, pp. 7–8).

A strongly associated idea is that of mathematical literacy, a term originally coined in the USA and used to underpin international large-scale assessments like TIMSS and PISA. Niss and Jablonka (2014) contrast mathematical literacy 'as a tool for solving nonmathematical problems' with mathematical competence being 'what it means to master mathematics at large, including the capacity to solve mathematical as well as nonmathematical problems' (p. 392). Here, nonmathematical problems describe those which are resolved from outside of mathematics, but could contain mathematical techniques in their solution. The distinction rests on a contrast between mathematical and nonmathematical problems, which is often problematic. In the PISA 2012 report, an example question is given in which Chris needs to choose between four cars with engine capacities (in litres) given as 1.79, 1.796, 1.82 and 1.783. Only one part deals with the engine capacities, being question 2, which asks: 'Which car's engine capacity is the smallest?' (OECD 2014, p. 42). It would be hard to see why knowing the smallest capacity of four engines of roughly 1.8 litres would be a component in the choice of a car. So, a reasonable conclusion would be that this is a purely mathematical problem placed in a story context.

Jablonka (2015) suggests that the teaching of 'more immediately useful mathematics [...] has been interpreted as a reaction to curriculum reforms associated with the 'new mathematics' [from the 1960s]' (p. 601), which being based on the formal mathematics of Bourbaki was a highly specialised presentation, 'aimed at identifying an essence of academic mathematics meant to be made accessible to all students' (p. 601). However, exercise problems were frequently placed in contexts exactly as in the PISA example above. This again suggests that the PISA question is an example of a purely mathematical problem.

The PISA question constructs action within school mathematics, being used as describing action in a process of car buying. The recognition rules for competence in car buying relate to a successful purchase, while within school mathematics a successful engagement with the mathematical ideas to be learned would be required. In school mathematics, we order decimals to demonstrate competence in the place value number system, whereas when buying a car we would consider all engines of roughly 1.8 litres as equivalent and look for other buying criteria. There has been a recontextualisation, and a mythological car buying practice has been constructed in its recruitment as a context within school mathematics. This is not in any sense real car buying, so the context of the car purchase and the engine sizes is a myth. Dowling (2010) refers to this as a push strategy in which the mathematical practice is pushed into the car buying practice and is privileged in doing so; it is the mathematical practical which gives purpose to this activity. He contends that 'school mathematics fails to provide transferrable competences in push mode' and the mathematical calculation provides no meaningful input to car buying. In contrast, in fetch mode, the mathematical practice is fetched from the car buying practice. So, the problem originating context is privileged and retains its recognition rules for competence within the problem-solving activity. However, 'it is an empirical question as to whether mathematical competences may be productively useable in fetch mode' (Dowling 2010, p. 5). Here, fetch mode gives rise to applied mathematics as one may expect it to be manifested in a professional setting, such as engineering, bridge building or medicine, where building a successful bridge or designing an effective drug are the privileged recognition rules. However, this is only very rarely seen in an educational setting. Mellin-Olsen's (1987) description of the multi-faceted, rich mathematical project he worked on with his students to inform parents about the issues associated with the extension of the nearby Bergen airport was certainly in fetch mode. However, a student, while happily engaged, is concerned that 'the other class is half-way through the book by now' (Mellin-Olsen 1987, p. 41). Fetch mode is hard to deploy in a crowded, exam-focused curriculum.

In school mathematics, inquiry is possible with problems set in purely mathematical contexts requiring an open, exploratory problem-solving approach. These are referred to as investigations in the UK. Here, there is no recontextualisation, the problem is solved entirely within the mathematical practice and hence the competences required are only mathematical. In contrast, open problems placed in a naturalistic setting but set within school mathematics are examples requiring mathematical modelling (Blum et al. 2007). Modelling generates an uncertain relationship between the problem-to-be-solved and the mathematical model.

Mathematical modelling is presented as a cyclical practice of action, critique and improvement between the originating practice (e.g. science or engineering) and the mathematical practice (Blum et al. 2007). In push mode, the originating practice is mythologised and the competences will be entirely mathematical. However, mathematical modelling provides the possibility for fetch mode, and in this case competences in both practices will be required.

We have described three types for inquiry-based learning: open problems solved within mathematics (investigation), open problems of application in push mode (which are purely mathematical questions set in context) and open problems in fetch mode (requiring mathematical modelling). The assessment implications of these three types are necessarily starkly different. Exercise questions set in context are a very standard feature of all public examinations in mathematics and, as we have seen, high-stakes international comparison tests. The answer is known and unambiguous and scored accordingly. In England, public examinations in mathematics were offered with up to 100% of the assessment allocated to 'coursework', teacher assessment, often including open tasks. Initially a very high level of freedom was accorded to teachers to decide what counted as coursework. However, the specification tightened considerably, and in its final incarnation an extended piece of open investigative work operating most commonly entirely within mathematics was graded according to three process strands (under the general heading of 'using and applying mathematics') with a multi-level statement bank offering performance descriptors under each strand. A formula then provided a score. However, by the later 1990s and 2000s, the difficulty of ensuring the security and reliability of independent teacher-assessed work in an era of greater government oversite leads to the abandonment of all coursework elements in high-stakes examinations. Mathematical modelling as described above has no heritage of assessment in high-stakes testing. However, the case study below describes an instance of an internal mechanism where such assessment could be developed.

In the recent EU-funded project ASSIST-ME (assistme.ku.dk/practical-examples/swiss-examples), students have engaged in inquiry-based problem-solving. For example, Swiss students in grades 4 and 6 worked on puzzle/game problems involving a board game with specified rules and end conditions. Students were invited to explore the game and describe the possible outcomes. Additionally, students in lower secondary level engaged with a problem of two cyclists who must share a bicycle to complete a journey, alternately walking and riding. In both cases a collection of assessed competences are given, one common to both activities being: 'exploring problems and making conjectures'. Competence here is demonstrated with mathematical recognition rules in the game case. The game is analysed purely as a mathematical practice. In the cycling case competence is recognised in push mode. Here, students must show competence in 'transferring problems into the 'mathematical world' (if necessary)'. The word *transferring* infers a modelling process, and indeed the problem potentially provides useful insights into the particular situation of sharing resources when walking and cycling. Competent conjecturing and exploration would need to be recognised in the cycling/walking setting as well as the mathematical realisation of it for the problem to operate in fetch mode and hence be an application. A range of assessment mechanisms are provided for in the materials: (1) on-the-fly teacher assessment of competence in the problem-solving process, (2) student-produced posters followed by classroom discussion of them, (3) written comment and peer feedback on student groups' oral presentations and (4) written comments and self-assessment of the competences which are given levels of achievement. All of these mechanisms operate within a framework described as assessment for learning (Black and Wiliam 2004; Hodgen and Wiliam 2006). The last item has the potential for an external assessment tool, but there is no suggestion that it could be used as such.

Mathematical competence is the successful deployment of capabilities in engagement with mathematical problems and with the language and tools of mathematics. It can be deployed in problem-solving within mathematics and from naturalistic settings. Competence is described as a collection of process categories sometime allied to content categories. However, mathematical engagement in naturalistic settings generates recontextualisations, frequently mythologising practices from outside of mathematics, as in the example of choosing a car on the basis of small differences in engine size. Problems in a mathematical setting can be used to directly develop and assess mathematical competences, whereas problems from nonmathematical settings require mathematical modelling in fetch mode to generate a credible application and hence a site where competence can be described.

Innovation Competence: A New Perspective

The present political context for education, in general, and science education, in particular, is permeated by the relatively new trend of seeing teaching as something that fosters students' innovation competence. This trend comes to the foreground in the twenty-first-century skills programme:

> Given the twenty-first century demands […] it should come as no surprise that creativity and innovation are very high on the list of twenty-first century skills. In fact, many believe that our current Knowledge Age is quickly giving way to an Innovation Age, where the ability to solve problems in new ways [will …] be highly prized. (Trilling and Fadel 2009, p. 56)

Indeed, key American, European and international policy organisations have called for changes to the educational systems that make future generations more innovative in order to secure sustained social welfare (European Commission 2010; OECD 2010; White House 2011). Clearly, the instalment of a buzzword such as *innovation* as a goal of teaching can frustrate science educators: exactly what do these new educational goals signify? Concrete understandings of innovation competence are still only emerging on the horizon. Unfortunately, the word innovation is often used in a way that connotes economical or financial gain, as a process that 'involves creating and marketing of the new' (Kline and Rosenberg 1986, p. 275). This pecuniary way of parsing innovation, however, seems ill-equipped as a learning goal for school science. A number of scholars have recently attempted to identify other ways of parsing *teaching-for-innovation* that are more suited to classroom teaching in the existing disciplines and which specify valuable skills and competences that cover various disciplines (e.g. Christensen et al. 2012; Nielsen 2015).

An important distinction should be made between entrepreneurship (understood here as the transformation of a service, product or process into financial gain) and innovation (which we could initially determine as the process of improving a field of practice by drawing on (inter-)disciplinary knowledge and skills (Nielsen and Holmegaard 2015; Rump et al. 2013). The focus here is on innovation, not on entrepreneurship.

In what seems to be the most detailed investigation yet of what innovation competence could be in a teaching context, Nielsen (2015) worked with groups of upper secondary school teachers from Denmark, who were experienced in designing teaching activities in their own disciplines that foster innovation. In Denmark, as well as in most other Nordic countries, innovation competence has been a focal point in educational policy for at least the last 20 years (see Danish Ministry of Education 1995; Nordic Council of Ministers 2011), but without a clear definition of what innovation competence is as a learning goal.

In the study of Nielsen (2015), from the teacher groups' talk about how to assess students' innovation competence in practice, there emerged a composite understanding of innovation competence that involves five dimensions: creativity, collaboration, (disciplinary) navigation, implementation and communication. For each dimension, Nielsen (2015) found key aspects in the teachers' talk that may be used as regulative ideals in assessing students along each dimension (for a more detailed outline of each dimension, see Nielsen 2015):

- Creativity involves students' ability to (1) independently find, or independently interpret a given problem issue from a field of practice, (2) generate a range of ideas or solutions to a problem rather than just one idiosyncratic type of idea and (3) to work with generated ideas in a critical fashion, e.g. by evaluating, sorting, revising and expanding the ideas of themselves or others. In this way the creativity dimension of innovation competence is in line with state-of-the-art notions of creativity in general, as an ability that involves both divergent (idea generating) and convergent (revising ideas in light of an end goal) processes (Cropley 2006). Further, this particular conception of creativity is in line with the more modern approach to creativity as a skill set that can be developed, rather than a stable trait of the individual (see, e.g. Jeffrey and Craft 2004).
- Collaboration involves students' ability to (1) take responsibility and facilitate that the collaborative group finishes its tasks, e.g. by being able to identify how the competences of the people in the group can complement each other, and (2) to include and be flexible in a collaboration, e.g. by being able to work with many different types of stakeholders or people, rather than just a limited number of people or classmates. This particular understanding of students' ability to collaborate resonates with recent attempts to formalise assessment of collaborative skills by OECD (2016) for PISA 2015.

- (Disciplinary) navigation involves students' ability to (1) interpret a specific problem from practice as a problem that can be approached from a disciplinary perspective, e.g. by being able to translate the problem into disciplinary language; (2) functionally handle knowledge, e.g. by handling a plentiful and heterogeneous information and sorting and prioritising which information is most important to go into detail with; and (3) master complex work process. Such aspects resonate well with recent attempts to formalise information literacy (Ainley et al. 2005; Binkley et al. 2012).
- Implementation (or action) involves students' ability to (1) make informed decisions about what actions to take in a set time in a work process, (2) take action outside their comfort zone (e.g. by seeking information outside the classroom) and (3) take risks and put themselves and others into play, e.g. by not stopping at the level of an idea, but carrying that idea out. As such, these aspects are in line with current trends in managerial education research (Oosterbeek et al. 2010), and they resemble what others have called implementation skills (e.g. The Conference Board of Canada 2013).
- Communication involves students' ability to (1) assess how to communicate (among themselves or to other stakeholders) in a given situation, (2) master a range of communication techniques and genres and (3) communicate in an engaging and convincing manner. Again, this set of aspects resonates with the communication aspects in the twenty-first-century skills programme (Binkley et al. 2012).

From such a perspective, innovation competence combines a multitude of sub-competences or skills that together could be determined as students' ability (alone or in collaboration with others) to (a) generate solutions to issues while drawing on their disciplinary knowledge and their analysis of the field of practice where the issue arises, (b) analyse and reflect on the value-creating potential and realisability of their ideas, (c) work towards implementing their ideas and (d) communicate about their ideas to various stakeholders (Nielsen and Holmegaard 2015). It should be noted that just like other generic competences (i.e. competences that are not endemic to one discipline alone) such as modelling or inquiry, there are aspects or dimensions of innovation competence that are also important in other competences.

Concluding Remarks

Summarising the development of the concept of competence in education in general as well as in the domains of science, technology and mathematics, it becomes obvious that the construct of competence is still difficult to define especially in relation to the concepts of Bildung and literacy. The conceptual delimitation stays rather vague. One reason is that competence is a complex construct relying on different constituents. An analytical perspective on competence thus naturally has limitations and is in danger of underrepresenting the construct. A more holistic perspective, on the other hand, better represents the construct.

In all three domains, science, technology and mathematics, curricula, standards, and assessment frameworks exist that define specific competences. These definitions can help to get a clearer picture of what competences are in distinct domains. These definitions also highlight that the concept of competence is similarly operationalised in these domains. However, in science and mathematics the concept of competence is still the subject of an ongoing debate that is not carried out in the same way in the field of technology. This might be due to the fact that the issues discussed, e.g. the role of contexts and the specificity of competencies in certain situations, are less controversial in technology because the essence of technology lies in the ways in which, e.g. societal contexts, students' knowledge and skills base and the processes of designing interact.

Furthermore, it also becomes obvious that the construct of competence in some respect goes beyond the concepts of Bildung and literacy. It is related much more to students' everyday life by focusing on complex problems that might be highly relevant to them. Although the focus on subject-specific contexts or everyday life situations is a similarity between the three domains, the nature of these contexts or situations differs because they reveal the intrinsic characteristic of each domain. In mathematics and also in science, the main objective of applying competences is to generate knowledge that is new to the students. This is why modelling is such an important competence in science education. In technology education the objective of applying competences is to develop a new product or to improve an existing one in order to meet societal or personal needs and expectations.

The concept of competence is seen as the answer to new developments and challenges in our society and world. Being competent means being able to address problems caused by these developments and challenges from a meta-perspective. Although it is a prerequisite, it is often not sufficient to be able to suggest and develop solutions for a specific problem. Students need to be able to generalise solutions or transfer them to different contexts; they need to realise when more information about a system or situation is needed or when potential risks need to be evaluated and traded versus potential benefits. Furthermore, they need to realise that no concrete solutions exist for some of the challenges facing humanity. Only then will they be able to react to our rapidly changing world and effectively participate in today's society and labour market. Here it becomes obvious that the concepts of knowledge and competences are intertwined. Reacting to problems in a competent way requires the application of existing knowledge. Otherwise, for example, it is not possible to interpret data to make evidence-based decisions.

New learning goals involve new teaching and learning approaches as well as new assessment methods. In view of teaching and learning approaches, it is important to develop students' competences by focusing on typical subject-specific thinking and working processes that help students solve problems in different situations (s. Chap. 2). In the classroom context, the authenticity of problems or situations often becomes inherently reduced. Even if some problems are highly relevant, they might be too complex for students to *solve* in a classroom situation. Therefore, teachers have to create learning situations that fit students' level of cognitive development and that are nevertheless as realistic as possible. Otherwise the gap between the classroom context and real-life situations might hinder meaningful learning and successful transfer of knowledge and competences. With respect to the assessment of competences, the complexity of the construct afflicts issues of validity and reliability. It is obviously not possible to measure the full range of students' competences in a valid and reliable way using only one test. Therefore, it is still a challenge to develop valid and reliable tests that cover different competences and that are useful for formative and summative assessment purposes (s. Chap. 3).

Altogether, defining, teaching and learning competences as well as the assessment of competences remain challenges for science, technology and mathematics education. One possibility to overcome the existing dichotomy between an analytical and a holistic view could be the more integrated framework proposed by Blömeke et al. (2015) that encompasses competences including indicators for cognitive, affective and motivational traits demanded in particular situations and related to the performance through a set of perceptual, interpretive and decision-making processes.

References

Ainley, J., Fraillon, J., & Freeman, C. (2005). *National assessment program: ICT literacy years 6 & 10 report*. Carlton South: The Ministerial Council on Education, Employment, Training and Youth Affairs.

American Association for the Advancement of Science. (Ed.). (1989, 1990). *Science for all Americans – Online*. Retrieved from www.project2061.org/publications/sfaa/online/intro.htm

Australian Curriculum, Assessment and Reporting Authority. (2016). *Design and technologies sequence of achievement F-10*. Retrieved from www.australiancurriculum.edu.au

Binkley, M., Erstad, O., Herman, J., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In I. P. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). New York: Springer.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validitiy in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215–232.

Black, P., & Harrisson, C. (2016). *Teacher education programme. SAILS – Strategies for assessment of inquiry learning in science*. London: King's College London.

Black, P., & Wiliam, D. (2004). The formative purpose: Assessment must first promote learning. *Yearbook of the National Society for the Study of Education, 103*, 20–50.

Black, P., & Wiliam, D. (2006). *Inside the black box: Raising standards through classroom assessment*. London: Granada Learning.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies. Competence viewed as a continuum. *Zeitschrift für Psychologie, 223*(1), 3–13.

Blum, W., Galbraith, P. L., & Henn, H. W. (2007). *Modelling and applications in mathematics education: The 14th ICMI study*. New York: Springer.

Bulte, A. M. W., Westbroek, H. B., de Jong, O., & Pilot, A. (2006). A research approach to designing chemistry education using authentic practices as contexts. *International Journal of Science Education, 28*(9), 1063–1086.

Bybee, R. W. (1997). Towards an understanding of scientific literacy. In W. Gräber & C. Bolte (Eds.), *Scientific literacy – An international symposium* (pp. 37–68). Kiel: Institut für die Pädagogik der Naturwissenschaften (IPN).

Christensen, T. S., Hobel, P., & Paulsen, M. (2012). Evaluering af projekt innovationskraft og entreprenørskab i gymnasiet i Region Hovedstaden. Innovation i gymnasiet. Rapport 3 og 4 [evaluation of the project innovation and entrepreneurship in high school in the capital region. Innovation in high school. Report 3 and 4]. Odense, Denmark: Institut for Filosofi, Pædagogik og Religionsstudier, Syddansk Universitet.

Cizek, G. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20*, 19–28.

Cropley, A. (2006). In praise of convergent thinking. *Creativity Research Journal, 18*(3), 391–404.

Danish Ministry of Education. (1995). *En samlet uddannelse strategies på iværksætterområdet [a comprehensive educational strategy on the area of entrepreneurship]*. Copenhagen: Danish Ministry of Education.

DeBoer, G. E. (2000). Scientific literacy: Another look at its historical and contemporary meanings and its relationship to science education reform. *Journal of Research in Science Teaching, 37*(6), 582–601.

Department for Education. (2015). *Design and technology: GCSE subject content*. London: Department for Education. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/473188/GCSE_design_technology_subject_content_nov_ 2015.pdf.

Departmentment for Education and Employment/Qualifications and Curriculum Authority. (1999). *Design and technology: National curriculum for England*. London: HMSO.

Doppelt, Y. (2009). Assessing creative thinking in design-based learning. *International Journal of Technology and Design Education, 19*(1), 55–65.

Dowling, P. (2010). *Abandoning mathematics and hard labour in schools: A new sociology of education and curriculum reform.* In: MADIF7, 2010–01-27 - 2010-01-27, Stockholm.

European Commission. (2010). *Europe 2020: A strategy for smart, sustainable and inclusive growth*. Brussels: EU-Commission.

Gilbert, J. K. (2006). On the nature of "context" in chemical education. *International Journal of Science Education, 28*(9), 957–976.

Gitomer, D., & Duschl, R. (1998). Emerging issues and practices in science assessment. In B. Fraser & K. Tobin (Eds.), *International handbook of science education* (pp. 791–810). Dordrecht: Kluwer Academic Publishers.

Grugeon-Allys, B., Godino, J., & Castela, C. (2016). Three perspectives on the issue of theoretical diversity. In B. R. Hodgson, A. Kuzniak, & J. B. Lagrange (Eds.), *The didactics of mathematics: Approaches and issues. A homage to Michèle artigue* (pp. 57–86). New York: Springer.

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

Harlen, W. (2013). *Assessment & inquiry-based science education: Issues in policy and practice*. Trieste: Global Network of Science Academies (IAP) Science Education Programme.

Hodgen, J., & Wiliam, D. (2006). *Mathematics inside the black box: Assessment for learning in the mathematics classroom*. London: Granada Learning.

International Technology Education Association. (2007). *Standards for technological literacy: Content for the study of technology*. Retrieved from www.iteea.org/File.aspx?id=67767&v=b26b7852.

Jablonka, E. (2015). The evolvement of numeracy and mathematical literacy curricula and the construction of hierarchies of numerate or mathematically literate subjects. *ZDM, 47*, 599–609.

Jeffrey, B., & Craft, A. (2004). Teaching creatively and teaching for creativity: Distinctions and relationships. *Educational Studies, 30*(1), 77–87.

Kauertz, A., Neumann, K., & Härtig, H. (2012). Competence in science education. In B. Fraser, K. Tobin, & C. McRobbie (Eds.), *Second international handbook of science education*, *Springer international handbooks of education* (Vol. 24, pp. 711–721). Dordrecht: Springer.

Kilpatrick, J. (2014). Competency frameworks in mathematics education. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 85–87). Dordrecht: Springer.

Kimbell, R. (1997). *Assessing technology: International trends in curriculum and assessment*. Buckingham: Open University Press.

Kimbell, R. (2012). The origins and underlying principles of e-scape. *International Journal of Technology and Design Education, 22*(2), 123–134.

Kimbell, R., & Perry, D. (2001). *Design and technology in a knowledge economy*. London: Engineering Council.

Kimbell, R., & Stables, K. (2007). *Researching design learning: Issues and findings from two decades of research and development*. Berlin: Springer.

Kimbell, R., Stables, K., Wheeler, T., Wozniak, A., & Kelly, A. V. (1991). *The assessment of performance in design and technology*. London: SEAC / HMSO.

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., ... Whitehouse, G (2009). E-scape portfolio assessment: A research & development project for the Department of Children, Families and Schools, phase 3 report. London: Goldsmiths College.

Klieme, E., Avenarius, H., Blum, W., Döbrich, P., Gruber, H., Prenzel, M., et al. (2004). *The development of national educational standards – An expertise*. Berlin: Federal Ministry of Education and Research (BMBF).

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Klieme, & D. Leutner (Eds.), *Assessments of competences in educational contexts* (pp. 3–22). Cambridge, MA: Hogrefe.

Kline, S. J., & Rosenberg, N. (1986). An overview of innovation. In I. R. Landau & N. Rosenberg (Eds.), *The positive sum game* (pp. 275–305). Washington, DC: National Academy Press.

Koeppen, K., Hartig, J., Klieme, E., & Leutner, D. (2008). Current issues in competence modeling and assessment. *Zeitschrift für Psychologie/Journal of Psychology, 216*(2), 61–73.

Layton, D. (1993). *Technology's challenge to science education: Cathedral, quarry or company store*. Buckingham: Open University Press.

Leuders, T., & Sodian, B. (2013). Inwiefern sind Kompetenzmodelle dazu geeignet kognitive Prozesse von Lernenden zu beschreiben? [To what extent can competence models describe cognitive processes?]. *Zeitschrift für Erziehungswissenschaften, 16*, 27–33.

McLaren, S. (2012). Assessment is for learning: Supporting feedback. *International Journal of Technology and Design Education, 22*(2), 227–245.

Mellin-Olsen, S. (1987). *The politics of mathematics education*. New York: Springer.

Ministry of Education. (2010). *Technology curriculum support*. Wellington: techlink.org.nz.

Mioduser, D. (2015). The pedagogical ecology of technology education: An agenda for future research and development. In P. J. Williams, A. Jones, & C. Buntting (Eds.), *The future of technology education* (pp. 77–98). Singapore: Springer.

Moreland, J. (2009). Assessment: Focusing on the learner and the subject. In A. Jones & M. de Vries (Eds.), *International handbook of research and development in technology education* (pp. 445–448). Rotterdam: Sense Publishers.

Moreland, J., Jones, A., & Barlex, D. (2008). *Design and technology inside the black box: Assessment for learning in the design and technology classroom*. London: GL Assessment.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Nielsen, J. A. (2015). Assessment of innovation competency: A thematic analysis of upper secondary school teachers' talk. *The Journal of Educational Research, 108*(4), 318–330.

Nielsen, J. A., & Holmegaard, H. T. (2015). Innovation and employability: Moving beyond the buzzwords - a theoretical lens to improve chemistry education. In I. Eilks & A. Hofstein (Eds.), *Relevant chemistry education – From theory to practice* (pp. 317–334). Rotterdam: Sense Publishers.

Niss, M. (2004). Mathematical competencies and the learning of mathematics: The Danish KOM project. In A. Gagtsis & Papastavridis (Eds.), *3rd Mediterranean conference on mathematical education, 3–5 january 2003, Athens, Greece* (pp. 115–124). Athens: The Hellenic mathematical society.

Niss, M., & Jablonka, E. (2014). Mathematical literacy. In S. Lerman (Ed.), *Encyclopedia of mathematics education* (pp. 391–396). Dordrecht: Springer.

Nordic Council of Ministers. (2011). Kreativitet, innovation og entreprenørskab i de nordiske uddannelsessystemer – Fra politiske hensigtserklæringer til praktisk handling [creativity, innovation, and entrepreneurship in the Nordic educational systems – From political intentions to practical action]. Copenhagen, Denmark: Nordic Council of Ministers.

OECD. (2010). *The OECD innovation strategy: Getting a head start on tomorrow*. Paris: OECD Publishing.

OECD. (2014). *PISA 2012 results: What students know and can do (volume I)*. Paris: OECD Publishing.

OECD. (2016). *PISA 2015 assessment and analytical framework. Science, reading, mathematic and financial literacy*. Paris: OECD Publishing.

Oosterbeek, H., van Praag, M., & Ijsselstein, A. (2010). The impact of entrepreneurship education on entrepreneurship skills and motivation. *European Economic Review, 54*(3), 442–454.

Qualifications and Curriculum Authority. (2007). *Science. Programme of study for key stage 3 and attainment targets*. Retrieved from https://www.stem.org.uk/elibrary/resource/28541

Ritz, J., & Reed, P. (2006). Technology education and the influence of research: A United States perspective, 1985–2005. In M. J. de Vries & I. Mottier (Eds.), *International handbook of technology education: Reviewing the past twenty years* (pp. 113–124). Rotterdam: Sense Publishers.

Roberts, D. (2007). Scientific literacy/science literacy. In S. K. Abell & N. G. Lederman (Eds.), *Handbook of research on science education* (pp. 729–780). Mahwah: Lawrence Erlbaum.

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education, 52*(2), 161–197.

Rump, C., Nielsen, J. A., Hammar, P., & Christiansen, F. V. (2013). *A framework for teaching educators to teach innovation.* Paper presented at the SEFI (European Society for Engineering Education) 2013 conference, Leuven, Belgien.

Rychen, D. S., & Salganik, L. H. (2003). A holistic model of competence. In D. S. Rychen & L. H. Salganik (Eds.), *Key competencies for a successful life and a well-functioning society* (pp. 41–62). Göttingen: Hogrefe & Huber.

Sadler, D. R. (2013). Making competent judgements of competence. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn, & J. Fege (Eds.), *Modeling and measuring competencies in higher education – Tasks and challenges* (pp. 13–28). Rotterdam: Sense Publishers.

Schecker, H., & Parchmann, I. (2007). Standards and competence models: The German situation. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Standards in science education* (pp. 147–164). Münster: Waxmann.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education, 22*(2), 205–226.

Sen, A. (1992). *Inequality reexamined*. New York: Russell Sage Foundation.

Shavelson, R. J. (2010). On the measurement of competency. *Empirical Research in Vocational Education and Training, 1*, 43–65.

Skolverket. (2009). *Syllabuses for the compulsory school*. Stockholm.

Skolverket. (2012). *Upper secondary school 2011*. Stockholm.

Snape, P., & Fox-Turnbull, W. (2013). Perspectives of authenticity: Implementation in technology education. *International Journal of Technology and Design Education, 23*(1), 51–68.

Stables, K. (2013). *Social and cultural relevance in approaches to developing designerly well-being: The potential and challenges when learners call the shots in design and technology projects*. Paper presented at the Technology Education for the future: A play on sustainability, Christchurch, New Zealand.

Stables, K., & Kimbell, R. (2000). *The unpickled portfolio: Pioneering performance assessment in design and technology.* Paper presented at the D&T International Millennium Conference: Learning from Experience: Modelling new futures, Institute of Education, University of London.

The Conference Board of Canada. (2013). *Innovation skills profile 2.0*. Ottawa: The Conference Board of Canada, Centre for Business Innovation.

Theyßen, H., Schecker, H., Gut, C., Hopf, M., Kuhn, J., Schreiber, N., et al. (2014). Modelling and assessing experimental competencies in physics. In C. Bruguière, A. Tiberghien, & P. Clément (Eds.), *Topics and trends in current science education: 9th ESERA conference selected contributions* (pp. 321–339). Dordrecht: Springer.

Theyßen, H., Dickmann, M., Neumann, K., Schecker, H., & Eickhorst, B. (2016). Measuring experimental skills in large scale assessments: A simulation-based test instrument. In J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto, & K. Hahl (Hrsg.), *Proceedings of the bi-annual conference of the European Science Education Research Conference (ESERA)* (pp. 1598–6006).

Trilling, B., & Fadel, C. (2009). *21st century skills: Learning for life in our times*. San Francisco: Wiley.

Turnbull, W. (2002). The place of authenticity in technology in the New Zealand curriculum. *International Journal of Technology and Design Education, 12*, 23–40.

Waddington, D., Nentwig, P., & Schanze, S. (Eds.). (2007). *Making it comparable – Standards in science education*. Münster: Waxmann.

Weinert, F. E. (2001). Concept of competence: A conceptual clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Seattle: Hogrefe & Huber.

White House. (2011). *A strategy for American innovation, securing our economic growth and prosperity*. Washington, DC: The National Economic Council, Council of Economic Advisors, and Office of Science and Technology Policy.

Williams, P. J. (2006). Technology education in Australia: Twenty years in retrospect. In M. J. de Vries & I. Mottier (Eds.), *International handbook of technology education: Reviewing the past twenty years* (pp. 183–196). Rotterdam: Sense Publishers.

Williams, P. J. (2013). Engineering studies. In P. J. Williams & P. Newhouse (Eds.), *Digital representations of student performance for assessment* (pp. 99–124). Rotterdam: Sense Publishers.

Chapter 2
The Teaching and Assessment of Inquiry Competences

Silke Rönnebeck, Jan Alexis Nielsen, Christopher Olley, Mathias Ropohl, and Kay Stables

Introduction

As described in Chap. 1, the educational standards in many countries reflect the transition from content-oriented towards competence-oriented learning goals. The sustainable implementation of these new learning goals, however, requires changes both in the teaching and the assessment of these goals. Within the field of science education, a fundamental approach of competence-oriented teaching is based on the concept of scientific inquiry or, as it is more recently called, scientific practices (e.g. Abd-El-Khalick et al. 2004; National Research Council 1996, 2012). From a European perspective, several high-level reports have identified scientific inquiry as one means to improve science teaching thus addressing the increasing discussion in Europe about the need to recruit more young people to careers in science and engineering in order to ensure economic development and welfare (e.g. European Commission 2004; Rocard et al. 2007). Despite its prominent role in science education research within the last 20 years, however, the concept of scientific inquiry is not uniquely defined (e.g. Abd-El-Khalick et al. 2004; Furtak et al. 2012). The

S. Rönnebeck (✉)
Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

Kiel University, Kiel, Germany
e-mail: sroennebeck@uv.uni-kiel.de

J.A. Nielsen
Department of Science Education, University of Copenhagen, Copenhagen, Denmark

C. Olley
King's College London, London, UK

M. Ropohl
Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

K. Stables
Goldsmiths, University of London, London, UK

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_2

situation becomes even more complex by looking across domains. In mathematics and technology education, inquiry-based approaches also exist but usually go under different names. In mathematics education, they are often related to problem solving, in technology education to design processes. Innovation can be regarded as a cross-curricular inquiry-based approach to the teaching and learning of twenty-first-century skills since it requires competences from different domains in order to solve problems from different areas of practice. Although no general definition of inquiry exists within and even less across domains, inquiry-based approaches share some common characteristics like the active engagement of students in the thinking and working processes of scientists with the aim of solving complex problems of a personal, societal, environmental or disciplinary nature.

In all domains, moreover, teaching for inquiry confronts teachers with new challenges. It implies a shift in emphasis away from teachers presenting information and covering content-related topics towards teachers as facilitators by creating 'environments in which they and their students work together as active learners' (National Research Council 1996, p. 4). Taking science again as an example, inquiry-based teaching requires teachers to constantly 'guide, focus, challenge, and encourage student learning' (National Research Council 1996, p. 33), e.g. by orchestrating discourse among students but also by modelling 'the skills of scientific inquiry, as well as the curiosity, openness […], and scepticism that characterize science' (ibid, p. 32).

Changing teaching practice, however, requires time and support from the educational system. An important aspect in this context is that changes in teaching need to be accompanied by changes in assessment in order to be sustainable. Assessment is one of the most important driving forces in education and a defining aspect of any educational system. Assessment signals priorities for curricula and instruction since teachers and curriculum developers tend to focus on what is tested rather than on underlying learning goals (Binkley et al. 2012; Gardner et al. 2010; Harlen 2007). Teaching and assessment goals thus need to be aligned, and assessment methods need to be developed that allow for the assessment of inquiry competences within the different domains.

The aim of this chapter thus is to shed light on the understanding of inquiry and inquiry-based teaching and assessment within and across the different domains. The following describes the concepts, teaching and assessment of scientific inquiry, mathematical problem solving, design processes and innovation. For each domain, the description is structured along three major questions: (1) How is the concept defined and which competences are students supposed to develop? (2) What changes in teaching are needed to support students in developing these competences? (3) What changes in assessment are needed to assess these competences? The chapter concludes with a discussion about commonalities and differences with respect to the implementation of inquiry in the three domains and innovation, respectively.

The Concept of Scientific Inquiry

Scientific inquiry is not a uniformly defined concept. Within the science education literature, a general disagreement and variation can be observed with respect to the meaning of inquiry (e.g. Abd-El-Khalick et al. 2004; Anderson 2002, Furtak et al. 2012; Hmelo-Silver et al. 2007; Kirschner et al. 2006). From a holistic perspective, scientific inquiry could be described as a teaching and learning approach that tries to imitate more or less authentic scientific investigations embedded in real-world contexts. Learners are presented with problems and questions and supported in identifying ways of solving these problems by applying scientific thinking and working processes like planning an investigation or constructing models and by drawing on their knowledge of scientific content and the nature of science with the aim of constructing new knowledge.

Against this background, curricula, frameworks and reviews often describe scientific inquiry as a set of activities and the underlying competences that these activities require (e.g. Bell et al. 2010; Linn et al. 2004; National Research Council 2012; Pedaste et al. 2015). The findings from a recent review, however, 'illustrate that the variability found in the research literature with respect to the definition and operationalisation of the holistic concept of scientific inquiry is also reflected at the level of single activities of the inquiry process' (Rönnebeck et al. 2016, p. 190). Nevertheless, the curricula, frameworks and reviews are reflective of different phases and steps in the inquiry process and lead to models of scientific inquiry that encompass subject-specific competences like planning investigations as well as more generic competences like communicating (Pedaste et al. 2015; Rönnebeck et al. 2016; cf. Fig. 2.1). Moreover, the model by Rönnebeck et al. (2016) explicitly acknowledges the importance of relating scientific inquiry to scientific knowledge and knowledge about the nature of science and scientific inquiry.

Typically, scientific inquiry is regarded as a process (e.g. White and Frederiksen 1998; cf. Fig. 2.1). In this process, students apply the underlying competences in a sequence of steps that build on each other, e.g. they start with formulating a question and then generate a set of competing predictions and hypotheses related to that question. The advantage of this understanding of inquiry as a process is that students have to 'reflect on both the limitations of what they have learned (which suggests new questions) and on the deficiencies in the inquiry process itself (which suggests how it could be improved)' (White and Frederiksen 1998, p. 4). The improvement leads students back to the beginning of the process with a new or refined question or a revised approach.

Innovations on the level of learning goals imply innovations in teaching and learning approaches in order to address newly defined competences. One innovation that resulted from the emphasis on scientific inquiry is a change in the pedagogy from passive, teacher-led instruction to active, student-driven and cooperative learn-

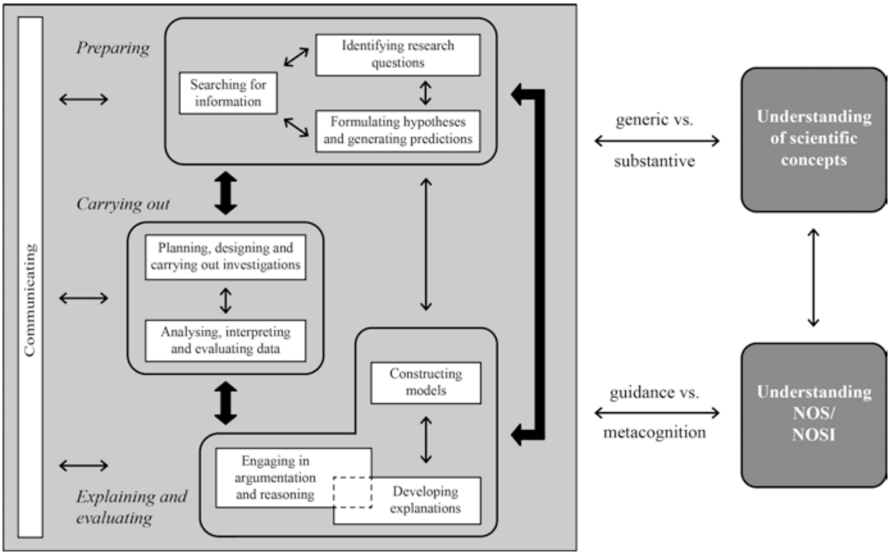

Communicating
Preparing
Searching for information
Identifying research questions
Formulating hypotheses and generating predictions
Carrying out
Planning, designing and carrying out investigations
Analysing, interpreting and evaluating data
Constructing models
Explaining and evaluating
Engaging in argumentation and reasoning
Developing explanations
generic vs. substantive
guidance vs. metacognition
Understanding of scientific concepts
Understanding NOS/ NOSI

Fig. 2.1 The concept of scientific inquiry as a model (Rönnebeck et al. 2016, p. 189)

ing (Barron and Darling-Hammond 2008; Pellegrino and Hilton 2012). Thus, students should be engaged in actively building their knowledge (cf. Furtak et al. 2012), while the teachers make students' thinking visible, guide small group work and ask questions to enhance students' self-reflection. Possible methodological approaches to address these issues are the combination of different activities, the use of open-ended tasks, the implementation of scaffolding and the realisation of self-directed learning.

Basically, teachers initiate the inquiry process by providing opportunities 'that invite student questions by demonstrating a phenomenon or having students engage in an open investigation of objects, substances, or processes' (Kessler and Galvan 2007, p. 2). During the inquiry process, teachers act as facilitators providing guidance and scaffolds where needed. In general, minimally guided instruction isn't likely to support student learning – in order to be effective, inquiry-based instruction requires active guidance from the teacher (Hmelo-Silver et al. 2007; Kirschner et al. 2006). For example, 'teachers will likely have to modify student questions into ones that can be answered by students with the resources available, while being mindful of the curriculum' (Kessler and Galvan 2007, p. 2). However, the role of the teacher in inquiry teaching is much more complex than simply being a facilitator. Crawford (2000) identified ten roles a teacher takes up in inquiry classroom settings: motivator, diagnostician, guide, innovator, experimenter, researcher, modeller, mentor, collaborator and learner. Instead of the traditional distinction between the teacher as the knowledge giver and the students as the knowledge receivers, in inquiry settings, teachers and students 'collaborate to develop conceptual understandings through shared learning experiences' (Crawford 2000, p. 933).

By describing teachers' roles, it becomes obvious that doing inquiry in school inevitably differs from the work of real scientists. One aspect of these differences can be seen in practical reasons, e.g. available time and materials or safety precautions. A more important aspect, however, is described by the concept of educational reconstruction. Inquiry activities in instruction are related to specific, predefined learning goals and especially designed for students' learning and understanding. This focus differs from the focus in general science research although the underlying principles of the inquiry processes in both settings are the same (or at least very similar). Inquiry in schools thus requires the transformation of an authentic research situation into an educational setting – which might at times result in less authenticity (Furtak 2006; see also Chap. 1).

Nevertheless, when engaging in inquiry, students should be given 'the opportunity to undertake 'research activities' instead of just carrying out routine 'cookbook experiments'' (European Commission 2004, p. 125). Doing scientific inquiry should not only require students to engage in hands-on but also in minds-on activities by providing meaningful and realistic problems that allow for multiple solutions and multiple methods for reaching these solutions (Barron and Darling-Hammond 2008). In order to make the application of multiple methods for students possible, it is important to teach them the process of scientific inquiry as a sequence of interrelated steps as well as each step separately. By this the students learn a repertoire of different ways of how to do inquiry. Hadfield's (1995) example of the so-called copper problem (see box below) represents such a teaching and learning situation where students undertake research activities in the sense of scientific inquiry.

This focus on learning how to do inquiry is reflected in recent years in the USA where the discussion has moved away from using the term inquiry in favour of emphasising the importance of engaging students in *scientific practices* as the means to 'establish, extend and refine' knowledge (National Research Council 2012, p. 27). The eight practices described in the Framework for K-12 Science Education include (1) asking questions (for science) and defining problems (for engineering), (2) developing and using models, (3) planning and carrying out investigations, (4) analysing and interpreting data, (5) using mathematics and computational thinking, (6) constructing explanations (for science) and designing solutions (for engineering), (7) engaging in argument from evidence and (8) obtaining, evaluating and communicating information (National Research Council 2012, p. 42).

The definition of scientific inquiry by the description of related competences implies that teachers need assessment methods for each of the defined competences that help them detect if students reached the learning goals or not. However, Hume and Coll (2010) as well as Shavelson (2011) emphasise the difficulty of assessing inquiry-related competences. The former conclude that existing 'standards-based assessments using planning templates, exemplar assessment schedules and restricted opportunities for full investigations in different contexts tends to reduce student learning about experimental design to an exercise in 'following the rules'' (p. 43). Shavelson (2011) argues that the more complex the learning goals, the more difficult they are to measure. The understanding of competences as the ability to cope with complex challenges in everyday life means that assessment methods have to focus on scientific knowledge and on scientific inquiry.

Inquiry-Based Learning Example by Malcolm Hadfield (1995)
The copper problem. Students in small groups hold a small piece of copper foil in the Bunsen flame using a pair of tongs. When the copper is red hot, they place it on a ceramic mat and allow the copper to cool. Then, the students describe their observations. Afterwards, they formulate hypotheses about the observable black layer on the copper. Common hypotheses are that the black layer is soot from the Bunsen flame, that it forms out of the copper itself or that it has something to do with the ambient air.

In a next step, the students plan an investigation that tests their hypotheses. For example, they could think about experimental setups that isolate the copper from the flame or from the ambient air (see Fig. 2.2). Then, the students conduct the investigation and observe what happens. Based on their observations, they draw conclusions from their observations and evaluate their hypotheses.

In this example, the teacher intervenes relatively seldom. It is important that the students plan their inquiry on their own. The teacher's role is mainly to ensure that safety regulations are respected.

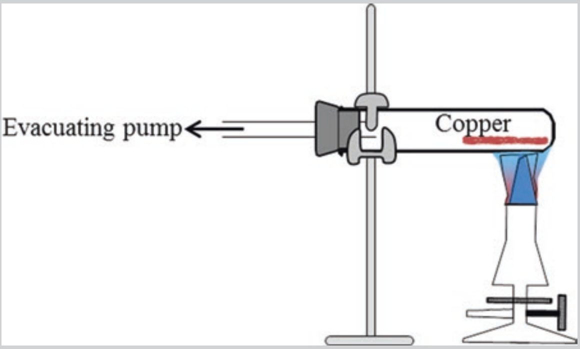


Fig. 2.2 Heating copper under conditions of a vacuum

In addition, the assessment methods should be related to everyday situations. Artificial tasks formulated as multiple choice test items can hardly measure inquiry-related competences. Usually, these items are used to assess students' understanding of scientific concepts. However, assessment methods should also focus on process-related aspects like students' competences in planning an investigation. Open-ended items or observations by the teacher seem to be more appropriate due to validity reasons. Compared to standards-based assessments, formative assessment has the great potential to address this issue of validity by focusing on process-related aspects (Barron and Darling-Hammond 2008; see also Chap. 3). Therefore, they are needed in addition to the above-mentioned standards-based assessments. The additional function of formative assessment is to give feedback to the students thus guiding their learning in the sense of scaffolding. Possible methods are rubrics, whole class discussions, performance assessments, written journals, portfolios, weekly reports and self-assessments (cf. Barron and Darling-Hammond 2008). It can be concluded that introducing scientific inquiry and formative assessment both require a considerable change in pedagogy (see also introduction of this book and Chap. 3 for details).

The Concept of Design Processes

As outlined in Chap. 1, competences in technology education link to its procedural nature, the core of which is design. This is recognised within technology education, 'Design is regarded by many as the core problem-solving process of technological development. It is as fundamental to technology as inquiry is to science' (International Technology Education Association 2007, p. 90), and also in science education, 'Technology as design is included in the [science] standards as parallel to science as inquiry' (National Research Council 1996, p. 24). In technology education, an inquiry approach involves presenting learners with challenges, problems and scenarios and supporting learners to identify ways of addressing these through iterative design processes that draw on critical thinking, creative and exploratory idea development and effective and thoughtful outcome resolution.

The early focus on design processes within technology education emerged in the late 1960s through a UK research project – the Design and Craft Education Project (Schools Council 1975). The project shifted the focus in what was a traditional *making* curriculum to *designing and making*, recognising that a design focus to teaching and learning enriched the subject greatly. The project was conducted at a time when design researchers were exploring professional design approaches, placing considerable attention on defining *the design process*. In the 1960s era of modernism, seeking the ultimate rational definition made sense and what emerged was a linear design process. This seemed like a logical sequence – identify a problem, conduct research, generate ideas, make a solution and evaluate its effectiveness. The stages in the process became a focus for teaching and, more significantly, for assessment with marks being allocated to each stage, creating what was, effectively, an early version of competences in technology education. The approach was embedded in a formal external examination for 16-year-olds (NWSEB 1970) and quickly spread to other assessment systems.

However logical a linear process may seem, it is more a management process than a representation of how designing takes place. The suggestion that a person somehow restrains from having any ideas until a *problem* is fully defined and all research is undertaken makes no sense – even if it was possible to prevent ideas from beginning to form. The notion that no evaluation needs to take place until the

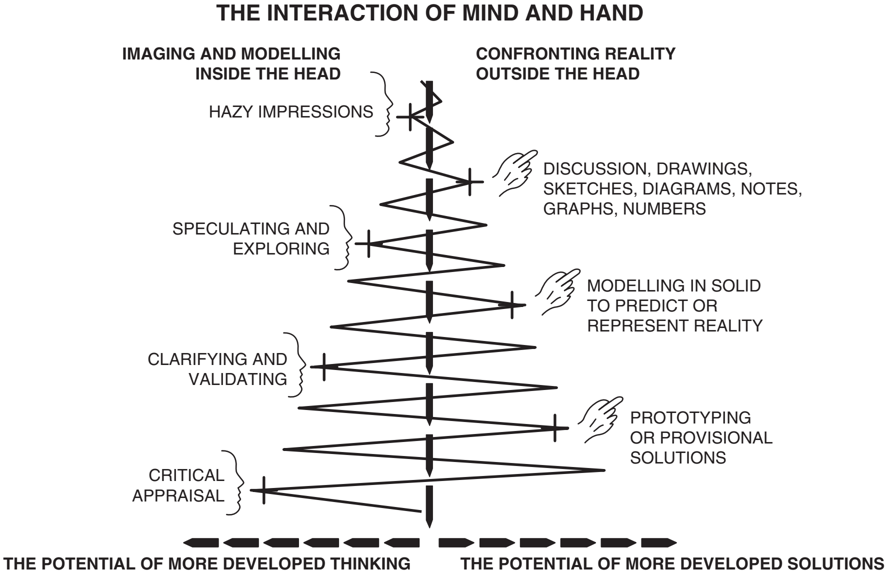

THE INTERACTION OF MIND AND HAND
IMAGING AND MODELLING INSIDE THE HEAD
CONFRONTING REALITY OUTSIDE THE HEAD
HAZY IMPRESSIONS
DISCUSSION, DRAWINGS, SKETCHES, DIAGRAMS, NOTES, GRAPHS, NUMBERS
SPECULATING AND EXPLORING
MODELLING IN SOLID TO PREDICT OR REPRESENT REALITY
CLARIFYING AND VALIDATING
PROTOTYPING OR PROVISIONAL SOLUTIONS
CRITICAL APPRAISAL
THE POTENTIAL OF MORE DEVELOPED THINKING
THE POTENTIAL OF MORE DEVELOPED SOLUTIONS

Fig. 2.3 The APU D&T model of iterative design (Kimbell et al. 1991, p. 20)

project is completed also makes no sense. Dissatisfaction with a linear model of the design process emerged both in the professional design world and in educational contexts. Addressing this dissatisfaction was an early challenge for a major research project commissioned to assess the design and technological capability of 15-year-olds in England, Wales and Northern Ireland (the APU D&T project, Kimbell et al. 1991). Drawing on the team's experience as design and technologists and teachers and on early empirical fieldwork, an alternative model was created (see Fig. 2.3). As its starting point, it took the spark of a hazy idea in the mind's eye, possibly provoked by a problem, maybe by an opportunity. The process was then a journey of taking action to develop the idea and iterating this with reflection, appraising developments to identify next steps, to-ing and fro-ing until a thoughtful, well-developed outcome was created. The iterative nature of the process was critical – action needed to be taken on the hazy internal idea to bring it out into the world by drawing, talking about it and modelling it in some way. In turn, the action provided models to reflect on, speculating on how it might develop, what information was needed to take it forward and so on.

The research team used the model as a framework for structuring and assessing design and technological capability (Kelly et al. 1987). The sample of learners generated samples of design and technology work that were initially assessed holistically and then analysed to identify qualities of performance. Holistic assessment allowed assessors to see the overarching qualities of the work, and analysis provided empirical evidence of the qualities within the work. Key attributes and competences became apparent through a combination of the two. Three clusters of procedural capability were identified: reflective qualities, active qualities and appraisal qualities that linked the two together. The holistic assessment revealed the extent to which high-quality performance was directly linked to the ability to iterate between action and reflection throughout the activity. It underscored the critical importance of an iterative process, a factor that was echoed by the national working party developing the first English/Welsh National Curriculum who stated that 'because Design and Technology activity is so integrative, the approach to the assessment of learners' performance in this area should ideally be holistic' (DES/WO 1988, para 1.30). Evidence of design iteration contributing to high levels of performance has been found in other projects, including recent research (Botleng et al. 2016; Crismond 2011; Strimel 2015).

Two research projects building on the APU D&T project clarified further detail. The Understanding Technological Approaches project (1992–1994) observed and documented in fine detail 80 live projects across ages ranging from 5 to 16. The project focused on the *in the moment* design and technological intentions of learners and how these were manifested (Kimbell et al. 1996). It identified the following facets of performance common across all age groups: investigating, planning, modelling and making, raising and tackling design issues, evaluating, extending knowledge and skills, and communicating. The Assessing Design Innovation project was commissioned to research ways of introducing creativity and innovation into high-stakes assessment projects (Kimbell et al. 2004). The research identified further elements of idea development – into the *having*, *growing* and *critiquing* of ideas within an iterative process. Once again, the interaction of these throughout a project was revealed through holistic assessment. Atkinson (1999) suggests that a holistic experience allows learners to understand how component parts of the process link together.

Iterative processes of design are now central in the English national curriculum, including in high-stakes assessment. The critique of linear processes and the shift to cyclical and iterative processes of design can be commonly seen in curriculum documentation in other settings (e.g. Department of Basic Education 2011; International Technology Education Association 2007; Ministry of Education 2010). In the USA, the International Technology Education Association (ITEA) Standards for Technological Literacy highlight an iterative process focusing specifically within engineering design. For them this is an important step to move away from a historic craft tradition and towards a stronger link with engineering futures, both in the work place, indicating a more instrumental ambition, and through a more general engineering literacy (International Technology Education Association 2007; Lewis and Zuga 2005). While engineering is being aligned with technology education in many national contexts, an engineering design process has a more narrow focus, engineering being but one of many significant contributions made through design.

Across technology education, there is relative agreement regarding the individual qualities that contribute to technological literacy and/or capability, such as addressing task and user needs, investigating, modelling ideas, applying and acquiring knowledge and skills and critiquing idea development. Furthermore, there is the fundamental ability of using the individual qualities in a responsive and integrated way through an iterative process and, as identified in Chap. 1, doing this within a societal context. Taken together, this repertoire of individual and integrative qualities is the challenge and focus of learning, teaching and assessment.

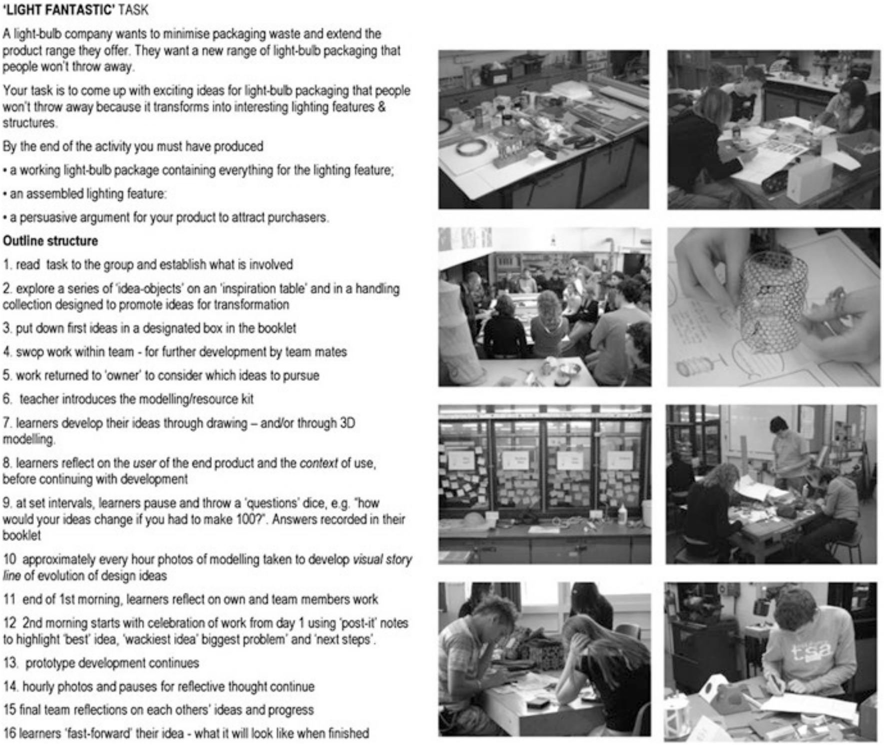
Technology teachers are generally comfortable with a project-based approach to learning and teaching, but manageability of time, resources and class sizes often leads to prescriptive projects that are dominated by teaching of knowledge and skills rather than being genuinely design led and based in socially and culturally relevant contexts. Reports in England have identified that, at worst, design and technology teaching can be too formulaic, too narrowly focused with projects that lack challenge and often result in unfinished outcomes. But they also identified that, at best, teachers have high expectations, set challenging and ambitious projects in relevant contexts that spark the learners' imaginations and create palpable excitement in the learning (Department for Education 2011; Design Commission 2011; Miller 2011; Ofsted 2011, 2012).

The importance of authenticity in project-based learning has been highlighted within and beyond technology education (e.g. Barak and Awad 2008; Merrill et al. 2010; Snape and Fox-Turnbull 2013, Stables 2013; Turnbull 2002). A major challenge for teachers is structuring and scaffolding projects that are set in authentic contexts that allow learners to take ownership of tasks. A broad and loosely defined context provides opportunities for ownership, but learners can get lost in the complexity and overwhelmed by a perceived enormity of the challenge. If it is too tightly specified, on the other hand, there may be oversimplification and too little room for personalisation (Jones 1997; Kimbell et al. 1991). A framework was created through the APU D&T project that identified levels of complexity and hierarchy in tasks from open contexts, to referenced scenarios, to specific briefs. The framework allows teachers to place a design challenge at an appropriate level for learners and then create the scaffolding to enable learners to move between the levels, keeping sight of the broader context and their own specific challenge (Kimbell et al. 1991; Kimbell and Stables 2007).

Barak and Awad (2008) provide insights into learners choosing of authentic tasks from within their own personal contexts, based on an area of specified challenge, for example, the creation of an information system, highlighting the motivational aspects of enabling personalisation in choices made. Snape and Fox-Turnbull (2013) also stress the importance of learner motivation in authentic tasks, seeing learner engagement as a dimension of the interweaving of elements of curriculum, suggesting that 'in order to elucidate authentic technological practice the dimensions of authenticity are woven together by rich contexts, social construction, meaningful connections and student engagement' (p. 60). They (along with others) also point to the value of a socio-constructivist approach through which knowledge is developed through social experience and collaboration and emphasise the value of a cognitive apprenticeship model. Moreland et al. (2008) regard teachers working alongside learners, modelling thinking and designing, as a valuable approach to support learning while providing feedback. Drawing attention to the requirements of supporting diverse projects and the resulting diverse learning needs these produce, Snape and Fox-Turnbull (2013) suggest that learner action plans enable teachers to manage a balance between *just in case* and *just in time* teaching.

To support metacognitive learning, documention and reflection are an important dimension of project-based learning. A curriculum-led assessment approach supports the dual value of making learning visible, both for the learner and the teacher, providing important insights for formative assessment. The use of project portfolios is ubiquitous in technology education. Portfolios based on selections of work are useful to help learners curate the documentation of their learning, but these *after-the-event*, container style portfolios have been critiqued for becoming products in their own right and acting as either a distraction for or displacement of learning, presenting ritualistic, prettied up documentation rather than the thought of an action as it took place. Mike Ive, former Chief HMI for design and technology in the UK, repeatedly referred to this as *neat nonsense*, reporting formally that 'many [learners] still spend too much time on superfluous decoration of their design folders rather than on real design development' (Ofsted 2002, p.4). McLaren (2007) suggests that this has resulted in assessment that fails to authentically model design processes hence resulting in the production of artificial documenting that demotivates learners.

An alternative to an *after-the-event* portfolio is a working portfolio where documentation is captured dynamically in real time as a project progresses. Spendlove and Hopper (2006) see working portfolios as liberating learners, opening up possibilities for creative dialogue. Digital tools used in e-portfolios have enabled this in a literal sense, building audio and video tools into portfolios in ways that provide opportunities to capture the learner's voice. The e-scape project (e-solutions for creative assessment in portfolio environments, Kimbell et al. 2009) explored this potential, by creating a web-based application that allowed teachers to structure learning activities through which learners documented their project work in real time, using a collection of digital tools including text, drawing, mind mapping, photo, audio and video, thus creating a 'trace of the thinking left behind' (Kimbell and Stables 2007, p. 222). Moreland et al. (2008) suggest that design and technology is an ideal place to exploit such multimodal approaches for assessment. Using digital tools has been found to support assessment for different learning styles, including learners with special educational needs (Stables et al. 2015). The e-scape system includes possibilities for peer assessment via text and drawing, using the concept of *critical friends*, thus also supporting collaboration. A linked project extended the initial range by adding an option for teachers to add formative feedback into the portfolios, including while the learners were working, 'comments/suggestions/ideas in exactly the way one would if talking directly to learners in the classroom' (McLaren 2012, p. 234). A further development currently being researched is the possibility of a built-in screen avatar taking a critical friend role (Stables et al. 2016). The assessment potential of e-portfolios has been exploited by examination boards enabling the submission of portfolios in digital format, often using an application such as Microsoft PowerPoint. The e-scape portfolio, being web-based, allowed for a further innovation through the use of adaptive comparative judgement that not only created high levels of reliability in assessment but also acted as a professional development tool for those engaged in the assessment process (Kimbell 2012; Pollitt 2012; McLaren 2012). It also enabled peer assessment,



'LIGHT FANTASTIC' TASK

A light-bulb company wants to minimise packaging waste and extend the product range they offer. They want a new range of light-bulb packaging that people won't throw away.

Your task is to come up with exciting ideas for light-bulb packaging that people won't throw away because it transforms into interesting lighting features & structures.

By the end of the activity you must have produced

- a working light-bulb package containing everything for the lighting feature;
- an assembled lighting feature:
- a persuasive argument for your product to attract purchasers.

Outline structure

1. read task to the group and establish what is involved
2. explore a series of 'idea-objects' on an 'inspiration table' and in a handling collection designed to promote ideas for transformation
3. put down first ideas in a designated box in the booklet
4. swop work within team - for further development by team mates
5. work returned to 'owner' to consider which ideas to pursue
6. teacher introduces the modelling/resource kit
7. learners develop their ideas through drawing – and/or through 3D modelling.
8. learners reflect on the *user* of the end product and the *context* of use, before continuing with development
9. at set intervals, learners pause and throw a 'questions' dice, e.g. "how would your ideas change if you had to make 100?". Answers recorded in their booklet
10 approximately every hour photos of modelling taken to develop *visual story line* of evolution of design ideas
11 end of 1st morning, learners reflect on own and team members work
12 2nd morning starts with celebration of work from day 1 using 'post-it' notes to highlight 'best' idea, 'wackiest idea' biggest problem' and 'next steps'.
13. prototype development continues
14. hourly photos and pauses for reflective thought continue
15 final team reflections on each others' ideas and progress
16 learners 'fast-forward' their idea - what it will look like when finished

Fig. 2.4 The structure of a *controlled assessment* iterative design assessment task from the Assessing Design Innovation project

explored in a small trial with 15-year-olds (Kimbell et al. 2009) and more extensively with undergraduate design students who found that it had 'the potential to increase thinking, learning and confidence, helping the student to establish the role and purpose of assessment' (Seery et al. 2012, p. 209).

Through the Assessing Design Innovation project, a structured design and technology assessment task framework was created, undertaken as *controlled assessment* (Isaacs 2010). An example of this is given here to illustrate how the task was structured (Fig. 2.4). The task was designed to take 6 h, ideally conducted over two consecutive mornings.

The Concept of Problem Solving

In Chap. 1 we considered the requirement within mathematical literacy to solve mathematical problems and contrasted those where the outcomes are validated within mathematics (sometimes referred to as investigations) and those where the validation came from outside the field of mathematics (mathematical modelling). Defining inquiry-based education (IBE) or inquiry-based learning (IBL) as a concept within mathematics education is relatively new and often associated with EU-funded projects (Maaß and Doorman 2013). In describing the rationale for the PRIMAS project, Maaß and Doorman (2013) define it as 'refer[ing] to a teaching culture and to classroom practices in which students inquire and pose questions, explore and evaluate' (p. 887). The nature and purpose of the problems-to-be-solved and how solutions might be validated poses significant issues for the competences that could be developed. Maaß and Doorman suggest IBE can support 'develop[ing] competences in such areas as attaining new knowledge, creative problem solving and critical thinking' (ibid, p.1). Alongside PRIMAS, the Fibonacci project aimed to 'contribute to the dissemination of IBL by designing, implementing, and evaluating a dissemination process' (Maaß and Artigue 2013, p. 788). This consisted of local and regional centres together with community involvement and an emphasis on collaboratively produced materials with attention on the diversity of contexts found in different centres.

The Danish KOM curriculum reform project roots new syllabus construction in mathematical competences. These seek to elaborate the competences involved in asking and answering questions (mathematical thinking/problem tackling/modelling/reasoning) and in mathematical language and tools (representing/symbol and formalism/communicating/aids and tools) (Niss and Højgaard 2011). The ordering seems significant in that asking and answering questions is made possible through the deployment of mathematical tools and language. It is notable that competences are required to both ask and answer questions in the sense of problems to be solved. So, the learner is involved in an engagement with the generation of the problem, and a clear distinction is made between problems within mathematics and those where mathematics is deployed in settings outside of mathematics. These are described as problem tackling and modelling, respectively. However, the notion that modelling is simply one of the eight competences above is critiqued by Niss himself who suggests that 'the entire domain of mathematical competencies must be perceived as a proper subset of the modelling competency' (Niss 2015, p. 2). This suggests that the totality of mathematics education is directed at problem solving (and posing) and that modelling is required to achieve this. IBE meanwhile presents a possible mechanism for achieving it. The description of the KOM project does not even contain the word *inquiry*, but the centrality of student activity in *investigation* and *modelling* occurs repeatedly, suggesting these terms describe a pedagogy comparable to IBE (Niss and Højgaard 2011).

In contrast, the PISA 2012 framework states that 'mathematical literacy is assessed in the context of a challenge or problem that arises in the real world' (OECD 2014, p. 37). This represents a more limited setting for mathematical problem solving, resonant with our initial notion of modelling, but not the overarching definition suggested by Niss, which would include problems within mathematics or investigations. Burke et al. (2016) propose a structuring of mathematical modelling practices enabling an analysis of practices such as those described above, when deployed as educational activities. This uses Dowling's notion of discursive saturation which determines the extent to which the principles of an activity can be determined in discursive forms (Dowling 2007). Where, for example, mathematics

Table 2.1 Four characterisations of mathematical modelling practice

Quantification rule (external syntax)	Mapping rule (internal syntax)	
	DS+	DS−
Ds+	Definitive mathematisation	Ad hoc mathematisation
Ds−	Derived mathematisation	Originative mathematisation

practices in general are highly discursively saturated, in that statements made within mathematics are very clearly determined by the language and syntax of mathematics, by contrast swimming is not generally available in discursive forms, and there is a weaker relationship between descriptions of it and the practice itself. Mathematical modelling requires an internal syntax in which mathematical terms and statements may be clearly constructed and amenable to proof (high discursive saturation/DS+). This is referred to as a mapping rule. It also requires an external syntax in which statements from the originating context can be clearly quantified and thus engaged with using the mathematics. This is the quantification rule. This establishes four characterisations of mathematical modelling practice (Burke et al. 2016, p. 4–5) as shown in Table 2.1.

The pedagogic aim of engaging with mathematical modelling would be to apprentice learners into a practice of definitive mathematisation. Yet, this mode is almost never present in problems used in the PISA tests. Either the problem fails to establish clarity internally or externally, making the practice ad hoc, or it establishes clarity within mathematics but no credible rules for quantifying data from the *real-world* setting. In maths textbooks this is commonplace with context-based exercise problems (Burke et al. 2016, p. 5–6). The setting of problems in an apparently real-world context and the requirement for a solution determined using some mathematical principles do not in itself provide a pedagogic activity in mathematical modelling/problem solving.

As described earlier, the PRIMAS project refers to a teaching culture which is supportive of student inquiry. In Chap. 1, we have suggested that the nature of the problem to be solved determines the nature of the problem solving process and hence the competences that can be developed though engagement with them. Thus, teacher practice is central to the possibility for inquiry. An important contribution to inquiry-based learning in mathematics was the CAME project in the UK started in 1993. This was built around the deployment of 30 *thinking maths* activities, with a strong emphasis on discussion, pair and group work and student presentation. Notably, there was a very strong element of teacher professional development. Shayer and Adhami (2007) in their post project retrospective state that:

> The mathematics teachers were encouraged, as part of their PD, to establish connections between the agenda of the CAME lessons, and the contexts of their ordinary mathematics lessons using the same reasoning patterns. […] In effect many of them were taking a ‘Thinking Maths’ approach into all their teaching, and by implication encouraging their students to take a thinking approach to their learning, which seems to have affected their learning in other subjects as well (Shayer and Adhami 2007, pp. 287–88).

A core element of the pedagogy, which they have developed from the inquiry-based learning and thus brought into their general teaching, is exemplified thus:

> At this point the teacher, rather than spending time going round to groups 'helping' instead listens, sees and notes where each group has got to, and, depending on the different aspects of working on the underlying mathematical ideas he finds, makes a plan of which groups, and in what order, he will ask to contribute to Act 3. He may occasionally throw in a strategic question if he sees a group is stuck (ibid, p. 275).

This is strongly resonant with the principles set out for assessment for learning initiated by Black and colleagues at the same institution (Black et al. 2004). Maaß and Doorman (2013) describe the teacher's role in PRIMAS: 'Teachers are proactive: they support pupils who are struggling and challenge those who are succeeding through the use of carefully chosen strategic questions' (p. 887). This intense multilayered approach was also the expectation with CAME. This is clearly complex and thus expensive, but the hope for student competences is also complex and as we have seen requires sophisticated task design and teaching to enable leaners to be apprenticed into the definitive mathematisations required for mathematical modelling and thus real-world problem solving, as in the example below.

Intriguingly, the teachers' beliefs of the nature of mathematics in itself seem to have an effect on their students' measured school mathematics achievement. Askew et al. (1997), studying primary school teachers' practice in the Effective Teachers of Numeracy project in the UK, reported that teachers who believed that mathematics was a multiply interconnected subject (referred to as *connectionist*) were most effective in terms of the student outcomes they supported. This is understood through the pedagogy that this thinking enabled: 'The connectionist teachers' lessons were generally characterised by a high degree of focused discussion between teacher and whole class, teacher and groups of pupils, teacher and individual pupils and between pupils themselves' (ibid, p. 46).

We have seen that PISA test practices do not incorporate all aspects of real-world problem solving despite a strong urge to do so. In Denmark, a wide ranging curriculum reform allows the possibility for corresponding developments in assessment systems. There is considerable discussion of the varied nature of an assessment system that would support the new competence-based syllabus, both new and old: 'However, there is still a great need for continuously devising, testing and new developments evaluating new test and examination forms' (Niss and Højgaard 2011, p. 144). However, the outcomes of the CAME project and the Effective Teachers of Numeracy project suggest that intense and long-term intervention in teacher development in inquiry-based learning has the potential to produce significant gains in students' general mathematical performance. In the PRIMAS and ASSIST-ME projects, there is a desire for change in assessment systems; however, there appear to be benefits from the inquiry-based approach developed in these projects even within existing systems.

Inquiry-Based Learning Example from the Assist-Me Project (http://assistme.ku.dk)

The Towers of Hanoi. Students in small groups solve this classic wooden puzzle, to move piles of different-sized discs from the one end of three pegs to the other end peg one at a time, never placing a larger disc on a smaller one (Fig. 2.5).

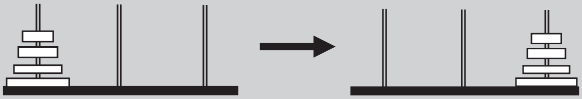


Fig. 2.5 The Towers of Hanoi

When successful they repeat the tasks sufficiently often that they can do this reliably and feel confident they have minimised the number of moves for a given number of discs. They capture the moves in diagrammatic/symbolic/textual form to report their *method*. They vary the number of discs looking for relationships between the number of discs and the minimum number of moves as a direct relation and as a recurrence relation. They look to explain why the direct relation must always hold true and generate embryonic proofs (potentially by induction). The teacher intervention is as characterised in the reports from CAME and PRIMAS, with small group work, focused discussion and student presentation. The teacher does not hint towards an outcome, but prompts for a process to continue. This generates competences *within* mathematics as described in the KOM project and is an example of a definitive mathematisation as a mathematical modelling practice, since the internal syntax generates a definitive proof and the relationship between the real-world (wooden puzzle) setting and its quantification (number of moves) is very clearly described.

The Concept of Innovation Competence

In Chap. 1, it was argued that one of the focal competences of the twenty-first-century skills programme is innovation competence. In this context, teaching for innovation is understood as mono- or interdisciplinary teaching activities in which students work on using their disciplinary knowledge and skills in order to improve on an authentic *field of practice*. Here, field of practice is meant in the broadest possible sense as ranging from the performance of very specific activities such as the practice of showering in the morning to complex clusters of activities such as the practice of getting rid of waste at music festivals. The previous chapter also described that innovation competence can be operationalised as students' ability (alone or in collaboration with others) to (a) generate solutions to issues while drawing on their disciplinary knowledge and their analysis of the field of practice where the issue arises, (b) analyse and reflect on the value-creating potential and realisability of their ideas, (c) work towards implementing their ideas and (d) communicate about their ideas to various stakeholders (cf. Nielsen and Holmegaard 2015). As a learning goal, innovation competence involves five dimensions: creativity, collaboration, (disciplinary) navigation, implementation and communication. Each dimension is described in Chap. 1.

In teaching for innovation, there is no theoretically correct answer to the tasks the students are doing, and the teacher is not the disciplinary expert who knows the way to solve the problem or do the task. Thus the teacher's role is as a supervisor, facilitator or guide very similar to the teacher's role in inquiry teaching. Indeed, similar to inquiry teaching, teaching for innovation shifts the focus of formative assessment in comparison to regular mono-disciplinary teaching. Innovation competence is very much a process competence. Therefore, the formative assessment should be directed at facilitating that students become more able to work in specific processes, rather than facilitating that students master a specific disciplinary content (e.g. Harlen 1999). Figure 2.6 shows one kind of model (the double diamond model) that can capture archetypical innovation work processes. The Polluted Seawater task provides an example of a comprehensive activity that roughly follows the double diamond model.

As argued in Chap. 1, innovation competence can be seen as a complex of five dimensions. This division can help teachers and educators to operationalise the competence for designing prospective activities and assess students' competence development formatively and summatively. A generic way of spelling out the five dimensions would be the following (a richer description was developed in Nielsen 2015a):

- Creativity:
  - The student independently finds or independently interprets a given problem issue from a field of practice.
  - The student generates a range of ideas or solutions to a problem rather than just one idiosyncratic type of idea.
  - The student works with generated ideas in a critical fashion, e.g. by evaluating, sorting, revising and expanding the ideas of herself or others.
- Collaboration:
  - The student takes responsibility for and facilitates that the collaborative group finishes its tasks, e.g. by being able to identify how the competences of the people in the group can complement each other.
  - The student includes others and is flexible in a collaboration, e.g. by being able to work with many different types of stakeholder or people, rather than just a limited number of people or classmates.
- Navigation:
  - The student interprets a specific problem from practice as a problem that can be approached from a disciplinary perspective, e.g. by being able to translate the problem into disciplinary language.

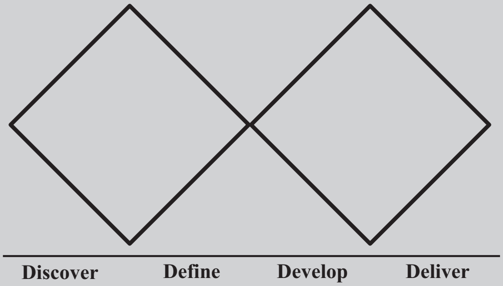
- The student functionally handles knowledge, e.g. by handling plentiful and heterogeneous information and sorting and prioritising which information is the most important to go into detail with.
- The student masters complex work processes.

- Implementation:
  - The student makes informed decisions about what actions to take in a specific time in a work process.
  - The student takes action outside his/her comfort zone (e.g. by seeking information outside the classroom).
  - The student takes risks and puts him/herself and others into play, e.g. by not stopping at the level of an idea but carrying out that idea.

The Double Diamond Model

This model, constructed by the Design Council (2005), can be seen as representative of typical teaching activities that aim to foster all five dimensions of innovation competence (the original model is more focused on design processes). In the *Discove*r phase, students make inquiries into the field of practice that they are working on with the aim of identifying factors and aspects of the specific problem. In teaching for innovation, this will involve inquiries into the relevant disciplinary subject areas as well as information relevant to the field of practice and its stakeholders, for example, information about what leads to the specific problem or how the problem is currently handled in the field of practice. In the *Define* phase, students converge on a focal factor or aspect that they would like to improve; here the aim is to delimit the problem area and define possible success criteria for improvement. In the *Develop* phase, students generate ideas for improving the delimited problem. This may involve multiple cycles of generating, testing and revising ideas (and possibly prototypes). In the *Deliver* phase, the proposed solution is finalised and handed over to the field of practice, typically by communicating an idea or presenting a prototype to relevant stakeholders (a number of other similar models for designing teaching for innovation are available at https://innovationenglish.sites.ku.dk).

Fig. 2.6 The double diamond model



- Communication:
  - The student assesses how to communicate (e.g. to stakeholders) in a given situation.
  - The student masters a range of communication techniques and genres.
  - The student communicates in an engaging and convincing manner.

Students can acquire competences and skills relevant for innovation in different degrees of comprehensiveness. It is entirely possible for a teacher to focus on one or two of the dimensions of innovation competence. In a recent Danish attempt to trial examination formats for innovation competence (Nielsen 2015b), teachers elected to have some activities cover all five dimensions, while other activities strategically focused on one or more dimensions. For example, one could imagine a class working intensively on developing the collaboration dimension by working in groups on some interdisciplinary content under the observation of the different teachers involved, with pauses in the group work where the teachers, based on their observations, can provide formative feedback to individual students on how they have observed the students' collaboration skills and provide improvement strategies for them.

Innovation-Oriented Learning Example

Polluted Seawater. The project is started by a marine biologist from the municipality who introduces the students to a problem related to seawater quality. Students work in groups experimentally and/or by means of data processing to investigate/document the problem, its cause and its extent. The groups must generate possible solutions to improve the seawater quality and discuss their practical realisability, benefits and consequences. The end product is a proposed solution from each group which is communicated (e.g. as a poster exhibit) to the marine biologist.

This activity is divided into two main phases:

Phase 1 (Biological inquiry): Inquiry process of possible causes to the problem. This phase could include measurement of nitrate (or phosphate) and *E. coli* concentration in water samples from selected sites (rivers and sea) possibly before and after rain, constructing plots for biochemical oxygen demand (BOD5) for selected streams of water into the sea and identifying experts and authorities who can provide knowledge and inspiration; the marine biologist can also be contacted if groups need more information or have questions about the local conditions.

Phase 2 (Generating solutions): The groups work on ideas for possible solutions. The task is to narrow in on the possible causes that each group wants to work with, in order to target the proposed solutions. This phase resembles an inquiry process, but aims to identify viable solutions to the issue and testing of or reflection on their realisability and potential for value creation.

Summary and Discussion

In the last decades, inquiry, or as it is more recently described, engaging students in scientific practices, has become a fundamental approach in science teaching and learning (National Research Council 1996, 2012). Its importance has been mirrored at the European level in the funding of several EU projects (e.g. S-TEAM, ESTABLISH and PRIMAS) aiming to support science teachers in implementing the approach. Despite its prominent role in science education research, however, no general agreement about the exact definition of scientific inquiry exists. From a holistic perspective, inquiry-based approaches to science teaching and learning generally try to imitate more or less authentic scientific investigations embedded in real-world contexts. They involve presenting learners with problems and questions and supporting them in identifying ways of solving these problems by applying the thinking and working processes of scientists and by drawing on their knowledge of scientific content and the nature of science with the aim of constructing new knowledge.

Looking at inquiry-based approaches across the domains of science, technology and mathematics, the concept seems to be strongly related to the field of science education (Ropohl et al. 2013). In mathematics and technology education similar concepts exist; however, they usually go under different names. In mathematics education inquiry is manifested in two different ways: firstly learners exploring mathematical problems, developing their own mathematics and working towards solutions and their proof or secondly learners using techniques of mathematical modelling to support elements in the process of solving problems originating from outside of mathematics. A clear distinction thus exists between those problems where problem and solution reside within mathematics and those where the problem originates outside of mathematics, the latter necessarily being examples of mathematical modelling. This approach has been referred to as problem-based learning which 'describes a learning environment where problems drive the learning' (Rocard et al. 2007, p. 9). A significant difference with scientific inquiry is the focus on the mathematical development towards deduction and proof, often with a corresponding lack of interest in the actual problem resolution, where the solution 'is presented as a deduction from what was given in the problem to what was to be found or proved' (Schoenfeld and Kilpatrick 2013, p. 908). In technology education, the closest connection to inquiry is provided by approaches to teaching and learning using the concept of design processes. Inquiry in technology education involves presenting learners with challenges, problems and scenarios and supporting learners to identify ways of addressing these through iterative design processes that draw on critical thinking, creative and exploratory idea development and effective and thoughtful outcome resolution. Scientific inquiry and design processes are closely related – Lewis (2006) even proposes that 'design and inquiry are conceptual parallels' (p. 255) since they converge on many dimensions like, e.g. they are both reasoning processes including elements of uncertainty and the need for testing, evaluating and decision making, they both depend on content knowledge and they both work under domain-specific constraints. The major distinguishing characteristic is a difference in purpose. Whereas pure science is inherently speculative, the purpose of technology is invariably instrumental: 'The goal of science is to understand the natural world, and the goal of technology is to make modifications in the world to meet human needs' (National Research Council 1996, p. 24; also Lewis 2006).

One of the focal competences of the twenty-first-century skills programme is innovation competence. In the case of teaching for innovation, inquiry involves presenting learners with real-world problems and supporting them to identify realisable and value-generating solutions to these problems through iterative processes that draw on idea generation, disciplinary navigation, collaboration, implementation and communication. Obviously scientific inquiry, design processes and teaching for innovation have a number of similarities. However, teaching for innovation differs significantly from inquiry teaching in science because the former always begins with a problem from an authentic field of practice in the real world that students work to solve or alleviate – there is so to speak always a user (a person in the field of practice) who is the main addressee of the students' work. This is not necessarily the case in scientific inquiry teaching. Teaching for innovation also differs from design processes in the sense that the latter is typically taught in a specific discipline, design, engineering or technology. Teaching for innovation does not fall under the purview of a specific discipline but is a possible extension of every existing discipline.

Despite these domain-specific differences in the understanding of inquiry, however, inquiry-based teaching and learning shares characteristics across domains that could be factored in a kind of *meta-definition* of inquiry: Scientific inquiry in science, problem solving in mathematics, design processes in technology and innovation as a cross-curricular approach all require students to become actively engaged in solving problems of a personal, societal, environmental or disciplinary nature by drawing on their disciplinary knowledge which involves both, knowledge about the content and the nature of their discipline, and by applying the domain-specific and generic thinking and working processes of scientists. As already mentioned in Chap. 1, this overarching understanding of inquiry across domains stresses again that competences and knowledge are always intertwined. Acting competently inevitably requires knowledge – the specific amount and type of knowledge that is necessary to solve a problem, however, may vary depending on the specific context in which the problem is embedded and the specific task that students are facing (see Chap. 1; Rönnebeck et al. 2016).

By involving students in inquiry processes or scientific practices, teachers can address complex subject-specific (e.g. carrying out investigations in science or designing a device addressing a specific need in technology) as well as more generic competences (e.g. developing explanations or arguments based on evidence or communicating efficiently). The role of the teacher thereby changes from primarily being the disciplinary expert and conveyor of knowledge to becoming a facilitator who guides the students through their learning providing disciplinary knowledge when needed.

In order to do this, the teacher takes on multiple roles like motivator, diagnostician and guide but also as collaborator, mentor, modeller and learner (Crawford 2000). In a similar way, the role of the students changes. Instead of being mere passive recipients of instruction, they need to become active participants in their learning processes. In inquiry settings, the traditional distinction between the teacher as the knowledge giver and the student as the knowledge receiver is replaced by the teacher working collaboratively with his or her students in order to construct understanding. To effectively support students' learning in inquiry settings, teachers need to actively guide their students through the inquiry process by creating opportunities to learn, encouraging students to become active learners and providing scaffolds and support when needed. Taking on this multitude of roles is demanding for the teacher and requires a high level of expertise, a great level of involvement and a willingness 'to embrace inquiry as a content and pedagogy' (Crawford 2000, p. 933).

New learning goals moreover require new forms of assessment that allow for the assessment of complex, process-oriented competences and acknowledge the active role of the students (see also Chaps. 1 and 3). Across all domains, developing and implementing such assessments, whether for formative or summative purposes, is a complex and challenging task. The challenges that researchers, teacher educators and teachers face include reaching a shared understanding of the learning goals and competence expectations, defining what counts as evidence of achievement as well as ensuring reliability and validity (see also Chaps. 1 and 3). Against this background, formative assessment could offer promising perspectives because of its inherent emphasis of active student engagement and its potential for supporting complex learning processes by defining learning goals and competence expectations and by providing feedback to students on that basis. Examples of formative assessment methods used to assess inquiry competences in the different domains will be presented in the following chapters.

References

Abd El Khalick, F., Boujaoude, S., Duschl, R. A., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., et al. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397–419.

Anderson, R. D. (2002). Reforming science teaching: What research says about inquiry. *Journal of Science Teacher Education, 13*(1), 1–12.

Askew, M., Brown, M., Rhodes, V., Johnson, D., & Wiliam, D. (1997). *Effective teachers of numeracy*. London: King's College London.

Atkinson, S. (1999). Key factors influencing pupil motivation in design and technology. *Journal of Technology Education, 10*(2), 4–26.

Barak, M., & Awad, N. (2008). *Learning processes in information system design*. Paper presented at the PATT 20: Critical issues in technology education, Tel Aviv, Israel.

Barron, B., & Darling-Hammond, L. (2008). Teaching for meaningful learning: A review of research on inquiry-based and cooperative learning. In L. Darling-Hammond, B. Barron, P. D. Pearson, A. H. Schoenfeld, E. K. Stage, T. D. Zimmermann, … (Eds.), *Powerful Learning. What we know about teaching for understanding.* San Francisco, CA: Jossey-Bass.

Bell, T., Urhahne, D., Schanze, S., & Ploetzner, R. (2010). Collaborative inquiry learning: Models, tools, and challenges. *International Journal of Science Education, 32*(3), 349–377.

Binkley, M., Erstad, O., Herman, J. L., Raizen, S., Ripley, M., Miller-Ricci, M., & Rumble, M. (2012). Defining twenty-first century skills. In P. E. Griffin, B. McGaw, & E. Care (Eds.), *Assessment and teaching of 21st century skills* (pp. 17–66). Dordrecht/New York: Springer.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the black box: Assessment for learning in the classroom. *Phi Delta Kappan, 86*(1), 8–21.

Botleng, V. J., Brunel, S., & Girard, P. (2016). *The digital fabrication laboratories (Fab labs) platform: A dynamic hands-on, minds-on and hearts-on approach to augment STEM education activities and 21st century skills.* Paper presented at the PATT 32: Technology education for 21st century skills, Utrecht, Netherlands.

Burke, J., Jablonka, E., & Olley, C. (2016). A firm foundation or shifting sands: Mathematisation and evaluation strategies. In G. Wake et al. (Eds.), *Modelling perspectives: Looking in and across boundaries*. Berlin: Springer.

Crawford, B. A. (2000). Embracing the essence of inquiry: New roles for science teachers. *Journal of Research in Science Teaching, 37*(9), 916–937.

Crismond, D. (2011). Scaffolding strategies for integrating engineering design and scientific inquiry in project-based learning environments. In M. Barak & M. Hacker (Eds.), *Fostering human development through engineering and technology education* (pp. 235–256). Rotterdam: Sense Publishers.

Department of Basic Education. (2011). *Curriculum and assessment policy statement: Grades 7–9 technology*. Republic of South Africa: Department of Basic Education.

DES/WO. (1988). *National Curriculum Design and technology working group interim report*. London: HMSO.

Design Commission. (2011). *Restarting britain, design education and growth*. London: Policy Connect.

Dowling, P. (2007). *Sociology as method: Departures from the forensics of culture, text and knowledge*. Rotterdam: Sense.

European Commission. (2004). *Increasing human resources for science and technology in Europe*. Brussels: European Commission.

Furtak, E. M. (2006). The problem with answers: An exploration of guided scientific inquiry teaching. *Science Education, 90*(3), 453–467.

Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. *Review of Educational Research, 82*(3), 300–329.

Gardner, J., Harlen, W., Hayward, L., Stobart, G., & Montgomery, M. (2010). *Developing teacher assessment*. Maidenhead: Open University Press.

Hadfield, M. (1995). Das kupfer-problem [the copper problem]. *ChemKon, 2*(3), 103–106.

Harlen, W. (1999). Purposes and procedures for assessing science process skills. *Assessment in Education: Principles, Policy & Practice, 6*(1), 129–144.

Harlen, W. (2007). *The quality of learning: Assessment alternatives for primary education*, *Primary review research survey 3/4*. Cambridge: University of Cambridge Faculty of Education.

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99–107.

Hume, A., & Coll, R. K. (2010). Authentic student inquiry: The mismatch between the intended curriculum and the student-experienced curriculum. *Research in Science & Technological Education, 28*(1), 43–62.

International Technology Education Association. (2007). *Standards for technological literacy: Content for the study of technology*. Reston: International Technology Education Association.

Isaacs, T. (2010). Educational assessment in England. *Assessment in Education: Principles, Policy & Practice, 17*(3), 315–334.

Jones, A. (1997). Recent research in learning technological concepts and processes. *International Journal of Technology and Design Education, 7*(1–2), 83–96.

Kelly, A. V., Kimbell, R. A., Patterson, V. J., Saxton, J., & Stables, K. (1987). *Design and technology: A framework for assessment*. London: HMSO.

Kessler, J. H., & Galvan, P. M. (2007). *Inquiry in action: Investigating matter through inquiry. A project of the American Chemical Society Education Division*, Office of K–8 Science: American Chemical Society. Retrieved from http://www.inquiry-inaction.org/ download/. Accessed 19 Sept 2016.

Kimbell. (2012). Evolving project e-scape for national assessment. *International Journal of Technology and Design Education, 22*(2), 135–155.

Kimbell, R., & Stables, K. (2007). *Researching design learning: Issues and findings from two decades of research and development* (Hardback ed.). Berlin: Springer.

Kimbell, R., Stables, K., Wheeler, T., Wozniak, A., & Kelly, A. V. (1991). *The assessment of performance in design and technology*. London: SEAC/HMSO.

Kimbell, R., Stables, K., & Green, R. (1996). *Understanding practice in design and technology*. Buckingham: Open University Press.

Kimbell, R., Miller, S., Bain, J., Wright, R., Wheeler, T., & Stables, K. (2004). *Assessing design innovation: A research and development project for the Department for Education & skills (DfES) and the qualifications and curriculum authority (QCA)*. London: Goldsmiths, University of London.

Kimbell, R., Wheeler, T., Stables, K., Shepard, T., Martin, F., Davies, D., et al. (2009). *E-scape portfolio assessment: A research & development project for the Department of Children, families and schools, phase 3 report*. London: Goldsmiths, University of London.

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*, 75–86.

Lewis, T. (2006). Design and inquiry: Bases for an accommodation between science and technology education in the curriculum? *Journal of Research in Science Teaching, 43*(3), 255–288.

Lewis, T., & Zuga, K. (2005). *A conceptual framework of ideas and issues in technology education*. Washington, DC: National Science Foundation.

Linn, M. C., Davis, E. A., & Bell, P. (Eds.). (2004). *Internet environments for science education*. Mahwah: Lawrence Erlbaum Associates Publishers.

Maaß, K., & Artigue, M. (2013). Implementation of inquiry-based learning in day-to-day teaching: A synthesis. *ZDM Mathematics Education, 45*, 779–795.

Maaß, K., & Doorman, M. (2013). A model for a widespread implementation of inquiry-based learning. *ZDM Mathematics Education, 45*, 887–899.

McLaren, S. V. (2007). An international overview of assessment issues in technology education: Disentangling the influences, confusion and complexities. *Design and Technology Education: An International Journal, 12*(2), 10–24.

McLaren, S. V. (2012). Assessment is for learning: Supporting feedback. *International Journal of Technology and Design Education, 22*(2), 227–245.

Merrill, C., Reese, G., & Daugherty, J. (2010). Mathematics education. In J. Ritz & P. Reed (Eds.), *Research in technology education: 59th yearbook of the Council on technology teacher education* (Vol. 59, pp. 172–191). Muncie: CTTE.

Miller, J. (2011). *What's wrong with DT?* London: RSA.

Ministry of Education. (2010). *Technology curriculum support*. Wellington: techlink.org.nz..

Moreland, J., Jones, A., & Barlex, D. (2008). *Design and technology inside the black box: Assessment for learning in the design and technology classroom*. London: GL Assessment.

National Research Council. (1996). *National science education standards*. Washington, DC: The National Academies Press.

National Research Council. (2012). *A framework for K-12 science education: Practices, crosscutting concepts, and core ideas*. Washington, DC: The National Academies Press.

Nielsen, J. A. (2015a). Assessment of innovation competency: A thematic analysis of upper secondary school teachers' talk. *The Journal of Educational Research, 108*(4), 318–330.

Nielsen, J. A. (2015b). *Rapport fra arbejdsgruppe for prøveformer der tester innovationskompetencer i gymnasiet [report from the working group on testing innovation competency in high school]*. Copenhagen: Department of Science Education.

Nielsen, J. A., & Holmegaard, H. T. (2015). Innovation and employability: Moving beyond the buzzwords - a theoretical lens to improve chemistry education. In I. Eilks & A. Hofstein (Eds.), *Relevant chemistry education – From theory to practice* (pp. 317–334). Rotterdam: Sense Publishers.

Niss, M. (2015). Modelling as a mathematical competency: A paradox? In G. Kaiser & H.-W. Henn (Eds.), *Werner blum und seine Beiträge zum modellieren im mathematikunterricht, Realitätsbezüge im mathematikunterricht [Werner blum and his contributions to modeling in mathematics instruction, relations to reality in mathematics instruction]* (pp. 269–276). Wiesbaden: Springer Fachmedien.

Niss, M., & Højgaard, T. (2011). *Competencies and mathematical learning: Ideas and inspiration for the development of teaching and learning in Denmark (IMFUFA tekst)*. Roskilde: Roskilde University.

NWSEB. (1970). *A course of study in design*. Manchester: North Western Secondary School Examinations Board.

OECD. (2014). *PISA 2012 results: What students know and can do (volume I, revised edition, february 2014)*. Paris: Organisation for Economic Co-operation and Development.

Ofsted. (2002). *Secondary subject reports 2000/01: Design and technology*. London: Department for Education and Employment.

Ofsted. (2011). *Meeting technological challenges? Design and technology in schools 2007–2010*. London: Ofsted.

Ofsted. (2012). *Making a mark: Art, craft and design education 2008/11*. London: Ofsted.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., et al. (2015). Phases of inquiry-based learning: Definitions and the inquiry cycle. *Educational Research Review, 14*, 47–61.

Pellegrino, J. W., & Hilton, M. L. (2012). *Education for life and work: Developing transferable knowledge and skills in the 21st century*. Washington, DC: The National Academies Press.

Pollitt, A. (2012). Comparative judgment for assessment. *International Journal of Technology and Design Education, 22*(2), 157–171.

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., & Hemmo, V. (2007). *Science education now: A renewed pedagogy for the future of Europe*. Brussels: European Commission.

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education, 52*(2), 161–198.

Ropohl, M., Rönnebeck, S., Bernholt, S., & Köller, O. (2013). *Report from the FP7 project: Assess inquiry in science, technology and mathematics education. A definition of inquiry-based STM education and tools for measuring the degree of IBE* [deliverable D2.5]. Copenhagen: ASSIST-ME project. Retrieved from http://assistme.ku.dk/project/workpackages/wp2/131015_del_2_5_IPN.pdf. Accessed 07 Nov 2016.

Schoenfeld, A. H., & Kilpatrick, J. (2013). A US perspective on the implementation of inquiry-based learning in mathematics. *ZDM – The International Journal on Mathematics Education, 45*(6), 901–909.

Schools Council. (1975). *Education through design and craft: Schools Council design and craft education project*. London: Edward Arnold.

Seery, N., Canty, D., & Phelan, P. (2012). The validity and value of peer assessment using adaptive comparative judgement in design driven practical education. *International Journal of Technology and Design Education, 22*(2), 205–226.

Shavelson, R. (2011). *An approach to testing and modeling competence*. Paper presented at the Bad Honnef Conference on Teachers' Professional Knowledge, Bad Honnef, Germany.

Shayer, M., & Adhami, M. (2007). Fostering cognitive development through the context of mathematics: Results of the CAME project. *Educational Studies in Mathematics, 64*, 265–291.

Snape, P., & Fox-Turnbull, W. (2013). Perspectives of authenticity: Implementation in technology education. *International Journal of Technology and Design Education, 23*(1), 51–68.

Spendlove, D., & Hopper, M. (2006). Using 'electronic portfolios' to challenge current orthodoxies in the presentation of an initial teacher training design and technology activity. *International Journal of Technology and Design Education, 16*(2), 177–191.

Stables, K. (2013). *Social and cultural relevance in approaches to developing designerly well-being: The potential and challenges when learners call the shots in Design and Technology projects*. Paper presented at the PATT 27: Technology education for the future: A play on sustainability, Christchurch, New Zealand.

Stables, K., Dagan, O., & Davies, D. (2015). Blended learning and assessment through dynamic digital portfolios: The e-scape approach. In S. Koç, X. Liu, & P. Wachira (Eds.), *Assessment in online and blended learning environments*. Charlotte: Information Age Publishing.

Stables, K., Kimbell, R., Wheeler, T., & Derrick, K. (2016). *Lighting the blue touch paper: Design talk that provokes learners to think more deeply and broadly about their project work*. Paper presented at the PATT 32: Technology education for 21st century skills, Utrecht, Netherlands.

Strimel, G. (2015). *Cognitive processes as indicators for student aptitude in engineering design*. Paper presented at the PATT 29: Plurality and complementarity of approaches in design and technology education, Marseille, France.

Turnbull, W. (2002). The place of authenticity in technology in the New Zealand curriculum. *International Journal of Technology and Design Education, 12*(2), 23–40.

White, B. Y., & Frederiksen, J. R. (1998). Inquiry, modeling, and metacognition: Making science accessible to all students. *Cognition and Instruction, 16*(1), 3–118.

Chapter 3
Exploring Relations Between Formative and Summative Assessment

Jens Dolin, Paul Black, Wynne Harlen, and Andrée Tiberghien

The goals of STEM education go beyond specific knowledge and skills. They include a wider range of competencies and contribute, as does work in other domains, to what are variously described as twenty-first-century skills or life skills necessary for personal fulfilment, employment, citizenship and social responsibility. For example, the countries of the European Union have agreed upon a framework of key competencies that are being reflected in the school curricula of most of these countries (European Commission 2011). They include competencies in communication, mathematics, science and technology, learning to learn, social and civic responsibility, initiative and cultural awareness. Similarly, in response to the rapid changes in technology that make information readily available, the countries of the OECD are emphasising the importance of competencies relating to selecting, synthesising, transforming and applying information, thinking creatively, collaborating with others and communicating effectively (OECD 2013).

These general educational trends have influenced the STEM learning goals (as described in Chap. 1). The generic competencies have to be learned while working with subject-specific content and problems, changing both curriculum and pedagogy. Simultaneously, the STEM domains have developed their curriculum goals and descriptions to focus on some domain-specific competencies, like modelling and investigating.

J. Dolin (✉)
Department of Science Education, University of Copenhagen, Copenhagen, Denmark
e-mail: dolin@ind.ku.dk

P. Black
School of Education, Communication and Society, King's College, London, UK

W. Harlen
Graduate School of Education, University of Bristol, Bristol, UK

A. Tiberghien
University of Lyon and Centre National de la Recherche Scientifique, ICAR (CNRS), Lyon, France

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_3

Achieving these new goals may require changes to be made, not only in curriculum content and pedagogy but also in assessment content and tools, given the well-established impact of what is assessed on what is taught and how it is taught (Harlen and Deakin Crick 2003). The central goal is to improve learning, and assessment is a tool to help achieve this goal; however, unless the full range of goals is assessed, those goals not included are likely to be given less attention. As a result, developing new competencies may well remain as an aspiration on paper unless appropriate assessment tools and processes are developed. Developing assessment approaches able to capture these new goals is therefore necessary and at the heart of many recent projects such as SAILS (Strategies for Assessment of Inquiry Learning in Science – an EU FP7 project), TAPS (Teacher Assessment in Primary Science – a Primary Science Teaching Trust project) and the ASSIST-ME project. The ASSIST-ME project has researched the implementation in the partner countries of the formative use of four assessment methods able to capture inquiry processes within STM subjects. Ideally, the high-stakes assessment that influences teaching practice should use similar approaches. As it was beyond the scope of ASSIST-ME to determine national examinations, the project pragmatically investigated how the formative use of the four assessment methods are in alignment with the more high-stakes assessments, especially the final examinations, and how current high-stakes assessments should be changed to promote effective formative assessment.

This chapter sets these investigations into a theoretical frame by characterising two key purposes of assessment, formative and summative, and considering how they are related and can be brought together in developing a dependable approach to summative assessment using evidence collected and used in formative assessment. The third purpose of assessment, accountability, is dealt with as a special use of summative assessment. The range of approaches to using assessment formatively – to influence learning activities as they take place – and summatively, to record evidence of what has been learned at certain times, is discussed. Some examples from the ASSIST-ME project illustrate the variety of approaches to assessment and the overlapping relations between formative and summative use of assessment.

The Characteristics of Formative and Summative Assessment

The distinction between formative and summative as applied to student assessment emerged in the 1960s, having originated in identifying the roles of programme evaluation in the development of new curriculum materials. Evaluation described as formative was conducted during the process of materials development to provide information about how to revise early drafts, while summative evaluation usually provided some measure of the effectiveness of the final draft. So, in the case of student learning, the main purpose of formative assessment is seen as helping learning, while the main purpose of summative assessment is to provide information about what learning has been achieved at a certain time. Even major texts on pedagogy have until recently paid almost no attention to the potential of formative assessment to assist the day-to-day learning of students (Black 2016).

Before we can discuss how to combine aspects of assessment for formative and summative purposes, it is helpful to define and identify the characteristics they have in common and their differences. It is important to recognise that formative and summative refer to different *purposes* of assessment and not to different kinds or *forms* of assessment. As we will see later, data of the same kind can, in some circumstances, be used both formatively and summatively. It is the impact of assessment in a particular instance and the use made of the data that identifies the assessment as formative or summative and not the form of the data. This distinction is more easily and more often overlooked in the case of formative assessment, so we begin by looking at the characteristics of assessment used to enhance learning. However, in recognition of the way assessment language is commonly used in practice, we will often use the phrase 'formative assessment' for the more correct 'formative purpose of assessment' or 'formative use of assessment evidence'.

Formative Assessment

Although the concept of using assessment to identify difficulties in learning and to support growth in learning has been embraced in many countries across the world, there is considerable diversity in how assessment is translated into practice and conflicting views about how it is defined. In particular there is a division, to which Bennett (2011) draws attention, between 'those who believe formative assessment refers to an instrument (e.g., Pearson 2005), as in a diagnostic test, an 'interim' assessment, or an item bank from which teachers might create those tests' (Bennett 2011, p. 6) and those who view it as a process. There are considerable practical and implementation differences between these two views. Expressed rather starkly, the 'instrument' view involves the use of a tool such as a test or task that will yield information about current competencies, often in the form of a score or judgement. Seen in this way, formative assessment is much less dependent on the ability of teachers to ask appropriate questions or make relevant observations than is the case for the process view, which produces 'a qualitative insight into students learning rather than a score' (Shepherd 2008). Of course, this exaggerates the differences between these views, and in practice there is overlap between the two positions of how formative assessment is conceived and practised. Nevertheless, the most widely used definitions of formative assessment emphasise the process view, for instance:

> *Formative assessment is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.* (ARG 2002)

> *Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes.* (CCSSE, in McManus 2008)

It can in these definitions be assumed that the assessment activities make use of a variety of methods (which could be seen as instruments) to gain insight into present levels of student competence, and the process is very much about using this method to get access to students' learning. The ARG definition focuses on the learner and the learner's learning process, while the CCSSE definition embeds formative assessment in instruction in order to improve both students' learning and teachers' teaching.

The focus on process is underpinned by an individual or sociocultural constructivist (see later) concept of learning, that is, one that views learning as making sense of new experience starting from existing ideas and competencies. Evidence of what students already know and can do, in relation to short-term goals of a particular lesson or group of lessons on a topic, informs decisions about the steps needed to make progress. The process involves teachers and students in:

- Establishing clear goals, and progression steps (including criteria) towards them, that are shared between teacher and students
- Designing or choosing a method for collecting data about student performance
- Using the method for gathering evidence of where students are in relation to these goals
- Interpreting this evidence to decide what is needed to help students towards the goals
- Deciding what action can help the student through the next progression step

In Fig. 3.1 these actions are connected by arrows into a repeated cycle of events (the arrows leading away from the cycle to the right are explained later). A, B and C indicate activities that lead students to work towards the goals. Evidence gathered in activity A is interpreted in relation to where students are in relation to the goals. This information facilitates decisions about appropriate next steps and how to take them, leading to activity B. The cycle is repeated, and the effects of decisions at one time are assessed at a later time as part of the ongoing process. How often this happens depends on the nature of the goals; a cycle might take place over several lessons or might take only a few minutes so that it is carried out several times in one lesson.

In practice the process may not be as formal and teacher directed as it appears in this description. The actions are not 'stages' in a lesson; rather, the cycle represents a framework for helping teachers and students to become more conscious of what is involved in learning and using this to help further learning and also to adapt teaching. The learners are at the centre of the process, and the two-headed arrows indicate their role as both providers of evidence and receivers of information to improve their learning. This gives students a role in their own assessment that helps them to come to understand the process of learning, to work towards explicit goals and standards and to modify what they do in relation to constructive task-related feedback

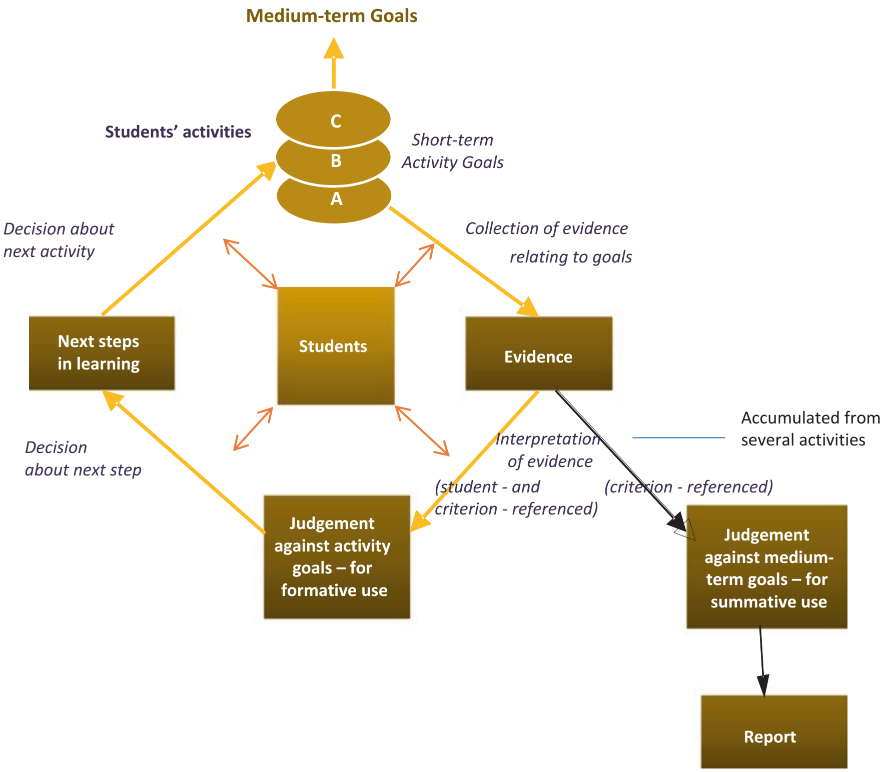


Fig. 3.1 The assessment process for both summative and formative purposes

from teachers and from other students and even from themselves. Besides enhancing learning, this involvement of the student in his or her learning process also engages and motivates students. This will further enhance learning and also guide students to a more self-directed learning (Wiliam 2011). Thus, activating students as instructional resources for one another will often be an important part of the teacher's task.

Parts of the formative cycle can be carried out in different ways depending on the type of activity and the age and ability of the students.

Ways in which evidence can be collected include:

- Observing students as they work, asking questions to probe their understanding, listening to their explanations and engaging in dialogue (ephemeral evidence often collected 'on-the-fly')
- Studying the outcome of their work, such as drawings, video and reports
- Embedding special tasks designed to require particular ideas and competencies in regular work
- Giving tests or quizzes (teacher made or externally produced)

Chapters 4, 5, 6 and 7 in this book describe different ways of collecting and judging evidence of student learning. The evidence gathered is judged in terms of what it indicates about existing ideas and competencies required to meet the lesson goals. The judgement may be made:

- By reference to a fixed set of predetermined criteria which provide concise descriptions of what students are expected to know and be able to do at a specific stage of their education (criterion-referenced)
- By reference to what is usual for similar students to be able to do – presupposing a normal distribution, that is, a bell-shaped curve, of students' performance (norm-referenced)
- By reference to expectations of the performance of particular students, taking into account what the student was previously able to do and the progress that the student has made over time (student-referenced or ipsative)
- A mixture of these

Judgement of the evidence in terms of progress towards the learning goals provides information to enable the teacher and students to decide what next steps are needed or, indeed, whether no immediate new action is needed if an activity is proceeding productively. Note that a distinction is made here between *evidence* (what was said, written or created) and *judgement* (interpreting what this means for the degree of goal attainment or for progress of individual or groups of students). This distinction is relevant to using assessment data for formative and summative purposes.

In formative assessment, taking the next steps involves feedback into the teaching-learning process. Feedback here, as in other situations, means giving responses to a product or process or event to improve performance. In the context of teaching, consistent with the social constructivist view of learning, feedback is from student to teacher, from teacher to student or from student to student. Feedback from student to teacher enables the teacher to know how students are responding to the learning activities and enables the teacher to know what action to take to adjust the opportunities and challenges provided to students. Such feedback may be used by teachers to adjust teaching in future planning of similar activities for other students. Feedback from teachers to students should give students information about how they can improve their work or take their learning forward. Just giving marks or grades that only indicate how well the work is judged to be is not consistent with the aim of using assessment to help learning. A controlled experimental study by Butler (1987, 1988) found that giving marks, with or without added comments, was less effective in improving students' work than giving comments only, while the extensive research studies of Dweck (2000) show that feedback that includes marks can have long-term negative effects on students' beliefs about their own ability to learn. What seems to happen is that students are looking for judgements rather than help with further learning; they seize upon marks and ignore any accompanying comments. Student to student feedback (peer feedback) may take place in the course of dialogue among students when they discuss the strengths and weaknesses of one another's work (see Chap. 6).

This 'constructive' use of formative assessment hinges on the ability of the teacher (or another provider of feedback) to actually give recommendations that are relevant and effective for improvement. Many of the findings in chapter four to seven point at this as the key issue in the implementation of formative assessment methods. The teacher needs to have a well-established pedagogical content knowledge, and the students need to learn how to give and receive feedback.

A Rationale for Formative Assessment

Formative assessment, unlike summative assessment, is not always a formal requirement of a teacher's role. Teachers generally have the duty to conduct some form of assessment in order to keep records of students' achievement and to report, for instance, regularly to parents. In some of the countries participating in ASSIST-ME, the formal requirements of teaching seemed more about judgement than about learning. If there is no similar imposed requirement for formative assessment, teachers and others may need to be persuaded of its value if they are to make the effort to include it in their practice.

Across the countries participating in ASSIST-ME, the hindrance most often expressed by teachers using formative assessment in their teaching was lack of time. Having been introduced to various assessment methods and been working with the implementation for formative purpose for some time, all teachers agreed upon the usefulness and effectiveness of formative assessment. They also agreed that students were satisfied and that learning was enhanced. But 'how do we get the time for it?' was a common question.

The teachers often perceived formative assessment as an add-on to the teaching, something extra instead of understanding it as a central and integrated part of teaching and learning. And in a way, they were right: learning takes time. Teachers' perception of not having time for formative assessment might simply be an illustration of the long-known fact that most curricula are overloaded, and in order to cover it, the focus is more on teaching than learning. But it might also be an illustration of what you choose to spend time on relates to what you think matters.

An important rationale for formative assessment practices follows from current perceptions of how learning takes place. For some time, it has been recognised that learners have an active role in constructing their understanding; it is not something that can be received ready-made from others, as in the more simple versions of the theory of learning described as *behaviourist.* Rather, we recognise that developing understanding requires active participation of learners in constructing their learning. This accords with a *cognitive constructivist* view of learning, that is, learners making sense of new experiences starting from existing ideas and competencies.

Recognition that learning is not entirely, or even mainly, an individual matter but takes place through social interaction is the basis of the *sociocultural constructivist* perspective of learning. In this view, understanding results from making sense of new experience with others rather than by working individually. In group situations the individual learner takes from a shared experience what is needed to help his or her understanding (internalises) and then communicates the result as an input into the group discussion (externalises). There is a constant to-ing and fro-ing from individual to group as knowledge is constructed communally through social interaction and dialogue. Physical resources and language also have important roles to play (James 2012). Since language, which is central to our capacity to think, is developed in relationships between people, social relationships are necessary for, and precede, learning (Vygotsky 1978).

From this perspective, learning is conceptualised as a social and collaborative activity in which people develop their thinking together. In classrooms the interaction between students is mostly face-to-face, but learning from and with others can also be through the written word. Feedback to students in writing can be an effective channel for dialogue between teacher and students, providing that the comments take students' learning forward and the students have time to read and reflect on the comments and perhaps make amendments or additions to their work in response to these comments. Thus we can promote progress towards learning goals by teachers in a range of ways, such as encouraging collaboration and group work, providing clear goals, giving feedback, enhancing dialogue and negotiating. Approaches such as these contribute to the process of formative assessment.

The Role of Learning Progression

A clearer definition of formative assessment and description of what it involves in practice is needed to establish sound formative practice. It was a common finding among the researchers involved in ASSIST-ME that teachers in general have a vague and implicit understanding and use of formative assessment. Many teachers didn't distinguish between formative and summative use of assessment, and they often used quite weak feedback practices. In particular, working with learning progression steps as a central prerequisite for feedback processes was quite new for many teachers.

A learning progression describes 'successively more sophisticated ways of reasoning within a content domain that follow one another as students learn' (Smith et al. 2006, p. 1, cited in Duncan and Hmelo-Silver 2009, p. 606). Basically, a learning progression is necessary to guide planning for feedback that seeks to identify and build towards next steps in learning. It is also crucial if students are to be involved in the formative process. Broad learning goals must be broken down into sub-goals with success criteria related to each sub-goal. Traditionally, teachers have not made these steps explicit. The steps were inside their heads, as they said when interviewed, as part of their professional knowledge. This works fine for teachers giving relatively informal feedback to students, but if students are to be involved, by self-assessment or peer assessment, the steps need to be explicit, and students need to know what they are. More structured feedback is also necessary if the formative assessment is to have a certain degree of reliability. Very often an increase in mastering a given area of knowledge or of a competence is described using words from a generic taxonomy like Blooms taxonomy or the SOLO taxonomy. This is, for example, the case in descriptions of most grading scales and curriculum statements. But when teachers in the project reflected on their students' learning path within a specific part of the discipline, based on their experience as teachers, they did not construct competence levels following a standard taxonomy across disciplines. Very often they built a sequence of building blocks, more like stepping stones spread over a muddy field where you have to stand on each stone in order to have covered the whole field. Accordingly, they found that they would not be able to reuse learning progressions from one discipline area to another but needed to design them independently for each new content. This is in accordance with the findings from a major science education project working with learning progressions (Alonzo and Gotwals 2012).

Teachers in the Local Working Groups (LWGs) found it time-consuming and hard work to make levels of attainment explicit to the students. They acknowledged the fact that it is valuable for teaching and for the students, to make explicit what they normally do without conscious thought, but it was simply too time-consuming. Also, some teachers experience a need to work with colleagues on formulating learning progressions, and there was not always time to do this.

As expressed by two Danish science teachers involved in ASSIST-ME:

> There is a big difference between working with learning progression and to be aware of learning progression. I do not work much with it but I am very conscious of it. I do think progression clearly into the way I structure my lessons and how I ask students, but I have never before this project made it clear for the students that we have this learning progression.
>
> I would normally never write progression steps to myself. We know them; they are deep inside of us. So, it should only be for the benefit of the students, to make the demands clear for them.

Most teachers in the project did not have clear criteria or did not make them explicit for the students, and they did not see (and did not use) formative assessment as a specific procedure with specific quality characteristics. This has serious implications for both the validity and especially the reliability of their judgements and hence the potential summative use of the formative processes.

Summative Assessment

The aim of summative assessment is generally to report on students' level of learning at a particular time, rather than to impact on ongoing learning, as in the case of formative assessment. In some cases an assessment can, to an extent, serve both purposes, as we discuss later, but first it is helpful to distinguish the separate characteristics of summative assessment.

Assessment for summative purposes involves collecting, interpreting and reporting evidence of learning. Interpretation of evidence is in relation to the goals that students are intended to have achieved at a certain point, such as the end of a year, semester or stage. These are goals that can be described as medium-term, as distinct from the short-term goals of particular lessons or topics and from long-term goals such as ‘big’ ideas which are achieved over the whole period of school education.

There are several different ways of collecting evidence for summative assessment: by administering tests or examinations, summarising observations and records kept during the time over which learning is being reported, creating a portfolio of work, embedding special tasks in regular activities, engaging in computer-based tasks or some combination of these. When choosing how to collect evidence, consideration has to be given to the uses to be made of the data. Uses vary from routine reporting of the achievement of individual students to parents, to other teachers and to students themselves, to keeping records of the performance of groups of students by age, gender, background, etc. Some of these uses have ‘high stakes’ for teachers, schools and students themselves, in that they are used for making evaluative judgements which affect student selection, students’ own decisions, school reputation and placement on league tables and even funding. Particular uses have implications for the degree of validity and reliability of the data used. It is important, therefore, to understand the concepts of validity and reliability and their interactions with each other.

Validity and Reliability

All assessment involves the generation, interpretation and communication of data (Harlen 2013). The same processes are involved whether the purpose is primarily formative or summative, and the main purpose of these processes is to provide inferences about the knowledge, skills and competencies that students possess. It is the way in which these processes are carried out, and the way that inferences are drawn that determine the quality of an assessment. Assessment quality is generally described in terms of two concepts – validity and reliability. Superficially we could say that validity has to do with the ‘what’ and ‘how’ of assessment, while reliability has to do with ‘how well’ (Johnson 2012). But there is far more to these concepts than this, as we now see.

Validity

It is usual to define validity of an assessment in terms of how well what is assessed corresponds with the behaviour or learning outcomes that it is intended should be assessed, i.e. about which evidence is required. Determining the extent to which this is the case is complex and involves judgements of various kinds. Different types of validity have been proposed depending on the kind of information used in judging the validity. For instance, *content validity* refers to how adequately the assessment covers the subject domain being taught and is usually based on the judgement of experts in the subject. However, content coverage is not enough to describe the full range of a test or other assessment tool. The important requirement is that the assessment samples all those aspects – but only those aspects – of students' achievement relevant to the particular purpose of the assessment. Including irrelevant aspects (construct irrelevance) is as much a threat to validity as omitting relevant aspects (construct under-representation) (Stobart 2008). An example of construct irrelevance could be that inquiry-based science education problems are often broadly described in real-life settings – demanding good reading skills. So, these problems assess reading competence as well as science competence. This gives a biased result – unless, that is, the construct assessed includes reading of, for instance, social scientific issues as part of the learning demands. This is why *construct validity* is an increasingly prevalent validity concept, subsuming many of the other validity measures. Construct validity is based on a description of the skills and competences to be assessed and their relations (a so-called framework or construct). It could be a theoretical understanding or a model of a competence. A test will then have construct validity if it is able to assess the underlying construct of competence.

Another form of validity, *consequential validity*, is not a property of the assessment instrument or procedure itself but is a judgement of how appropriate the assessment results are for the uses to which they are put. The validity of an assessment tool is reduced if inferences drawn on the basis of the results are not justified. For example, a test of arithmetic may be perfectly valid as a test of arithmetic but not valid if used to make judgements about mathematical ability more generally. On a more general level, a test may say something about student's knowledge within a specific area of knowledge but is not a valid basis for predicting the student's study success in further education.

So, validity is not a property of an assessment method or instrument regardless of the circumstances in which it is used. This situation is formally expressed in the definition by Messick (1989, p. 13) of validity as 'an integrative evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment'.

Reliability

The reliability of an assessment refers to the extent to which the results can be said to be of acceptable consistency or accuracy for a particular use. There are many factors that can reduce the reliability of an assessment. For example, reliability is reduced if the outcomes are dependent on who conducts the assessment (which teacher or oral examiner), on who rates the students' assessment performances (which scorer or observer or external examiner at oral examinations) or on the particular questions used in a written test when these can only test a sample of all the different topics and levels of learning included in the curriculum being tested. The particular occasion on which the assessment takes place, and the circumstances under which it is undertaken, can also affect the assessment outcome and contribute to reduced reliability. Thus reliability is often defined as, and measured by, the extent to which the assessment, if repeated, would give the same result. For summative assessments which are conducted on a national scale and which have important consequences for all involved, extensive and sophisticated psychometric analyses have been used to explore their reliability (Baird and Black 2013). The most common used is Cronbach's α, measuring internal consistency, and Cohen's κ, measuring degree of agreement between two observers.

Reliability is important in relation to both formative and summative assessment. In the case of formative assessment, reliability is often lower, and it matters less than for summative assessment. This is because the notion of making a repeatable judgement and treating all students in the same way is not equally relevant when the purpose is to support well-founded decisions about next steps for individual students who may be at different stages in their learning and require different kinds of feedback. But even if the judgement of specific evidence should be student sensitive, the interpretation of student evidence must be based on the same criteria, using the same indicators. This is important because the judgements that constitute the field at hand form the landscape of meanings, definitions and importance in which the students should find themselves (Dahler-Larsen 2014). If formative assessment is systematically flawed, that is a problem because it could give students a wrong image of the field they have to learn and the demands they have to live up to. The reason why we sometimes live with less reliable formative assessment is that it is local and not systematically biased in a large scale. But to have the potential to be used for summative purposes, also the formative judgements need to be reliable.

Validity and Reliability Interactions

It follows from the definition of reliability that tests comprising questions where students choose between fixed alternative answers (multiple choice), that can be machine marked, are more reliable than ones that require answers to be created and that require some judgement in the scoring, such as open response format tests. However, the latter may be a more valid assessment if the purpose is to find out what answers students can produce rather than asking them to choose from specified alternative answers. It has also been shown that if students are asked, in a multiple-choice test, both to choose the best response and to explain their choice, many who choose the correct response are unable to give a valid reason for that response (Osborne et al. 2016).

Although it is desirable for summative assessment to be both highly reliable and valid, in practice there is a limit to optimising both. Normally, there is a conflict between high reliability and high validity, so that it is necessary in practice to adopt a compromise between them. An assessment which is highly reliable (like a simple multiple-choice test) is often low in validity. This is especially true for assessment of competencies and more advanced skills. Assessing complex demands validly is complicated and time-consuming, meaning that it is difficult to do in a reliable way. However, if an assessment is low in validity, it can have damaging effects, because those using the results will make incorrect inferences about the learner's capability. So, for example, students may infer, from a strong result, that they will succeed in advanced study of a subject but may discover when committed to that subject that its demands are radically different from those which they had been able to meet in the invalid summative assessments.

One aim of ASSIST-ME was to investigate these relations between formative and summative use of assessment in order to produce guidelines for changing the summative use of assessment so that it would be in alignment with the formative use. A key issue is that if the same evidence used formatively is also going to be used summatively, it needs to be (re)interpreted reliably. This means that the assessment should fulfil some psychometric measures for reliability often linked to summative assessment, like the Cronbach α for internal consistency. Here, the distinction between *evidence* and *judgement*, mentioned earlier, becomes important. For formative assessment, the evidence is interpreted both in relation to the goals (criterion-referenced) and in relation to the progress of a student towards the goals of a particular lesson or sequence of lessons, the next steps relating to where a student has reached and knowledge of the student's capabilities (student-referenced). For summative assessment, the evidence collected over a period of time needs to be judged only in relation to more general, or medium-term, goals (criterion-referenced) that apply over that period of time and to all students. We will later take up further what this means in practice, but it is relevant to note here that this dual use of evidence involves teachers in setting up longer-term goals and working towards them through breaking down the longer-term goals into sub-goals for lessons or sequences of lessons with clear goals and criteria for each sequence. This turned out to be one of the key challenges for the teachers participating in the ASSIST-ME project. As previously identified, working with implementing formative uses of assessment very quickly turned out to be a question of making learning progressions explicit within the actual subject matter and steering the teaching/learning processes in the light of these progressions. The ASSIST-ME teachers were struggling, together with the researchers, in formulating learning goals. The teachers were provided with a template with space for sub-goals and a three-step progression for each sub-goal and a list of verbs from the SOLO taxonomy and Bloom's taxonomy. But reality turned for most teachers out to be more complicated, as two Danish science teachers explained:

> It is difficult to see which progression steps fit the students' actual performance – at the end of the day it will always be an interpretation with a lot of uncertainty. You can 'explain' or 'discuss' on many levels.

> The same generic taxonomy will be different from subject to subject – the one formulated in biology may not be usable in mathematics. 'Explain' can be on a very high level in mathematics while it might be low in other subjects.

The generic taxonomies did not fit the classroom. The categories were too abstract and too far from subject practice. As mentioned earlier, the teachers found that their teaching (and students' learning) did not follow a path described via a generic taxonomy, for instance, like climbing a ladder step by step beginning at the first step. Rather, it followed a subject-specific logic that has evolved through the teachers' professional life. It could be described as crossing a river using stepping stones spread across the river bed or putting a jigsaw together. This made it quite time-consuming to establish the progression steps needed for the formative feedback processes – they had to be constructed for every subject sequence.

More problematic than the time it took, was the influence it had on the teaching, the whole structure of the lessons.

> (Progression) templates to use in the daily teaching have to be very precise and you need one for almost every lesson or sequence, so, you risk to lose the overview. You focus on individual issues and lose the whole picture. … If you give this very comprehensive material to students they will give up. And it has to be very concrete – students can't use very general concepts.

> The more systematically you do it, the less freedom you have.

> I find it unfortunate if you teach with the progression steps totally described in any detail. You lose something – it tends to be about performance instead of learning. The more detailed, the more focus on: Now the two and the five competences are achieved. If we constantly focus on how they can assess themselves, then we remove the curiosity and the joy and focus on what you have to perform.

> From a basic human point of view something inside me is against this visible learning, where you represent people as rational thinking beings. I don't think this is the whole story. I more think that it is a question of curiosity and something quite impalpable. We lose a lot by doing this – and what about the Bildung – where is that in the progression plan?

Some teachers, though, were more positive and pointed at the fact that the progression steps describe something we want students to learn. They are often important competencies and through visible learning help students steer their own learning processes. But it could be a high price to pay to obtain this. Students risk to be pushed towards a more performance-oriented regime at the cost of their mastery of the subject (Midgley et al. 2001). It turned out to be a very delicate balance.

Issues in Using Tests

Tests are often the method of choice for gathering information for summative assessment on the grounds of 'fairness,' since they appear to treat all students in the same way. However, giving all students the same task is not the same as giving them equal opportunities to show what they understand or can do. This becomes clear when looking at test items. The subject matter used to pose a question, the language used, the amount of reading and writing involved and familiarity with the test situation are among the many factors that will advantage some students and disadvantage others. Research clearly shows how students' performance is influenced by the situation or context in which the task is set – students perform differently in different situations and especially low performers benefit from rich, authentic assessment situations (Dolin and Krogh 2010). Other well-known problems with tests stem from the necessarily limited number of items that can be included, meaning that, as pointed out above, a selection of contexts and problems has to be made and that a different selection is likely to lead to different results.

Further, when test results are used for high-stakes judgements, of students, teachers or schools, the tendency to 'teach to the test' in striving for high scores has a narrowing impact on the curriculum content and on pedagogy. Moreover, the validity of the assessment is threatened when there is a stark contrast between the mode of learning in inquiry-based activities and the mode of testing. Students, who are used to working in groups, sharing ideas and reviewing their own work in relation to comments and reactions of others, may find it difficult to express their understanding and their acquired competencies in a conventional test or examination setting.

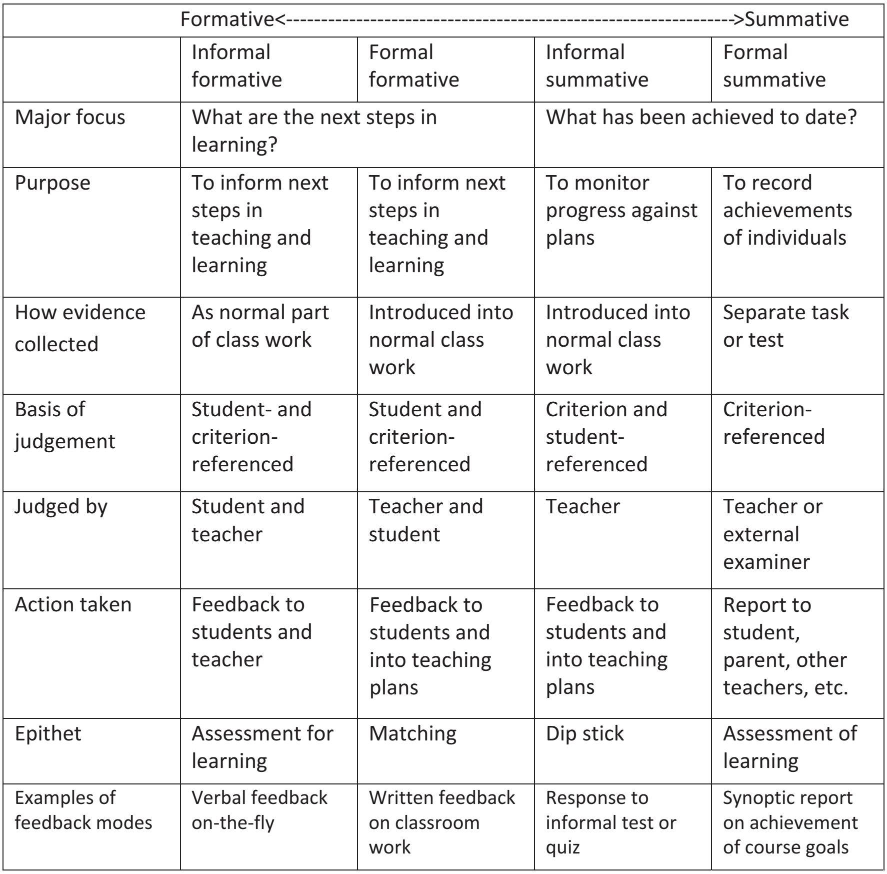
Many of the most severe problems of tests arise from the use made of the results. This is an important factor in determining the degree of stress that students may feel and indeed may influence the perception of the amount of testing they experience. For example, when test results are used for high stakes, where decisions affect students' opportunities (as in the end of primary tests in some countries with highly selective secondary education and end of high school examination in others), students, and their parents, know that the results matter and become stressed by the process. Teachers might add to this stress by giving repeated practice tests. However, stress can also become an issue when the tests matter a great deal for the school although not for the individual students, as in the case of the national tests in England at the end of primary school (which are entirely about accountability). Insofar as teachers are stressed about the results, they may transfer this stress to students, spending considerable time directly preparing students by creating more tests for practice. This adds to the experience of students of school work as being dominated by testing. It can be argued that it is the use rather than the nature of testing that is at fault. However, until the policy that leads to this use can be changed, it is unfortunately difficult for the potential of tests to provide useful guidance to teacher and students to be realised.

The overarching criterion, by which any combination of assessment methods is to be judged, is the degree to which they both support and reflect student learning. Such learning has different dimensions, ranging from learning through interactive dialogue to learning to tackle realistic complex tasks which require the integration of different elements of one's learning (Black 2016, p.733f).

Formative and Summative Assessment: Dimension or Dichotomy?

As the examples from the project in Chaps. 4, 5, 6 and 7 show, there is considerable diversity in the ways in which formative and summative assessment is practised. For instance, feedback can be offered to students as very informal and unstructured comments, 'on-the-fly', or more formal written comments on a structured piece of work or classroom test. Some of the features of these events, such as the use of tests, may seem – at least superficially, until their use is examined more carefully – closer to a summative than a formative purpose. This suggests that the relationship between formative and summative assessment might be better described as a dimension rather than a dichotomy as suggested in Fig. 3.2 (based on Harlen 2012).

At the extremes are the practices and uses that most clearly distinguish between assessment for learning and assessment of learning. At the purely formative end,



	Formative<-->Summative			
	Informal formative	Formal formative	Informal summative	Formal summative
Major focus	What are the next steps in learning?		What has been achieved to date?	
Purpose	To inform next steps in teaching and learning	To inform next steps in teaching and learning	To monitor progress against plans	To record achievements of individuals
How evidence collected	As normal part of class work	Introduced into normal class work	Introduced into normal class work	Separate task or test
Basis of judgement	Student- and criterion-referenced	Student and criterion-referenced	Criterion and student-referenced	Criterion-referenced
Judged by	Student and teacher	Teacher and student	Teacher	Teacher or external examiner
Action taken	Feedback to students and teacher	Feedback to students and into teaching plans	Feedback to students and into teaching plans	Report to student, parent, other teachers, etc.
Epithet	Assessment for learning	Matching	Dip stick	Assessment of learning
Examples of feedback modes	Verbal feedback on-the-fly	Written feedback on classroom work	Response to informal test or quiz	Synoptic report on achievement of course goals

Fig. 3.2 Dimensions of assessment purposes and practices

assessment is integral to student-teacher interaction and is also part of the student's role. The teacher and student consider work in relation to the activity goals that are appropriate for the particular learner, and so the judgements are essentially student-referenced. The central purpose is to enable teacher and students to identify next steps in learning and to know how to take these. At the purely summative end of the dimension, the purpose is to give an account of what has been achieved at certain points. For this purpose, assessment should result in a dependable report on the achievements of each individual student. Although self-assessment may be part of the process, the ultimate responsibility for giving a fair account of how each student's learning compares with the intended learning goals should rest with the teacher, whether or not the evidence is obtained by external tests or by judgements of regular work.

Between these extremes, it is possible to identify a range of procedures having various roles in teaching and learning. Indeed, there is in practice considerable overlap across the vertical boundaries inside the above table – they are not rigid barriers. For instance, many teachers would begin a new topic by finding out what the students already know, the purpose being to inform the teaching plans and maybe to identify the point of development of each individual. Similarly, at the end of a section of work, teachers often give an informal test to assess whether new ideas have been grasped or need consolidation.

Within both the more formative and more summative parts of the dimension, there are considerable variations. Classroom observations of teachers' formative assessment practices by Cowie and Bell (1999) were interpreted as indicating two forms of formative assessment: 'planned' and 'interactive'. Planned formative assessment concerns the whole class, and the teacher's purpose is to find out how far the learning has progressed in relation to the lesson goals. Information gathering, perhaps by giving a brief class test or special task, is planned and prepared ahead; the findings are fed back into teaching. This is similar, in many ways, to 'informal summative'. The Structured Assessment Dialogue described in Chap. 5 is an example of such a tightly structured assessment method that can be used for both formative and summative purposes without further adaptation. 'Interactive' formative assessment is not planned ahead in this way; it arises from the learning activity (on-the-fly). Its function is to help the learning of individuals or groups; feedback is both to the teacher and the learners and is immediate. It has the attributes of 'informal formative'. In practice, the two forms are difficult to distinguish. A teacher might ask an open question and encourage interactive dialogue across the whole class – the interaction is planned, in part, because the teacher might believe that engaging in dialogue is itself a way to develop students' capacity and confidence as learners – a belief that is supported by the work of Alexander (2008). Further examples are given in Webb and Jones (2009).

At the more summative end of the dimension, there are also different degrees of formality, from the examination taken under strictly controlled conditions to the more informal exercises set and judged by the teacher. What is described as 'informal summative' in Fig. 3.2 may involve similar practice to 'formal formative'. However, the essential difference is the use made of the information. Here, it is useful to consider the two kinds of feedback referred to earlier: feedback from teacher to students (affecting their learning) and feedback from students to teacher (affecting their teaching). In those cases, where summative assessment takes place at the end of a course or unit or the end of a semester or year, it may not be possible for the assessment result to be used to help these particular students improve their work. But the teachers may still use the result to adjust their approach when teaching similar activities to other students in the future (i.e. formative for teachers rather than formative for learners). So, if the results are used to adapt teaching or improving the student's learning, then it is 'formal formative'. If there is no feedback into teaching or learning, then it falls into the category of 'informal summative', even though the evidence may be the same classroom test. Ideally, of course, summative tests should be given, while there is still time for students to learn from their mistakes. When this is the case, any assessment can have a formative function, in accordance with the principle that 'Assessment of any kind should ultimately improve learning' (Harlen 2010, p. 31).

However, the process sequence represented in Fig. 3.1 should not be understood as implying that evidence to be used for summative judgements is only collected and recorded at the end of a teaching episode. Teachers may collect summative assessments en route and so build up for each student a collection of data, including pieces of students' work produced on several occasions, which will form the basis, perhaps with a formal test, for the final summative assessments. This will be expanded later in the chapter.

The Effect of Country Differences in Existing Assessment Practice

It is important to acknowledge differences among countries in respect of how assessment for formative and summative purposes is practised. Some of these differences relate to a country's traditions and existing practices. For example, whether students are assessed at the end of secondary school by external or school-based examinations has implications for classroom practice and the extent to which teachers see assessment as serving formative and summative purposes. Other differences are evident in how criteria to assess students on knowledge, skills and competencies are made explicit in the official curriculum documents and how teachers use these criteria. Black (2015) analyses eight different country practices of formative assessment and the difficulties in implementing them, often due to domination of summative assessment. As for linking formative and summative use of assessment, these variations seem relevant:

- Recognition of the need to share goals and assessment criteria with students, which may require teachers to reformulate the goals and criteria so that they can be understood by students, is a procedure that is not common practice in all countries.

- Clarity about the formative or summative status of any assessment event. In some countries it is the expectation of the students, and their parents, to be told when the student's product is to be used for summative assessment. This information affects the way students respond to a task; if a more formative purpose is intended, errors are seen as something that can help learning, while if the task is seen as for summative assessment, errors are seen as indicating that a learning goal has not been met. To put this another way, students may perform in one way if they understand that their products will be used (and judged) in order to improve their learning but may perform in a quite different way if they know the product will be judged for summative purposes. This links to the so-called didactical contract. If students know that the report has a formative purpose, they will be more experimental, more open and less tactical and will strive less for perfection, for flawless work.
- The balance between assessment for accountability purposes and assessment for learning or the dominance of a testing regime on everyday practice. National tests and the stakes of the examinations have a strong influence on teachers' teaching and students' behaviour. A survey among Danish upper secondary students revealed that 55% of the tested students have stress symptoms – due to a daily fight for high grades (Nielsen and Lagermann 2017). Their level of stress is similar to that of the 20% most stressed people in the Danish labour market (ibid.). This gives a general context for formative assessment. No matter how much teachers emphasise the formative purpose of an assessment, the summative element does not disappear. Students do seldom believe that any assessment only has a formative purpose, given today's reality in contemporary educational systems.

Approaches to Linking Formative and Summative Assessment

Given the shortcomings of tests outlined earlier, and given the wash-back effect of summative assessment on the teaching and learning, it is worthwhile looking for ways to diminish the negative effects on learning from summative use of assessment and to improve the formative use of assessment. In particular, we need ways that are capable of gathering data about the full range of competencies and understandings that are the aims of inquiry-based science, technology and mathematics education. It is, fundamentally, a question of optimising the alignment between the learning goals, the pedagogy and the assessment method (used for both formative and summative purposes). Here, linking the formative and the summative uses of assessment is a key issue. How to do this best will depend on the national and local context, but basically there are two different approaches to linking formative and summative purposes of assessment:

- *Connecting* the formative and summative use of assessment evidence. Here any assessment is used for either a formative or a summative purpose, but the methods are very alike giving relatively high validity to the summative use and relatively high reliability to the formative use.

- *Combining* the formative and summative use of assessment evidence, i.e. using evidence from an assessment for both formative and summative purposes. Using evidence from a summative assessment for formative purposes is relatively unproblematic and well-known. More innovative and full of perspective is the use of formative assessment processes for summative purposes.

Connecting the Formative and Summative Use of Assessment

The formative and summative use of assessment can be connected through:

(a) Designing a summative assessment method, an examination or a test, that is in alignment with the formative processes in everyday class. The assessment method must reflect the central goals and be able to assess the competencies in a valid way. Such an examination will replicate everyday teaching. During normal teaching students have become familiar with the criteria for fulfilling the learning goals via various formative processes. At the examination students are put in a semi-authentic situation and given a fixed time span (4 h, 8 h, 24 h, 48 h - depending on the subject) to go through (most of) the inquiry (including experimental) framework. The work is monitored and assessed (using a template with criteria known from everyday practice) and may be supplemented with an individual oral examination (both with an external examiner). Many Danish science and technology examinations have been designed and introduced following these principles, and they have the same reliability properties and the same cost level as ordinary oral examinations. Even advanced generic competences like innovation can be validly assessed for summative purpose using such authentic approaches (Nielsen 2015). Implementing an examination format in alignment with the learning goals will turn the often inexpedient wash-back effect from examination into a more productive effect.

(b) Using similar or the same assessment methods separately for both formative and summative purposes distributed throughout the year and very much performed close in time. Each assessment event is clearly identified as being used either for formative or for summative purposes. The summative assessment events are aggregated together for accountability purposes. The formative assessments are both similar to those for summative purposes but also supplemented with more valid methods able to measure more advanced competences. This is consistent with the didactical contract. Students can be confident that they will be clear about the assessment purpose – whether what they produce and present will only be formatively assessed or whether they should prepare for the summative assessments. This approach acknowledges the present accountability regime and the difficulties in changing the examinations. But it makes a clear distinction between the two purposes of assessment and thus minimises the distortion of the summative assessment.

There is considerable future potential to enhance assessment by integrating IT-based elements. The TELS (Technology-Enhanced Learning in Science) project is an early example of an IT-based system designed to improve learning and deliver assessments for summative purposes. The project has designed a Web-based Inquiry Science Environment (WISE) as an online platform for designing, developing and implementing science inquiry activities (Linn and Chiu 2011). Such new third-generation assessment systems are emerging, trying to serve both institutional (summative) purposes and individual learning (formative) purposes. They typically build a system of complex computer simulations and other performance tasks into a blended learning system. As an example, Bennett (2015) describes how Bennett and Gitomer (2009) designed an integrated system of assessments built into a demonstration programme, CBAL (Cognitively Based Assessment of, for and as Learning). The system would consist of distinct summative and formative assessment components each directed at achieving different purposes.

As Bennett (2015) writes, '… a key characteristic of third-generation assessments will be the use of complex simulations and other interactive performance tasks that replicate important features of real environments, allow more natural interaction with computers, and assess new skills in more sophisticated ways' (p. 387). Students can work with such continuously embedded assessment systems during the year, and frequent sampling of evidence of learning can then be used formatively and automatically scored and aggregated for accountability purposes. This needs to be organised in a way that does not violate students' premises for their work and their learning engagement, e.g. with a clear distinction between formative and summative use of the data and with other ethical considerations, as described by Timmis et al. (2016). The teacher's control over the system and the system's flexibility and its potential for adaptation to the teacher's own teaching approach and to the needs of the class are other important aspects of new integrated assessment systems. To some extent, there is a danger that these systems may lead to relatively rigid teaching practices. A more flexible platform was developed within the ASSIST-ME project. It was designed to support the development of students' experimental competence by providing guidance and feedback during a learning cycle. The platform is described in Chap. 9 in this book.

A tight alignment between everyday practice and the test or examination is only preferable if the test is in valid accordance with the learning goals or if the two purposes of the test are clearly separated. If this is the situation, they will minimise some of the more pernicious wash-back effects from tests and examinations. But summative assessment situations will always have some unfortunate side effects. The unusual and high-stakes-related situations will make many students nervous to a degree that affects their performance, and many of the resources invested in tests could be better used for learning activities. This is why there is a huge potential in trying to combine the formative and the summative uses of assessment evidence via extracting data for summative use from formative processes.

Combining the Formative and Summative Use of Assessment

Good assessment requires valid evidence, relevant criteria (goals) and a reliable method of judging the extent to which criteria have been met. If students are learning through inquiry, then the evidence of their learning exists in what they do in inquiry-based activities. When teachers are using assessment formatively, they are using this evidence to support learning, as in the formative assessment cycle. The activities that enable students to develop understanding and competencies are at the same time opportunities for collecting data about progress in their learning. This evidence can be accumulated from lessons on a topic and brought together at the time when achievement is to be reported and recorded.

A suggestion for how this might be done is represented in the complete Fig. 3.1, where the black arrow leading to judgement for summative use originates from the evidence gathered in the formative assessment cycle. There are several key points to make about this model.

First, formative judgements of what action, if any, to take are made in relation to the lesson or activity goals. For summative assessment judgements, however, the goals are the medium-term goals that are the targets for achievement over a period of time such as a semester or year. That is, evidence selectively accumulated over several activities is judged for summative reporting against medium-term goals stated in a national curriculum document or in a school's curriculum plan for a particular year or stage. For example, lesson or activity goals in relation to a topic on sound might be to recognise 'how sounds can be made in different ways and can be changed in loudness and pitch'. The medium-term goal, achieved after several lessons with different activity goals, might be to understand 'that sound is caused by vibration in a material'.

A second point is to emphasise that it is the evidence gathered for formative assessment that is used as evidence for summative assessment, not the formative judgements about what action to take. This is because, as noted earlier, formative assessment judgements are often not strictly criterion-referenced, and formative judgements take into account the circumstances that affect individual or groups of students. It is entirely appropriate for this to be the case in formative assessment since it means that students have the tailored help they need to take them from where they are towards the learning goals. But evidence used to report on learning for formal summative purposes has to be judged in the same way for all students, not influenced by student-related considerations.

Third, the evidence used is the 'latest and best' selected from what has been gathered, by the methods discussed earlier, over the time for which achievement is summarised and reported. The use of 'best' evidence is in recognition that over the course of a semester or year of work – at the end of which a summative report is needed – students make progress, and evidence from earlier work will be superseded by later work. 'Latest' recognises that some evidence will have been collected early in the year when particular topics were the focus of activities and may not have been revisited. There is no need to retain every piece of evidence produced during that time, which would in any case be too burdensome. Rather, the aim is to create a collection (which could be in various forms, such as a portfolio, folder or computer file) of what is considered to be the *best* evidence of achievement. As work proceeds during the year, better work may replace earlier attempts. Involving students in making this collection helps their understanding of the quality of work that is expected and adds to the formative value of the assessment process. It also means that students understand how their work is judged for formative and summative purposes.

Fourth, formative evidence that can be used in this way has to be in a tangible form. Given the considerable diversity in the ways in which formative assessment is practised, a question arises as to whether all kinds of evidence are suitable for use in summative assessment. In the very informal methods, evidence is picked up by teachers by observation as they interact with individuals or groups of students; judgement is almost immediate, and decisions is taken quickly about any action needed. The evidence in this case is ephemeral, with no record being made, but nevertheless it provides information that is used in promoting further learning. In more formal methods, the teacher will have planned to obtain evidence, by asking students either to give written answers to questions designed to probe their knowledge and ideas or, as they conduct an inquiry, to explain their plans at intermediate stages in the work – thereby composing their own diaries of an inquiry. In between these different ways of collecting evidence are other means through which students record their work and ideas. If evidence from formative assessment is to be used for summative purposes, it follows that some, at least, of this has to be in a tangible form rather than an ephemeral such as picked up informally. This argues for using a range of ways of gathering evidence for formative assessment.

Fifth, the process of summative judgement should include measures to ensure reliability appropriate to the use of the summative assessment. At the time when a summative report is required, the best evidence is brought together and summarised by scanning for evidence of meeting the criteria indicated in the medium-term goals. The judgement can be reported in various ways. For some uses at some stages, it is only necessary to report whether goals have or have not been met. In other cases, the report may distinguish different degrees of meeting the goals, in which case there is a need to establish and exemplify the criteria for partial achievement.

Whether the judgement of accumulated best evidence is made by the teacher or by someone else depends on the intended use of the summative data. The two main purposes (leaving aside national and international monitoring for which only a sample of students are assessed) are for reporting on individual students' achievement (low stakes for teacher but high stakes for students) and recording the progress of groups of students as part of school evaluation and accountability procedures (high stakes for teachers and schools).

It is usually students' own teachers who make the judgements for regular reporting to parents and information for school records. For these audiences, scores, levels or grades give little information about what has actually been achieved. Rather, information in summary form about the extent to which goals have been achieved is best accompanied by a narrative, identifying what has been learned well, what needs more attention and other information that will help future learning. Face-to-face interviews with parents and students can use examples of work that show what is needed to achieve certain goals.

Where the result is to be used for making decisions affecting a student's future, as in the case of some external examinations or for use in high-stakes accountability measures having another teacher or external assessor involved in the judgements, or using some moderation procedure may be considered necessary for confidence in the result. However, only if we had assessments that captured the full meaning of the learning goals, without construct over- or under-representation (see previous section on validity and reliability), and did so with optimum reliability, would it be acceptable to use students' achievement as the sole measure of school effectiveness or the only basis for accountability of schools and teachers.

Application to Inquiry-Based Education in Science

Inquiry-based science education aims to help students to develop an understanding of key ideas about science and of science (i.e. the nature of science) and ability to conduct science inquiries. Besides helping to develop these science-specific competencies, the assessment processes serve to build students' more generic competencies, such as innovation and creativity and other life and learning skills mentioned at the beginning of this chapter. The model of summative assessment based on evidence from classroom activities requires the identification of clear goals for each activity, contributing to medium-term goals for particular stages, which in turn contribute to overall long-term goals.

During the activity the teacher and students work towards the specific activity goals, gathering and using evidence from the ongoing work in cycles of formative assessment. With the medium-term goals in mind, the teacher also helps students make connections between the ideas emerging from different activities. In turn, these medium-term goals are part of the story of progress towards the overall aims of developing 'big ideas' and inquiry skill competencies. The progress towards the big ideas of science education has been mapped out in the publication *Working with Big Ideas of Science Education* (Harlen 2015).

The model of gathering data from the range of classroom activities is particularly appropriate for inquiry skills or transversal competencies such as 'investigation' or 'argumentation'. The reason for this is that the ability to apply such skills or competencies is highly dependent on the nature of the subject matter and content in which they are used, and so evidence from a range of different inquiries is needed for reliable assessment. In particular, for some inquiry topics, a clear understanding of the concepts involved may be essential, whereas for others such understanding is not needed – and indeed work to test explanations for an observed effect may help develop understanding of the relevant concept.

However, there is less clarity about progression in developing competencies relating to conducting inquiries than there is in the case of the development of scientific conceptual understanding. Thus for the model to be applied, there is a need for further work to identify criteria to be used in judging the accumulated evidence relating to development of inquiry skill (described as practices in the US framework (NRC 2012)).

There is obviously a problem in that national, high-stakes, assessments which are restricted to the use of externally set tests, to be completed in writing in a strictly limited time, cannot give results which are a valid measure of inquiry competences. This problem could only be overcome if a proportion of the high-stakes results were to be based on teachers' own assessments using a variety of assessment occasions, with evidence collected over time in the form of student portfolios. This approach has been used in a few state systems, notably in Australia (Wyatt-Smith et al. 2010). The findings of that work and those from a limited attempt to develop such work with a group of teachers in England (Black et al. 2011) show (a) that it can take one or more years for teachers to develop the necessary skills, (b) that interschool collaboration is essential in order to guarantee overall comparability of results and (c) that teachers involved have found the work rewarding as it has a positive effect on their whole approach to learning and assessment.

Challenges and Benefits

The combination between formative and summative assessments may take a large variety of forms and in most of the countries raises questions. This has been discussed in the ASSIST-ME Local Working Groups in each participating country.

Most of the teachers involved in the project consider it possible to use their summative assessments (not external final exam) formatively. Using formative assessment to summative purposes raises much more divergence. In two countries (England and Slovakia), there is a large agreement, for example, teachers 'demonstrated an ability to use their formative evidence from assessment conversations to make a summative judgment in relation to inquiry competences' (cf. meeting minutes). On the other side, the law of one country (Switzerland) prohibits 'the use of information collected for formative purposes in grades'. Most of the countries, even if this practice is possible or more recommended by the official instructions, acknowledged difficulties. The main one is the 'didactic contract' in the sense that the reciprocal expectations of the teacher and the students are different depending on the way the teacher uses the students' productions to make a judgement either to help them or to give a mark. Another difficulty concerns the coherence between the criteria involved in summative and formative assessments, and with what is taught. The need of coherence is largely recognised but also the difficulty to get it. Several reasons are proposed: some competences, in particular experimental, are difficult to assess summatively with written and pencil tests, or the criteria for formative assessment are adapted according to the students, whereas in summative case, they should be the same for all. To get coherence can necessitate a long back and forth analysis between the teaching goals in terms of knowledge and competences, how they are involved in the effective teaching and their use in the different assessments.

At last the question of time is raised in all countries. The teaching constrains in terms of content do not allow the teacher to take time enough for formative assessment and to relate it to summative ones in particular in the case of learning progressions which necessitate very regular assessment on different aspects of the teaching content. Thus, if all teachers are in favour of enhancing the combination between summative and formative assessments, different ways are possible according to the culture of the countries and the practices; nevertheless all ways imply a coherence of assessment criteria. All these questions lead all the participants to suggest teacher professional development on the combination of summative and formative assessments.

Without doubt, the easiest, technical, solution to linking formative and summative use of assessment is to connect them by designing summative assessment methods able to validly and reliably assess inquiry processes. Teaching to the test would then make sense, since the test mirrors the educational goals. Such examinations have been designed (Nielsen 2015) and have been proved to fulfil standard requirements for reliability.

However, fulfilling the important requirement of validity is difficult, and school assessments in many systems seem to have a tradition which would not be justified in any other context. As Black (2016) puts it:

> the common practice of basing such assessments on a student's performance produced over only a few hours in isolation, from memory, and responding to demands that they may not have seen before, to which an immediate response in writing is required, is strange. No business enterprise would think of assessing employees' progress in this way. It is hard to justify a decision by a school, which has known about many pieces of work by a student, produced in a range of contexts and on different occasions, to base key decisions on short terminal tests.

There are several challenges to be faced in combining formative and summative use of assessment by implementing summative assessment based on evidence from classroom activities that is used for formative purpose. Among these is the need to ensure that:

- There are clear goals for classroom activities that will lead to medium-term and long-term goals.
- Teachers are assisted in translating the goals into inquiry-based activities.
- Teachers understand the role of formative assessment in learning and use a range of methods to collect relevant evidence of learning.
- Students are involved in assessing their progress and are helped to see assessment as having a positive role in learning.
- All involved in providing, collecting and using assessment data recognise that any summative assessment is necessarily an approximation.
- Judgements of teaching and the effectiveness of schools are based on a wide range of relevant evidence and not only on measures of student achievement.

However, the benefits to be gained make it essential to face such challenges. Given the weighty evidence of the value for learning – and particularly for learning through inquiry – of using formative assessment, it is important to protect its practice from overbearing and outmoded summative assessment that dominates and dictates teaching and learning. Summative assessment that uses evidence from learning activities not only enables the full range of goals to be assessed but encourages, indeed requires, the practice of formative assessment.

References

Alexander, R. (2008). *Towards dialogic thinking: Rethinking classroom talk* (4th ed.). York: Dialogos.

Alonzo, A. C., & Gotwals, A. W. (Eds.). (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam: Sense Publishers.

ARG (Assessment Reform Group). (2002). *Assessment for learning: Ten principles.* http://www.hkeaa.edu.hk/DocLibrary/SBA/HKDSE/Eng_DVD/doc/Afl_principles.pdf Accessed 27 August 2016.

Baird, J.-A., & Black, P. (2013). Test theories, educational priorities and reliability of public examinations in the England. *Research Papers in Education, 28*(1), 5–21.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education, 18*(1), 5–26.

Bennett, R. E. (2015). The changing nature of educational assessment. *Review of Research in Education, 39*(1), 370–407.

Bennett, R. E., & Gitomer, D. H. (2009). Transforming K-12 assessment: Integrating accountability testing, formative assessment, and professional support. In C. Wyatt-Smith & J. Cumming (Eds.), *Educational assessment in the 21st century* (pp. 43–61). New York: Springer.

Black, P. (2015). Formative assessment – An optimistic but incomplete vision. *Assessment in Education: Principles, Policy and Practice., 22*(1), 161–177.

Black, P. (2016). The role of assessment in pedagogy – And why validity matters. Ch.45, pp. 725–739 in D. Wyse, L. Hayward & J. Pandya (eds.) *The Sage Handbook of curriculum, pedagogy and assessment.* London U.K.: Sage *see particularly pp. 733–734.*

Black, P., Harrison, C., Hodgen, J., Marshall, M., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education., 18*(4), 451–469.

Butler, R. (1987). Task involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest and performance. *Journal of Educational Psychology, 79*(4), 472–482.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology, 58*, 1–14.

Cowie, B., & Bell, B. (1999). A model of formative assessment in science education. *Assessment in Education, 6*(1), 101–116.

Dahler-Larsen, P. (2014). Constitutive effects of performance indicators: Getting beyond unintended consequences. *Public Management Review., 16*(7), 969–986.

Dolin, J., & Krogh, L. B. (2010). The relevance and consequences of Pisa science in a Danish context. *International Journal of Science and Mathematics Education, 8*, 565–592.

Duncan, R. G., & Hmelo-Silver, C. E. (2009). Editorial: Learning progressions: Aligning curriculum, instruction, and assessment. *Journal of Research in Science Teaching, 46*(6), 606–609.

Dweck, C. S. (2000). *Self-theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press.

European Commission. (2011). *Education and Training in a smart, sustainable and inclusive Europe COM* 902 final, Brussels. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=COM:2011:0902:FIN:EN:PDF.

Harlen, W. (2010). What is quality teacher assessment? In J. Gardner et al. (Eds.), *Developing teacher assessment* (pp. 29–52). London: Open University McGraw Hill.

Harlen, W. (2012). On the relationship between assessment for formative and summative purposes. In J. Gardner (Ed.), *Assessment and learning* (pp. 87–102). London: Sage.

Harlen, W. (2013). *Assessment and inquiry-based science education: Issues in policy and practice.* Trieste: IAP. Available for free download from http://www.interacademies.net/File.aspx?id=21245. Accessed on 22 September 16.

Harlen, W. (2015). *Working with big ideas of science education*. Trieste: IAP SEP. Available for free download in English from http://www.interacademies.net/File.aspx?id=26736 and in Spanish from http://interacademies.net/File.aspx?id=28260

Harlen, W., & Deakin Crick, R. (2003). Testing and motivation for learning. *Assessment in Education, 20*(2), 169–207.

James, M. (2012). Assessment in harmony with our understanding of learning: Problems and possibilities. In (ed.) J. Gardner. Assessment and learning, 2ndnd edn. London: Sage 187 – 205.

Johnson, S. (2012). *Assessing learning in the primary classroom*. London: Routledge.

Linn, M. C., & Chiu, J. L. (2011). Combining learning and assessment to improve science education. *Research and Practice in Assessment, 5*, 5–14.

McManus, S. (2008). *Attributes of effective formative assessment*. Washington, DC: Council for Chief State School Officers (CCSSO).

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement 3rd edition* (pp. 13–103). Washington, DC: America Council on Education.

Midgley, C., Kaplan, A., & Middleton, M. (2001). Performance-approach goals: Good for what, for whom, under what circumstances, and at what cost? *Journal of Educational Psychology, 93*(1), 77–86.

National Research Council (NRC). (2012). *A framework for K-12 science education* (p. 41). Washington, DC: The National Academies Press.

Nielsen, J. A. (2015). Assessment of innovation competency: A thematic analysis of upper secondary school teachers' talk. *The Journal of Educational Research, 108*, 318–330.

Nielsen, A. M. & Lagermann, L. C. (2017). *Stress i gymnasiet - Hvad der stresser gymnasielever og hvordan forebyggelse og behandling virker med 'Åben og Rolig for Unge*. DPU, Aarhus Universitet. (In Danish. Title in English: Stress in upper secondary – what stresses high school students and how to prevent and treat).

OECD. (2013). *Synergies for better learning: An international perspective on evaluation and assessment*. OECD Reviews of Evaluation and Assessment in Education, OECD Publishing: Paris. http://dx.doi.org/10.1787/9789264190658-en

Osborne, J. F., Henderson, J. B., MacPherson, A., Szu, E., Wild, A., & Yao, S.-y. (2016). The development and validation of a learning progression for argumentation in science. *Journal of Research in Science Teaching, 53*, 821–846.

Pearson. (2005). *Achieving student progress with scientifically based formative assessment: A white paper from Pearson*. Referenced in Bennett 2011.

Shepherd, L. A. (2008). Formative assessment: Caveat emptor. In C. A. Dwyer (Ed.), *The future of assessment: Shaping teaching and learning* (pp. 279–303). New York: Erlbaum.

Smith, C., Wiser, M., Anderson, C. W., & Krajcik, J. (2006). Implications for children's learning for assessment: A proposed learning progression for matter and the atomic molecular theory. *Measurement, 14*(1&2), 1–98.

Stobart, G. (2008). *Testing times. The uses and abuses of assessment*. London: Routledge.

Timmis, S., Broadfoot, P., Sutherland, R., & Oldfield, A. (2016). Rethinking assessment in a digital age: Opportunities, challenges and risks. *British Educational Research Journal, 42*, 454–476.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological process*. Cambridge, MA: Harvard University Process.

Webb, M., & Jones, J. (2009). Exploring tensions in developing assessment for learning. *Assessment in Education: Principles, Policy & Practice, 16*(2), 165–184.

Wiliam, D. (2011). *Embedded formative assessment*. Bloomington: Solution Tree Press.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: A study of standards in moderation. *Assessment in Education., 17*(1), 59–75.

Part II
Practice

Chapter 4
Assessment On-the-Fly: Promoting and Collecting Evidence of Learning Through Dialogue

Christine Harrison, Costas P. Constantinou, Catarina F. Correia, Michel Grangeat, Markus Hähkiöniemi, Michalis Livitzis, Pasi Nieminen, Nikos Papadouris, Elie Rached, Natasha Serret, Andrée Tiberghien, and Jouni Viiri

Introduction

Learning in STEM classrooms has undergone a radical change in the last two decades. Science and technology has made many advances, and information on many aspects of these subjects is readily available through a quick search on the Internet. Schools therefore need to move beyond the dissemination of information to students towards helping students generate the skills that enable them to use and apply STEM ideas in a range of contexts. This approach is central to inquiry learning, where students raise questions or notice problems and seek resolution and understanding. Such an approach replicates how students learn outside of school and so has greater authenticity than more traditional classroom learning as it encourages students to develop skills that will become useful in life and applicable across many school subjects. Inquiry learning models a range of approaches that scientists and technologists engage in as they work to examine problems, seek solutions and challenge perspectives (Duschl et al. 2007; Gillies et al. 2014).

C. Harrison (✉) • C.F. Correia
King's College London, London, UK
e-mail: christine.harrison@kcl.ac.uk

C.P. Constantinou • M. Livitzis • N. Papadouris
University of Cyprus, Latsia, Cyprus

M. Grangeat • E. Rached
Laboratoire des Sciences de l'Education, Université Grenoble Alpes, Grenoble, France

M. Hähkiöniemi • P. Nieminen • J. Viiri
University of Jyväskylä, Jyväskylä, Finland

N. Serret
Nottingham Trent University, Nottingham, UK

A. Tiberghien
l'ENS de Lyon, Lyon, France

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_4

The aim in inquiry classrooms is to move away from set routines that allow students simply to witness phenomena towards coherent explanations of the phenomena in context (Kawalkar and Vijapurker 2013). Laursen et al. (2011) found that inquiry learning in mathematics led to students having greater persistence, independence, enjoyment and confidence than students taught in a more traditional approach. While there is a broad range of approaches to inquiry in science classrooms, it is widely accepted that there are a number of teaching and learning elements that need to be present (Lederman et al. 2013; Minner et al. 2010) to distinguish it from traditional approaches (Grangeat 2016). For example, teachers need to use questions and tools to seek and recognise student thinking during the course of the inquiry activities and strategically scaffold students' thinking (Gillies and Nichols 2015) in order to capitalise on teachable moments that arise during instruction (Haug 2013). These points arise when students do or say something, during an activity, that alerts the teacher to some gap in student knowledge or skills. The teacher is then able to help the student by providing the information or demonstrating the skill or directing the student towards a resource.

Inquiry learning encourages classrooms where talk plays a key role in communicating and developing ideas. In inquiry classrooms, interactions between teacher and students, and within student groups, fashion and shape the meaning that learners form of ideas. In inquiry classrooms, conversations with students as they work on an inquiry task provide teachers with a rich bed of evidence of students' current thinking that is collected in real time as students engage with inquiry activities. This means that the teacher can respond to student ideas as they develop during the inquiry activity rather than provide feedback after the inquiry has been completed and documented in a report. While inquiry work provides teachers with an insight into student understanding, at the same time, the dialogue creates a rich breeding ground for student ideas to develop and evolve. Such an approach allows students more flexibility in the approach they take with an inquiry and also encourages them to consider at several points, during their inquiry, if their choices of method and apparatus are providing results that support or challenge their ideas. This creates a wealth of opportunities for teachers to survey and consider the assessment data they have collected throughout the activity and make judgements about this evidence. Teachers can then take formative action through making informed decisions regarding next steps in student learning. The learning evidence that can be collected through interactions on-the-fly depends on how these assessment conversations are set up, on what type of questions are used and how the teacher interprets and acts upon this evidence.

Assessment in the Inquiry Classroom

The roles of assessment for learning (AfL) and 'learning how to learn' (Hazelkorn et al. 2015; James et al. 2006) are increasingly being seen as means to facilitate learning and gather evidence of student learning as part of the teaching process. It is argued that through a formative approach to assessment, learners are able to take a more proactive role in the learning process. It also enables learners to develop valuable higher-order thinking, such as analysis and evaluation, and self-regulation skills, such as to set their own goal or persevere with a challenging task, as well as other prized twenty-first-century competences, which are often considered to be necessary in a rapidly changing world (Hazelkorn et al. 2015; OECD 2005; Rocard et al. 2007). An AfL approach to teaching and learning fits well with an inquiry-based approach where the teacher's role is more about mediating the learning rather than directing the students along a specific route. Within inquiry classroom activities, students are presented with an idea to explore either by raising questions themselves or seeking answers to questions that teachers raise. This means that students are given some choice about the approach they take, and teachers are able to collect information on student understanding from the way learners discuss ideas with one another, the questions they raise and how they seek answers to questions as well as the quality of the work they produce.

When teachers pay attention to assessment information collected during the course of learning, they are able to interpret and make decisions about such assessment data in a timely fashion that can drive future planning and support student learning, for example, through feedback (Black et al. 2003). Through classroom interactions during an inquiry activity, both teachers and students are able to get better information to assess where individual and groups of students are in their understanding and focus on what is needed to take their learning forwards. For teachers, these feedback loops provide evidence to structure and tailor guidance and lesson direction. For students, the focus on feedback helps them determine where they need to apply their effort, so that they work on skills or concepts that need development or consolidation. During inquiry activities, there is a two-way feedback process in that teachers notice what students do and say and also provide guidance for their students in response to this information. Through this reciprocal feedback process, students' understanding and performance can be developed and important assessment evidence gathered as part of the teaching and learning process.

In inquiry learning, students take an active role in decision-making (Crawford 2000). This might be about the question they are interested in exploring or in the methods they choose to answer a question or in the ways they decide to interpret the data collected. This more open approach to experimentation in the classroom makes it more difficult for the teacher to plan lessons in detail and in advance, and instead teachers need to develop skills and techniques that enable them to make prudent interventions, to support student learning during an inquiry. It is clearly important, in inquiry activities, that teachers find ways of tapping into student thinking, as these ideas are developing, and so teachers need to watch carefully how students explore inquiry ideas and develop techniques that probe student thinking concurrently. This requires a more dynamic approach to question and answer sessions in the classroom with teachers probing student thinking as ideas are emerging and developing.

Interactions on-the-fly are informal formative assessment conversations which are not planned beforehand but take place spontaneously when the teacher recognises appropriate opportunities to support students in advancing their learning (Shavelson et al. 2008). These conversations can take place between the teacher and the whole class, within a group of students during activities or with a single student in a group or class situation. All of these events set up feedback loops, the content and emphasis of which may be utilised by the individuals engaged in the interaction or by those listening in to what is being discussed. These informal assessment conversations provide good opportunities for formative action by the teacher as they provide inroads into student thinking.

In this context a simple initiation-response-evaluation (IRE) sequence would not be considered as good formative practice because evaluative feedback provides insufficient information regarding next steps that can advance learning. In our view, good formative assessment practice relies on teachers making use of the learning evidence that was gathered and goes beyond recognising what students can and cannot do, to making informed decisions on next steps and acting on them. It is therefore proactive and responsive in its approach. The challenge for the teacher is recognising that evidence, deciding what that evidence indicates in terms of learning and then forming an appropriate response that takes learning forwards.

The ASSIST-ME Project

This chapter explores how groups of teachers and researchers in four countries (England, Finland, Cyprus and France) tried to use assessment approaches designed to capture and interpret oral data collected during inquiry activities. Our aim was to investigate how classroom interactions supported learning and assessment during inquiry activities. This informal dialogue is referred to as interactions on-the-fly and includes episodes of dialogue selected from the classroom talk, where teachers use questions to probe student thinking during the inquiry activity, which the teacher can then respond to in various ways.

Each partner country was interested in the following research questions:

1. What are the characteristics of interactions on-the-fly in inquiry classrooms? And how are these interactions implemented?
2. How can interactions on-the-fly in inquiry classrooms be supported with tools and professional development?
3. What are the affordances, dilemmas and constraints of interactions on-the-fly in inquiry classrooms?

Within each of the four countries, interaction on-the-fly data was collected from classrooms by audio or video recording and transcribed. While each of the four countries collected these data from classroom settings, the style of inquiry activity differed across different countries, subjects and school contexts. Because the data was collected during the normal course of teaching and learning, the researchers selected sequences from lessons where they judged that development or use of inquiry competencies were taking place. So, for example, the French data were captured during whole class discussions, while the English data were generally collected as the teacher moved around observing and discussing with small groups of students. While these different classroom scenarios, to some extent, shape the ways in which the talk proceeds, nevertheless, the roles of the teacher and the students are much the same. The teacher is working to generate student talk that reflects their ideas and thinking so that they might assess current understanding to inform lesson planning. Meanwhile, students are able to hear how their peers are interpreting and thinking, about either a similar or the same inquiry activity, and so the classroom talk provides both a springboard for ideas and a means of checking how individual ideas are progressing.

The data were then analysed using a protocol developed by Ruiz-Primo and Furtak (2007) who modelled assessment conversations as a four-step cycle, where the teacher **E**licits information from the student by formulating a question, the **S**tudent responds, the teacher **R**ecognises the student's response and then **U**ses the information collected to inform student learning. This is known as the ESRU cycle and has been empirically derived from analysis of informal assessment conversations that took place during inquiry lessons in some US science classrooms.

The ESRU cycle provides a useful framework to characterise interactions on-the-fly because it enables a differentiation between instances where teacher is eliciting information from the student (e.g. formulating a question to probe student's ideas) and instances where the teacher is using the information that was collected to push the learning forwards (e.g. building on a student's answer). In their study, the students' performance on a multiple-choice questionnaire administered pre- and post-instruction and, on embedded assessments, was higher for classes with teachers who used more complete ESRU cycles. Ruiz-Primo and Furtak argue that the more cycles a teacher completes, the more likely it becomes that the information gathered from students was used in a helpful way for learning purposes.

In the ASSIST-ME study, the classroom interactions were analysed using the ESRU main categories which were applied to each speaking turn (one speaking turn is defined as the non-interrupted speech utterances of one single individual; this could be either the teacher or student) within the interaction on-the-fly data. From this analysis we were able to identify the number of complete cycles that arose in each classroom and also investigate how incidents of complete cycle and part cycles fitted with events within the inquiry discourse.

The vignettes from the different countries illustrate some of our findings and elaborate on these ideas and provide insights into how informal interactions in the inquiry classroom can be used to create feedback loops that support learning while providing some indications of the progress or lack of progress that specific students or groups of students are making within an inquiry scenario.

Vignettes

England

Interactions on-the-fly were studied in England in the context of primary science, secondary science and secondary mathematics classrooms.

For primary science, the Local Working Group consisted of nine teachers. It was decided from the start that structured bounded inquiries, as opposed to open inquiries (Wenning 2007), would best support the group, whose experience with and confidence in primary science were mainly low. The inquiry activities and foci for assessment were developed by the King's College team, tried by the teachers at the meeting and then implemented in their classrooms.

For secondary level, a local working group of five science teachers and one for eight mathematics teachers were set up. The science teachers were experienced in inquiry learning, having been involved in a previous EU inquiry project, and confidently developed some new activities to be trialled in their classrooms. For this project, the teachers developed three new inquiry activities – the double-decker bus, electromagnetism and exploring phenotypes. These were tried over three rounds of implementation and improvements and variations on the original themes developed, alongside a consideration of which inquiry competences could be assessed within each activity and how this might be done. The secondary mathematics teachers were less experienced in inquiry learning and were encouraged to use problem-solving activities from the nrich.maths.org website, such as the Towers of Hanoi and Peaches. For this study, a total of 18 science and mathematics lessons were analysed using the ESRU coding. The following two episodes illustrate some of the ways teachers used assessment on-the-fly in inquiry classrooms in England.

Episode 1

The following episode arose in a lesson about conditions for mould to grow on bread in a primary class of 30 children, aged 9–10 years. This lesson occurred towards the start of the year (November 2014). The episode took place after children had looked at bread samples they had placed in different environmental conditions.

The inquiry was set within an engaging context, authenticated through an email read-out by the teacher from someone requesting information about mouldy bread. This inquiry opened up with a whole class discussion where the success criteria, a vehicle for communicating what effective learning in science inquiry looks like, were shared with the class. The idea was that they should use their observational skills and inferences to work out which condition had most effect on mould growth. In the episode, the teacher is collecting results and ideas from different groups in a whole class discussion. The class were then sent off to interrogate, in groups of five, a range of data sources. These key characteristics of practical, collaborative, group

Table 4.1 The coding at the end of each line indicates the ESRU analysis

Turn	Speaker	Transcript	ESRU code
1	Teacher	Lovely and then Aesha's team?	E
2	Student	We found out that [inaudible] fungi actually grown [inaudible] and that when it starts off, it actually grows from the mould but you can't really see them [inaudible].	S
3	Teacher	Oh right, okay then and it works its way outwards, that's why it looks like wave effects, doesn't it?	R
4	Teacher	Do you think it's the same fungi or mould on the bread?	U
5	Student	No.	S
6	Teacher	How can you tell me that's true or not true, [inaudible], what clues have you picked up that tells me that maybe what's on there is different?	U
7	Student	Because some of them are different colours.	S
8	Teacher	Okay, so colour would be an observation that you've made; you might want to put that down in your notes.	R,U

inquiry set within meaningful, engaging and authentic contexts were a strong, common feature of all the primary inquiry ASSIST-ME lessons and helped to set up assessment conversations (Table 4.1).

During this assessment conversation, the teacher invited different learners to share their initial analysis of the range of data sources. These included some written texts about fungi and some sealed samples of bread containing mould for close observation. In this particular assessment conversation, there is evidence of the teacher attempting to work with student contributions, use these and build on these to develop a whole class understanding of how the observations they make during an inquiry (e.g. the wave effect, different colours of mould) could relate to and develop their conceptual understanding (different moulds or fungi on one bread sample). The teacher employed a combination of closed questions (e.g. 'Do you think it's the same fungi or mould on the bread?') and open questions (e.g. '…what clues have you picked up…'). As a consequence, the teacher used the evidence from the answer to the closed question to elicit further thinking through their more open question, creating an extended ESRU cycle. This then elicited individual student thinking, within a whole class discussion, thus making learning about the nature of science inquiry (in this case the relationship between observation, evidence and conceptual development) explicit to the whole class.

In the primary inquiry lessons, the interlinking of ESRU cycles was common with the use (U) of one cycle setting up the elicitation (E) for the next. This flow and merging of cycles tended to occur when the teacher was trying to help students share ideas and build on one another's findings. The teacher supported the classroom discussion by helping make connections and linking the ideas of different students or different groups with an aim to try to bring a reasoned argument to explain the results. In this way, the teachers' use of elicited ideas was framed to create more discussion through comparison, reason and evaluation, and so the teacher was deliberately taking a more divergent response and not closing down the cycle with a response that affirmed or rejected an idea. In this way the teacher encouraged students to think more deeply about their findings.

Episode 2

The following vignette comes from a lesson carried out by one of the secondary science teachers with a class of 18 students aged 15–16 years. This 90 min lesson focused on an inquiry into factors influencing the strength of an electromagnet. One of the learning aims of the lesson was to get students to derive a mathematical relationship from the data they had collected and use it to support the formulation of evidence-based conclusions (Table 4.2).

In this episode, students are writing their report of their inquiry in class. The teacher circulated and checked on some aspects of their reports as they are being written. She responded to one student's written conclusion and noticed that the stu-

Table 4.2 Students write their report of their inquiry in class

Turn	Speaker	Transcript	ESRU code
1	Teacher	[Teacher reads aloud students' written conclusions] These results show that the independent variables are directly proportional. Does it?	E
2	Student	Yeah… because as this increases this increases and as this increases this increases [pointing to values in a table].	S
3	Teacher	Yeah? What does directly proportional mean?	<u>R, U</u>
4	Student	That they increase together.	S
5	Teacher	Is that all it means?	E
6	Student	Is that something to do with equal?	S
7	Teacher	Right. If something is proportional, it means that as one increases the other, one increases in a straight line. Ok? If it is directly proportional, it means that as one increases the other, one increases in a straight line that passes through the origin.	R, U
8	Student	Oh!	S
9	Teacher	{Pointing to the values in the table}Oh! Do you know that this is a straight line?	E
10	Student	Yeah.	S
11	Teacher	How do you know that?	E
12	Student 1	Because as this increases this increases and…	S
13	Student 2	No, it won't be an exact straight line…because of this, look! {pointing to a specific value in the table}	S
14	Student 1	Oh!	S
15	Teacher	That curve shows as this increases, this increases, yeah? [teacher draws different examples on a sheet of paper]and that curve shows that as that increases that increases, yeah? Neither of those are proportional or directly proportional.	U

dent makes a claim of a direct proportionality relationship between the number of coils of a copper wire around an iron rod and the number of paper clips picked by the electromagnet. The teacher questioned the student and gradually brought the rest of the group into the discussion. The discussion evolved as the teacher probed student understanding on the concept of direct proportionality and guided the discussion using a combination of open-ended (turn 11) and closed questions (turns 3, 5 and 15), as well as exposition of information (turn 7). This guided the students in a particular direction, which, in this case, centred around collecting more data points before plotting the data into a graph.

In this episode, the teacher's attention to detail, as the students interpret and explain their findings, enabled the teacher and students to look more closely at the inquiry findings and the student interpretation of events. This allowed the students to rethink their initial interpretation of their results and the teacher to guide them towards an improvement in their thinking. Again, in terms of the patterns in the classroom interactions, the talk moves away from simple ESRU cycles to extended and linked cycles, where the use (U) of one cycle is positioned to elicit (E) another cycle.

Interestingly, the talk towards the end of this episode moves from alternate teacher-student to student-student, with student 2 redirecting student 1 to reconsider specific data points. This move from teacher-directed talk to student-directed talk became more common in the second and third implementation trials as some of the teachers became more skilled both in recognising likely avenues that students may take with an inquiry and in asking more open questions. These moves represent an increase in student agency that plays an important role in inquiry learning. It also creates more opportunity for the teacher to collect evidence of student thinking and make judgements about their learning through inquiry activities.

In this interaction, we observe two ESRU cycles that in fact can be considered as a single extended or interlaced cycle (turns 1–7). The important feature here is that the teacher is using the information gathered through eliciting questions to ask follow-up questions and give information that aims to push the learning forwards. By asking the students about what they think their data is telling them challenges the students to reflect on the questions they are trying to find an answer to and whether their results are providing evidence towards an answer (turn 11). The teacher wants to know current student thinking as they attempt to communicate their findings and whether they are simply presenting results or thinking more deeply about what the results are telling them about the phenomena of electromagnetism. In this way, the teacher is implicitly directing the students to utilise their inquiry competency of analysis, as they wrestle with the activity and she garners their current performance level in doing this. In this episode, the students have not thought through whether the results they are collecting are valid and the teacher's redirection, once she has ascertained what they are finding problematic, offers them a way forwards to consider. The teacher is not directly correcting work here but realises that the students have made assumptions rather than carefully analysing their data. In doing this she redirects their inquiry practice to offer them a way forwards. What is interesting here is that the teacher is allowing a considerable amount of student autonomy in the

Table 4.3 Data from different subjects and topics

Subject	Competence	Topic	Educational level	Teachers	Students	Lessons
Math	Problem-solving	Division, geometry	Primary	2	42	8
Physics	Investigations	Electricity, light and waves	Low secondary	2	45	7
Math	Problem-solving	Geometry, integral	Upper secondary	2	52	7

decisions they make and yet is ready and waiting to guide ideas and suggest quite direct next steps, once she notices what she considers is a discrepancy. The dilemma between being the assessor and being the guide during the inquiry activity is sometimes a difficult and tricky tightrope for teachers to take.

Finland

All the Finnish teachers who participated in the study had more than 10 years' teaching experience in school. However, they were not very experienced with inquiry-based learning. They often teach using practical work and hands-on activities, but these are typically teacher-guided closed tasks instead of open inquiry. The inquiry activities which were implemented in the study were designed through co-operation between the teachers and researchers to fit in with the teachers' course curriculum. Then the activities were revised together at the lesson plan level (Table 4.3).

Twenty-two lessons were analysed using the ESRU coding. Analysis indicated that there were 235 (56%) full ESRU cycles and 185 (44%) partial cycles. While there was a range of different inquiry scenarios, analysis of the ESRU cycles gave some indication of the nature of the inquiry-focused talk in these classrooms. In the Finnish lessons, ESRU cycles were often identified as partial due to lack of U-component. In these cases, the teacher, for one reason or another, did not use information from student learning. Another commonly missing component in the ESRU cycles is elicit (E). In many of these cases, the teacher does not need to elicit explicitly, because he/she is already getting information from the student's question or statement.

Episode 1

The first example is from a primary mathematics lesson (third grade, 9-year-old students); a student initiated (S) the following discussion. The teacher (Maria) asked students to divide 24 pieces of pasta between three people, using real objects. For this and other similar problems, they were asked to explain their solution on posters which would be presented later. Division as a topic was not taught formally before the inquiry activity (Table 4.4).

Table 4.4 Student-initiated question

Turn	Speaker	Transcript	ESRU code
1	Student	Maria, how should we do it? Because [inaudible] and eight. Divided by eight	S (E)
2	Teacher	Mm. By eight?	R
3	Student	Yes.	S
4	Teacher	Are there eight of you?	U
5	Student	No	S
6	Teacher	No. Read the task one more time. How many?	R
7	Student	Everyone gets eight pasta pieces	S

We see here, in turn 1, that a student initiates the question. Student-initiated conversations seem to be quite common in Finnish classrooms. In three closely analysed physics lessons, 36% of on-the-fly interactions were initiated by students (Correia et al. 2016). This suggests that the foci in these lessons have considerable student agency and direction.

Episode 2

The second example comes from a physics lesson (seventh grade, 13-year-old students), in which students have worked with an inquiry task for plane mirrors. They have made a prediction and observation, and they are working with their explanation. In the following excerpt, the teacher probes, many times, until he gets information about students' ideas. The teacher then helps students to express their idea and prompts them to write it down (Table 4.5).

The E-S-E-S structure continues for turns 1–9 where James, the teacher, recognises students' ideas. Turn 11 is an evident point at which he starts to use information to help students to express their ideas. So there is an extended elicitation section, in the classroom interaction, before the teacher uses the student responses. In this way, the teacher is surer that he has understood student thinking before taking formative action. In this way, the formative action is likely to be more appropriate to the learning needs of the students. However, we noticed that in the majority of lessons teachers' further probing actions, through extended elicitation, seem to be less frequent than quick interpretations of student response in the majority of classrooms.

ESRU main codes can be used for identifying the basic elements of on-the-fly formative assessment in the level of speaking turns, but they do not reveal the meaning of conversations per se. Thus, we performed data-driven analysis in order to characterise the nature of formative assessment at the episode level in one physics and one mathematics lesson (Nieminen et al. in press). First, these episodes were ESRU coded and then categorised. We found that there were two dimensions which characterise teachers' on-the-fly practices: (1) the speed with which teachers elicit information from students (quick interpretation or further probing) and (2) whether

Table 4.5 An example of the 2B type of interaction comes from a physics lesson

Turn	Speaker	Transcript	ESRU code
1	Teacher	What is going on here?	E
2	Student	Well, when… if we put that… that when…	S
3	Teacher	Are you working with conclusions?	E
4	Students	Yeah	S
5	Teacher	Okay. Do your prediction and your observation differ?	E
6	Students	Yeah.	S
7	Teacher	In what way?	E
8	Student	We thought as if I sit here. Then the mirror is there, so it would be mirrored to this corner but it is in that corner.	S
9	Teacher	Hm. So, you were here. Here is the object. You sat here. And you see the image… you supposed that it is here but is on this side.	R
10	Student	Yes.	S
11	Teacher	Well. So, you can answer: do your prediction and observation differ?	U
12	Student:	Yeah.	S
13	Teacher	Yes. How does the person's location influence the location of the image?	U
14	Student	It is mirrored like… if you look from the side the object is mirrored to that corner where you sit or stay.	S
15	Teacher	To that corner? So, you considered that the image is like on the different side from the object than your eye, but it is on the same side…	R
16	Student	Yeah.	S
17	Teacher	…than your eye. Okay. Yes it is. Yes. Then just write that.	R

Table 4.6 Four kinds of on-the-fly formative assessment interactions

	1 Quick interpretation	2 Further probing
A Help to take the next step	1A	2A
B Help to express thinking	1B	2B

or not teachers use that information for helping student to proceed or to express students' thinking. This results in four kinds of interactions as shown in Table 4.6.

In the vignette about sharing pasta pieces, the interaction is an example of the first conversation type 1A (quick interpretation and help to take the next step). In this discussion, the student mentions incorrectly 'dividing by eight' even though, in the situation, 8 is not the divider. Maria, the teacher, recognises the confusion and guides the student to notice that 8 is the answer to the task. In this case, the teacher did not ask further questions, such as why they are thinking about dividing by 8, and so her interactions are not probing but action based.

An example of the 2B type of interaction is the physics lesson above, where the teacher purposefully instigates thinking through his role in the interaction. The questions that the teacher asked challenge the student's explanation of the phenomena and required the learner to make decisions that he then had to justify and unpack. Twice the teacher returned to the same question – do your prediction and your observation differ? (Turn 5 and Turn 11) – and this highlighted not only to this student what is important in this inquiry but also communicates the idea to the rest of the class.

The ESRU framework by Ruiz-Primo and Furtak suggests set steps in the formative process that is usually initiated by the teacher. With our Finnish teachers, we found that sometimes the ESRU cycle is partial due to lack of eliciting (E), because a student initiates a conversation, rather than asks a question. Students' initiations are very important, because the teacher gets directed to student thinking and that can be information which the teacher may not get otherwise. Hence, it is important that the teacher aims to generate and maintain an atmosphere where students have the courage to ask questions, make statements and reveal their thinking.

France

Assessment of inquiry competences was studied in the context of two lower secondary schools in a socially and economically deprived area of the outskirts of Grenoble (France). The project teachers were three mathematics, three physics, two biology and two technology teachers (six females and four males). Together with teacher educators and researchers, the project teachers formed a working group that had been cooperating for three academic years (September 2012–July 2015); the data analysed in this research had been collected during the final year.

Because we were interested in capturing evidence directly from the lesson, the formative assessment tools were designed as a means for the teachers to structure their interaction with the students, particularly during the classroom discussion. Our focus was on the on-the-fly conversations that happened between the teacher and the class or the teacher with a group of students. We developed a tool, that we called a 'progress chart', which allowed teachers to make explicit their goals and students to locate themselves using detailed criteria ranging across four levels: newbie, beginners, experienced and expert. This tool helped the teacher to specify their goals and to anticipate students' attitude and more specifically allow them to anticipate what kind of answer may fit with these goals. This 'progress chart' was given to the students as a means of helping them understand the learning goals, to situate their knowledge and understanding regarding these goals and to outline the next steps for achieving them.

We analysed the on-the-fly assessment practices of six teachers. The teacher-student interactions took place during the episode dedicated to argumentation. Students in groups presented and discussed in front of the whole class their predictions and their respective explanations.

Episode 1

The first illustration came from a biology classroom. Students were determining the environmental factors that are essential for a plant's germination. During this episode, the teacher led her students to understand the phrase 'absolutely necessary', so that they could distinguish between those variables that had a direct effect on germination and other variables which did not. Understanding this aspect was a crucial issue for their design and planning (Table 4.7).

This short extract shows that the teacher had chosen not to directly ask her student to repeat the definition provided by the textbook. On the contrary, she asked a very open question that asks the student for her interpretation of the answer (Turn 1). Then she repeats the student answer to affirm that the student had understood the idea, and, eventually, she engaged the students in deepening this answer (Turn 5). This short episode resulted in the involvement of the whole class. While the teacher conversed with one student, the others looked on and reflected on how they might have answered the questions. Through these interactions, both the teacher and the students were able to make explicit their own understanding of a construct that is central for the current inquiry. This enabled them to envision more precisely the next steps of this inquiry.

Table 4.7 On-the-fly feedback with an open question

Turn	Speaker	Transcript	ESRU code
1	Teacher	Which environmental factors are…?	E
2	Student	Absolutely necessary	S
3	Teacher	Absolutely necessary.	R
		What does it mean 'essential' Sarah? Something that is absolutely necessary is something…	U
4	Student	That is something we need for doing something.	S
5	Teacher	For doing something.	R
		In our case, what do the seeds need that is absolutely necessary to germinate?	U
		Sarah.	E
6	Student	They need water	S
7	Teacher	Yes, water.	R
		And then …	U
8	Student	They need light	S
9	Teacher	Are you sure?	R
		Look at the table (indicates class result table)	U

Episode 2

The second episode came from a mathematics classroom dialogue about students' solutions to a problem of area measurement. Students had to work out the area of a park drawn on grid paper (Table 4.8).

This teacher chose to promote the exploration of the issue by his students instead of giving them the right answer or asserting that they are wrong. In this episode, we see that a student confused the squares within the diagram boundary with the unit of measurement itself (m^2) when measuring the area of a park. During the classroom discussion, he gave room for each student to report and justify their solution. In this part of the activity, the teacher provides opportunity for students to change their solution, in response to alternatives proposed by their peers. So we see a change in the teacher's role here from expert to guide, as he allowed students to provide feedback to one another.

In each classroom, all students participated actively in the inquiry, even those who had been reluctant to take risks in previous school activities. For supporting these complex interactions, the teachers used a 'progress chart' as a formative assessment tool that described four levels of achievement per competence, pertinent to the inquiry task. The teachers reported, during interviews, that the 'progress chart' was useful in the assessment process. First, it helps teachers to locate, during the lesson itself, each student's knowledge and competences amongst different alternatives. Second, it allows students to locate themselves with regard to the teacher's expectations. And thirdly, it encourages both teachers and students to envisage goals and strategies that are attainable in short-term scales.

In general, our results indicated that all teachers succeeded in implementing on-the-fly assessment strategies during whole class discussions in inquiry lessons. Most teachers (5/9 teachers) enacted complete ESRU cycles more often than ESR or ES cycles. Two teachers enacted ESRU and ESR at a similar level, while the remaining two teachers had interactions where ESR cycles were dominant. This suggests that our project teachers were able to collect assessment evidence during the inquiry activity, and most were able to act on this information in a formative manner.

Table 4.8 On-the-fly feedback with teacher promotion of an exploration

Turn	Speaker	Transcript	ESRU code
1	Teacher	So what happened in your case?	E
2	Student	Well I was counting the number of squares that could be put in the park…	S
3	Teacher	Okay … doing that …	R
		You mixed what and what?	U
4	Student	The unit and the squares …	S
5	Teacher	Yes, you have mixed unit and squares, but you was not so far off. Apart from that we wanted all the measures in units and not in squares. What do you need to change?.	R U

Cyprus

Researchers and teachers in Cyprus formed local working groups, where teachers were trained on applying specific methods of formative assessment and then, along with researchers, designed implementations in their classes.

Evidence was collected from the classroom interactions, particularly during whole class discussions, where the teacher raised specific issues relevant to the experimental design, the inquiry itself or interpretation of the results. To some extent, these discussions were planned, in the sense that the teacher identified beforehand issues that she/he should be aiming to bring into focus.

We were interested in identifying instances in which the ESRU cycle happened to break for some reason and elaborate on the different reasons that we are able to identify. We anticipated that this would provide evidence of challenges and intricacies associated with sustaining ESRU cycles. Our aim was to reveal possible trends or patterns in order to identify the variation as to what has caused the ESRU cycles to break. On the other hand, we were also interested in any patterns that might reveal possible factors that have facilitated the completion of ESRU cycles.

In an attempt to shed more light into the intricacies underlying the teachers' attempt to employ interactions on-the-fly as a formative assessment method, we focused on instances where either important information (i.e. contributions made by the students) went unnoticed during the discussion or was used in a nonoptimal manner. We describe two types of challenges that were identified in two episodes.

Episode 1

In this episode, the teacher has to make a choice as to whether to address certain issues that come up during the classroom discussion that may not have been planned for or to stay within the lesson plan (Table 4.9). This can be a difficult decision as sometimes students raise ideas that may have a profound influence on understanding and the ways in which they interpret the ideas in the lesson. This is illustrated in the following excerpt from the classroom dialogue about the forces that are applied on an object placed on a table.

In this case, through the discussion, the students had the opportunity to express their ideas regarding the interaction between two objects. The idea of action/reac-

Table 4.9 The teacher has to make a choice as to whether to address certain issues

Turn	Speaker	Transcript	ESRU code
1	Teacher	What can we think (which forces are applied on the object)?	E
2	Student 1	The Earth and the table apply forces on the object.	S
3	Student 2	And the object applies a force on the table.	S
4	Teacher	Let's leave that for now. What did we say we are interested in now?	R
5	Student 1	The forces that are applied on the object.	S

tion seems to be underlying how students actually conceptualise the subject under discussion, but the teacher led the students away from this to focus only on the forces applied to the object (Turns 5 and 6). While, for the teacher, there was a clear distinction between the 'forces being exerted by an object' and 'the forces being experienced by an object'; this distinction seemed unclear for the students and is therefore likely to underlie their interpretation throughout the discussion. This is evidenced by the fact that this specific idea emerged repeatedly at different points in the classroom discussion. For instance, the excerpt below shows how this idea surfaced again later in the same episode. Again, the teacher instructed the students to focus on the forces exerted on the object and not consider balanced forces (Table 4.10).

The selection to avoid discussing this particular idea (action-reaction) might reflect a strategic choice made by the teacher, or it could be that the teacher found it impossible to think about the idea from the students' viewpoint. One possible reason for this strategic choice could be related to the view that teaching evolved in a linear manner, which did not encourage deviations from the lesson plan prescribed by the textbook (which, in this case, did not include any reference to the idea of action/reaction). It is a challenge for teachers to overcome these views that might prevent them from realising the influence of students' misconceptions on developing scientific ideas. Teachers need support in recognising ways to take students' partial understanding forwards in a way that acknowledges and addresses such misconceptions as they arise in the classroom.

Episode 2

Interpreting the discourse that unfolds in the classroom is complicated by the possibility that the inputs contributed by individual students or the teacher could be resting on tacit assumptions, not necessarily shared by all members of that specific learning community. This might influence how the discussion evolves; thus, there is a challenge associated with detecting such cases when they occur. This could have an effect on the demands that are placed on the teacher in terms of detecting and

Table 4.10 The teacher chooses to avoid discussing a particular idea

Turn	Speaker	Transcript	ESRU code
1	Teacher	So this force is due to the direct contact with the table. Because the table is in contact with the object, some said that it exerts a force downwards and others upwards. Who is correct?	E
2	Student	Once the object is still, it means that the body exerts a force on the table, and the table exerts a force of the same magnitude on the object.	S
3	Teacher	You are right, but these forces are not exerted on the object. We are only interested in the forces exerted on the object. Which are these?	R U

Table 4.11 A misconception is identified

Turn	Speaker	Transcript	ESRU code
1	Teacher	Hmm. If I left it, it falls… why? Is it because of the ground?	E
2	Student	The gravity.	S
3	Teacher	What is gravity?	U
4	Student	The force towards the Earth's centre.	S
5	Teacher	Towards the Earth's centre. So the object falls because of the gravity?	R U
6	Student	Yes	S
7	Teacher	Where does the gravity come from? From the ground?	E
8	Student 2	From the Earth's centre	S
9	Teacher	So if we somehow remove Earth's centre, won't there be gravity?	U
10	Student 3	From the Earth in general	S
11	Teacher	From Earth in general. Ok…	R

managing such instances. In the following classroom dialogue, a crucial misconception arose that gravitational force is not an instance of interaction between masses but is an inherent property of a single object (e.g. the Earth) (Table 4.11).

In this discussion, some students seemed unclear about gravity and assumed it is a property of the Earth (Turn 8 and 10). This subtle point, which was not addressed during the discussion, despite the opportunities for doing so (Turn 11), came into play later as illustrated below, when the teacher suggested ignoring the air in order to avoid discussion about its resistance when the ball is falling. However, because the teacher did not appreciate the students' point of view, the student idea was ignored, and the discussion continued along the lines the teacher had chosen (Turn 3). Overall, this example illustrates how challenging it is for teachers to identify hidden assumptions that students have about particular phenomena and respond at the time, in a way that could help learners overcome their misconceptions.

Discussion

Inquiry in STEM classrooms creates rich opportunities for talk and for making connections with learning from previous activities, both practical and theoretical. Language is at the heart of the learning process. The learner uses talk to engage with new knowledge and to try to understand it within their own personal frameworks, through interactions with other learners and their teacher. Part of this, they achieve through comparison with their previous thinking in that area, but the major part of this learning is in negotiating common meaning with others that are also engaged in the learning experience. In this way, new knowledge is socially constructed (Vygotsky 1978), and communication through dialogue is essential in achieving this. Studies on classroom talk show that dialogic teaching contributes to enhanced student participation and learning (Alexander 2004; Mercer and Dawes 2014).

Mercer (2000) and Alexander (2004) both focus on the talk repertoire that teachers utilise in their classrooms. Both present hierarchical repertoires in which the lower levels centre on teachers' telling or using questions and prompt that demand students to recall what has already been encountered. In the higher levels of dialogue, Mercer introduces the category of exhortation, which he describes as teacher talk that encourages students to think. Alexander, on the other hand, describes higher levels of talk as discussion and scaffolded dialogue, which he takes pains to explain are different from the 'bedrock of teaching by direct instruction' (p23). Through discussion and scaffolded dialogue, talk moves from exchange of words to development of ideas, from interaction to shared social meaning and from knowing to understanding. Through rich dialogue during inquiry activities, the teachers on our project were able to create opportunities where their questions and interactions probed student ideas and encouraged thinking. Through encouraging students to voice ideas and make decisions about the questions they investigate, the apparatus they select, the method they take or the strength of the data collected, teachers learn a lot about how students are thinking within that domain. The key to interactions on-the-fly is sifting the evidence of learning from the many utterances that are made in the inquiry classroom in order to analyse students' developing understanding and realise what support students need to sort out or consolidate their ideas.

Our findings, across all four countries, indicate that teachers can and do want to use assessment evidence in a formative manner in their classrooms. On-the-fly interactions make student thinking and reasoning visible to enable formative assessment by teachers (Osborne et al. 2004). The evidence that they encounter through on-the-fly interactions, with their students, provides teachers with information to guide how they shape various aspects of the lessons and how they plan for future teaching. At the same time, students are continually being challenged about their progressive ideas both from teacher questions and from the discussions the teachers or peers are having with fellow students. In an inquiry classroom, the talk provides immediate feedback to the learners in the classroom that helps them steer their inquiry approach.

Working formatively in inquiry classrooms requires skill and focus from the teachers. The feedback to students is enhanced in classrooms where teachers notice specific solutions, problems or innovative approaches in the inquiry activities and decide to initiate a conversation to highlight what has attracted their attention. In these situations, the teacher uses questioning both to check why the inquiry has been done in a specific way and to act as a stimulus for other students to consider their inquiry activity and scaffold their thinking. As the teacher probes one student's ideas, the situation is set up where other students, in that class or within that group, can think about the teacher's question and their peers' answers. Witnessing this type of interaction provides the impetus for all students to reflect on their understanding of the idea being discussed by the teacher and their peer and, in some cases, reconsider how the aspect under discussion is working within their inquiry activity. In this way, the on-the-fly discussion becomes a technique and stimulus for providing immediate feedback and scaffolding for students. Within the classroom, the public spectacle of the on-the-fly interaction allows learners to explore their ideas about inquiry and look for evidence that either supports or challenges their current thinking.

Through sharing ideas with other learners, steered by the interactions initiated by their teacher, students engage with exploratory talk (Scott and Asoko 2006) that helps develop their understanding within a social context (Duit and Treagust 1998). Through this approach, teachers create learning environments where students can learn 'with and from one another' (Heritage 2007, p. 144). This is very different from the 'delivery' approach sometimes seen in STEM teaching, where the teacher presents ideas as though they are facts to be learnt and where practical work is used to illustrate relationships, laws and theories. This is particularly important for science classrooms, where discussion results in selecting results that fit with a prediction (Roth and Roychoudhury 1992), rather than careful consideration of the events and ideas developed in the inquiry. Roth and Roychoudhury (1992) found that, while the developing knowledge in science classrooms might be socially constructed, the process undertaken may result from a desire to agree or disagree with an initial idea, rather than careful consideration and reconsideration of the developing evidence. The teacher's role is key here in steering student thinking towards consideration of evidence and inferences, rather than simply allowing them to agree or disagree with an initial prediction or choose between two explanations. The teacher focuses the learners on the task, the question they are exploring and the strengths or weaknesses of evidence presented, so that the talk is not a ritual (Jimenez-Aleixandre et al. 2000) but a vital component of the inquiry learning. Such an approach moves away from the reported authoritative modes of discourse in many science lessons that leave students with naive images of science (Driver et al. 1996).

Assessing through evidence from classroom talk works best in classrooms where students share ideas with one another and are willing to listen to and take note what their peers are saying. In other words, such classrooms demand effective group work. Lemke (1990) puts forward the idea that 'learning science means learning to talk science' (p22), which suggests a move from teacher-dominated talk to classrooms where students discuss the ideas. What is essential here is that the talk is genuinely stimulated by student ideas rather than being a teacher-led discussion.

One of the dilemmas for teachers is how to set up and develop collaborative group work behaviour in their classes. Surveys of such work have shown that group work can lead to enhancement of learning but that it often fails to do so because the group do not interact in a productive way (Johnson et al. 2000; Mercer et al. 2004; Blatchford et al. 2006). Many teachers do not have the skills nor can recognise opportunities in tasks where they can develop collaborative group work, where students have to reconsider their interpretation of, or attempts at, a particular task. Teachers therefore need help in creating classroom environments and selecting tasks that encourage and enhance collaborative group work, if assessment is to focus on the classroom talk.

The ESRU cycle, proposed by Ruiz-Primo and Furtak (2007), helps teachers focus on the purpose of their classroom interactions by highlighting the need to both elicit and use evidence. While there was relatively high incidence of complete cycles in classrooms in all four countries, which indicated that teachers were acting upon the assessment evidence that they were collecting, we found that there was some variation between teachers and between activities and subject areas. Because no obvious patterns emerged as to whether experience, subject area or some other factor correlated with the number of complete ESRU cycles, this encouraged us to look more closely at the patterns in the talk that was emerging within lessons. Interestingly, all four countries found that there were occasions when the teachers used incomplete cycles in their interactions. So sometimes evidence was elicited, but the teacher did not respond and make use of that evidence. This was evident in the vignette from Cyprus about forces. In the flow of the classroom interactions, the teacher may notice some aspects and decide to respond, while in others they pass over and ignore the evidence, as they may feel extending that specific idea may not be fruitful learning. Also, in some instances a teacher may miss evidence that arises in the interactions, either because they misinterpret what a student is meaning or because the evidence from the talk is too vague to decipher at that point or because their own understanding of that aspect is called into question by the thinking under the discussion. In all of these cases, the ESRU cycle is left as incomplete. However, we need to remember that the talk is only the surface indicator of the thinking that is happening within that classroom, and it may be that the evidence is being used either by the teacher or the learners or both, even when it is not obvious to an observer in the class. Indeed, it could be that later in the talk or even within another lesson, the specific evidence produced is revisited because it is more timely or better understood by the teacher or the learners.

In some situations, the ESRU cycle appears conflated, so that there are runs of ESR or ES or US. These situations tended to occur in classrooms where the teacher was eager to collect more evidence around an idea, either by further probing of one student's ideas or widening the conversation to include more students in the discussion. We see this in the 'Mouldy Bread Inquiry' in the English vignette, where there is a run of US. The teacher here is keen to build on the previous answer that the student gave, trying to develop the talk into a reasoned argument of why the student had made the inference that they did. This convergent approach to the discussion, where the teacher's question leads the dialogue along a particular path, was common within all of our data, with teachers taking evidence from the students' inquiries to help illustrate or highlight ideas and concepts that they wanted the class to focus on. So while the teachers wanted to create an open thinking environment within the inquiry, there were times when they wanted to draw students along a particular line of thinking and so adopted a more convergent approach in these instances.

The Finnish vignette illustrates the opposite effect. Sometimes the teacher asks a series of probing and follow-up questions to try and bring clarity to the student description and explanation of their ideas. So the analysis shows extended runs of ES, where each answer is used to elicit more detail, with formative action only being taken when the teacher is more sure of the student meaning. So the pace changes, with teachers sometimes taking time to probe extensively before taking formative action and, at other times, reacting quickly to a student comment. It is clear that both approaches are needed in classrooms, and it is the skill of the teacher in deciding which to initiate. In some cases, quick interpretation and help to take the next step are needed to keep the whole class work process alive. In the case of quick interpretations, the guidance is based on anticipated correspondence of the students' and the teacher's ideas; sometimes teachers get this right, and sometimes they get it wrong. At other times, through further probing, the teacher gets more information about the students' learning, and, at the same time, the students get feedback to explain more. In the further probing, guidance is based more on the student's ideas. Thus, quick interpretations are more indicative of authoritative guidance, while teacher probing leads to more dialogic guidance, which chimes with the ways that Mortimer and Scott (2003) and Lehesvuori et al. (2013) interpreted talk in science classrooms.

Seeking evidence through interactions on-the-fly is not an easy task because the teacher is exposed to various inputs made by the students, usually as a reaction to the teacher's question, and the teacher needs to make an immediate judgement on how to respond to these. Some of the student contributions are more likely to support the evolution of the classroom dialogue in a productive and useful manner, while others seem unfruitful. At the same, it is always possible that elaborating on even apparently useful contributions might end up taking the discussion far afield and distracting or confusing other students and possibly distracting the class from their inquiry or the teacher from their lesson plan. Also, not responding may discourage some students from playing an active role in future discussion or in thinking about their inquiry or both. These features pose an important challenge for teachers, which entails two aspects:

(a) How to make optimum decisions in real time as to which students' contributions to draw upon, in the sense that they are more likely to lead to productive discourse, and what aspects to suppress, at least temporarily
(b) How to ensure a stance that is responsive and attentive to students' contributions but also allows steering the discussion in an effective manner, towards the learning goals

Making decisions, in real time, about the next moves in the evolving dialogue, constitutes a very challenging and demanding task for the teacher.

In ASSIST-ME, we found that teachers needed support in carrying out on-the-fly assessment. First, they needed to check that the inquiry tasks selected, or their approach to interpreting the inquiry task in the classroom, allowed opportunity for classroom assessment. This meant that the teachers needed to be confident about what evidence to look for at specific points in the inquiry activity and how to encourage students to be explicit in providing that evidence. In some cases, this meant helping teachers develop questions specific to the inquiry task or the inquiry process. For the former, teachers needed to be aware of common alternative conceptions in student thinking for the topic being considered. For the latter, teachers needed to identify common myths in inquiry practice, such as 'averages suggest accuracy', or ways of opening up the discussion to encourage the students to add more to the discussion – 'Why do you think that?' or 'Does that happen in all cases?' or 'Who can add to that idea?'

Secondly, teachers needed help in deciding 'next steps' after they had reached a judgement about current understanding. This was achieved in Cyprus by providing differentiated criteria for each inquiry activity and in France by rubrics and progress charts. Both in the development of these tools and helping teachers understand how to use them, a considerable amount of professional development time was required.

An added complication is that, in reality, teachers do not use one source of evidence to make judgements in the classroom and tend to use a range of assessment methods to collect evidence of learning and combine these different pieces of information to inform their practice. When focusing on one assessment method within a professional development programme, it is important to help teachers understand how this aspect fits within their classroom assessment repertoire. Teachers will engage with inquiry-based pedagogical approaches and formative assessment with varying levels of confidence and expertise. The design and content of any professional development experience must respond to this. Key characteristics of professional learning arising from ASSIST-ME, across all four assessment methods, focus on providing teachers with an opportunity to:

- Develop an understanding of inquiry and how this understanding is translated into classroom practice.
- Develop an understanding of classroom assessment.
- Recognise and experience the kinds of teaching and learning activities that apprentice learners into inquiry and formative assessment. Inquiry-based pedagogy and formative assessment emphasise learner autonomy.
- Suggest approaches that can create formative assessment opportunities within an inquiry lesson.
- Become familiar with and implement frameworks that enable a teacher to interpret learning and to evaluate progress in inquiry based on this.
- Explore teachers' personal understanding of the specific subject matter that underpins each individual inquiry. Most importantly, the professional development needs to address how personal subject knowledge can open up and limit formative interactions with learners.
- Identify and critically reflect upon their formative practice. This might entail sharing some specific examples of evidence of learning along with their evaluative judgement based on this evidence, next steps for this learner and how this was communicated.
- Develop and share classroom materials for enactment of inquiry activities and assessment of inquiry competences. Teachers need support on how to adapt existing materials and develop new ones that support their teaching and assessment practices.

Evidence of learning that can be collected through interactions on-the-fly depends on how these conversations are set up, on what type of questions are used to initiate and develop talk and to what extent are the conversations based on students' ideas and evidence. Although there is a growing body of knowledge on what good classroom talk looks like, previously, little has been on the characteristics of classroom talk in the context of inquiry lessons and on how these interactions on-the-fly contribute to good formative assessment practice. Our findings contribute to this area in recognising how teachers can set opportunities for formative practice in inquiry lessons and can interrogate evidence with students so that students recognise what constitutes appropriate thinking and evidence of inquiry. They also highlight the need for substantial professional development of teachers and provision of resources and materials to enable teachers to take these ideas forwards and make them effective in STEM classrooms.

References

Alexander, R. (2004). *Towards dialogic teaching: Rethinking classroom talk*. York: Dialogos.

Black, P., Harrison, C., Lee, C., Marshall, B., & William, D. (2003). Assessment for learning- putting it into practice. Maidenhead: Open UniversityPress.

Blatchford, P., Baines, E., Rubie-Davies, C., Bassett, P., & Chowne, A. (2006). The effect of a new approach to group-work on pupil-pupil and teacher-pupil interaction. *Journal of Educational Psychology, 98*, 750–765.

Correia, C. F., Nieminen, P., Serret, N., Hähkiöniemi, M., Viiri, J., & Harrison, C. (2016). *Informal formative assessment in inquiry-based science lessons*. In J. Lavonen, K. Juuti, J. Lampiselkä, A. Uitto & K. Hahl (Eds.), *Electronic Proceedings of the ESERA 2015 Conference. Science education research: Engaging learners for a sustainable future*, Part 11 (co-ed. J. Dolin & P. Kind), (pp. 1782–1791). Helsinki, Finland: University of Helsinki.

Crawford, B. A. (2000). Embracing the essence of inquiry: New roles for science teachers. *Journal of Research in Science Teaching, 37*(9), 916–937.

Driver, R., Leach, J., & Millar, R. (1996). *Young people's images of science*. Buckingham: Open University Press.

Duit, R., & Treagust, D. F. (1998). Learning in science: From behaviourism towards social constructivism and beyond. In B. J. Fraser & K. G. Tobin (Eds.), *International handbook of science education*, *Part 1* (pp. 3–25). Dordrecht: Kluwer.

Duschl, R., Schweingruber, H., & Shouse, A. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Research Council.

Gillies, R., & Nichols, K. (2015). How to support primary teachers' implementation of inquiry: Teachers' reflections on teaching cooperative inquiry-based science. *Research into Science Education, 45*, 171–191.

Gillies, R. M., Nichols, K., Burgh, G., & Haynes, M. (2014). Primary students' scientific reasoning and discourse during cooperative inquiry-based science activities. *International Journal of Educational Research, 63*, 127–140.

Grangeat, M. (2016). Dimensions and modalities of inquiry-based teaching: Understanding the variety of practices. *Education Inquiry, 7*, 4.

Haug, M. C. (2013). *Philosophical methodology: The armchair or the laboratory?* Oxford: Routledge.

Hazelkorn, E., Ryan, C., Beernaert, Y., Constantinou, C., Deca, L., Grangeat, M., Welzel-Breuer, M. (2015). *Science education for responsible citizenship* (No. EUR 26893). Brussels: European Commission – Research and Innovation.

Heritage, M. (2007). Formative assessment: What do teachers need to know and do?. *Phi Delta Kappan, 89*(02), 140–145.

James, M., Black, P., Carmichael, P., Conner, C., Dudley, P., Fox, A., & McCormick, R. (Eds.). (2006). *Learning how to learn: Tools for schools*. London: Routledge.

Jimenez-Aleixandre, M. P., Rodriguez, A. B., & Duschl, R. A. (2000). "Doing the lesson" or "doing science": Argument in high school genetics. *Science Education, 84*(6), 757–792.

Johnson, D. W., Johnson, R. T., & Stanne, M. E. (2000). *Cooperative learning methods: A meta-analysis*. Minneapolis: Cooperative Learning Centre, University of Minnesota.

Kawalkar, A., & Vijapurker, J. (2013). Scaffolding science talk: The role of teachers' questions in the inquiry classroom. *International Journal of Science Education, 35*(12), 2004–2027.

Laursen, S., Hassi, M., Kogan, M., Hunter, A., & Weston, T. (2011) Evaluation of the IBL mathematics project: Student and instructor outcomes of inquiry-based learning in College Mathematics. University of Colorado, Boulder. Retrieved 28 Feb 2016 from: http://www.colorado.edu/eer/research/documents/IBLmathReportALL_050211.pdf

Lederman, N. G., Lederman, J. S., & Antink, A. (2013). Nature of science and scientific inquiry as contexts for the learning of science and achievement of scientific literacy. *International Journal of Education in Mathematics, Science and Technology, 1*(3), 138–147.

Lehesvuori, S., Viiri, J., Rasku-Puttonen, H., Moate, J., & Helaakoski, J. (2013). Visualizing communication structures in science classrooms: Tracing cumulativity in teacher-led whole class discussions. *Journal of Research in Science Teaching, 50*(8), 912–939.

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Norwood: Ablex Publishing Corporation.

Mercer, N. (2000). *Words and minds: How we use language to think together*. London: Routledge.

Mercer, N., & Dawes, L. (2014). The study of talk between teachers and students, from the 1970s until the 2010s. *Oxford Review of Education, 40*(4), 430–445.

Mercer, N., Dawes, L., Wegerif, R., & Sams, C. (2004). Reasoning as a scientist: Ways of helping children to use language to learn science. *British Educational Research Journal, 30*(3), 359–377.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—What is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474–496.

Mortimer, E., & Scott, P. (2003). *Meaning making in secondary science classrooms*. Maidenhead: McGraw-Hill.

Nieminen, P., Hähkiöniemi, M., Leskinen, J., & Viiri, J. (in press). Four kinds of formative assessment discussions in inquiry-based physics and mathematics teaching. Proceedings of Finnish Mathematics and Science Education Research Association 2015.

OECD. (2005). *Education and training policy. Teachers matter. Attracting, developing and retaining effective teachers*. Retrieved from, 28 August 2015, from http://www.oecd.org/edu/teacherpolicy accessed.

Osborne, J., Erduran, S., & Simon, S. (2004). Enhancing the quality of argumentation in school science. *Journal of Research in Science Teaching, 41*, 994–1020.

Rocard, M., Csermely, P., Jorde, D., Lenzen, D., Walberg-Henriksson, H., & Hemmo, V. (2007). *Science education now: A new pedagogy for the future of Europe*. Brussels: European Commission.

Roth, W.-M., & Roychoudhury, A. (1992). The social construction of scientific concepts or the concept map as device and tool thinking in high conscription for social school science. *Science Education, 76*(5), 531–557.

Ruiz-Primo, M. A., & Furtak, E. M. (2007). Exploring teachers' informal formative assessment practices and students' understanding in the context of scientific inquiry. *Journal of Research in Science Teaching, 44*(1), 57–84.

Scott, P. H., & Asoko, H. (2006). Talk in science classrooms. In M. Hollins (Ed.), *ASE guide to secondary science education*. Hatfield: Association for Science Education (ASE).

Shavelson, R. L., Young, D. B., Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., Tomita, M. K., & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education, 21*, 295–314.

Vygotsky, L. S. (1978). *Mind in society: The development of higher mental process*. Cambridge, MA: Harvard University Press.

Wenning, C. J. (2007). Assessing inquiry skills as a component of scientific literacy. *Journal of Physics Teacher Education Online, 4*(2), 21–24.

Chapter 5
The Structured Assessment Dialogue

Jens Dolin, Jesper Bruun, Sanne S. Nielsen, Sofie Birch Jensen, and Pasi Nieminen

The Importance of Dialogue in Assessment Practice

Teacher-led classroom dialogue is one of the most common instruction practices worldwide (Wiliam and Leahy 2015). In a school science context, science teachers' ways of managing the dialogue play a central role in mediating between students' everyday language, current understanding and alternative concepts on the one hand and the scientific language, explanation and concepts on the other hand (e.g. Lemke 1990).

In addition, a large part of the information that teachers obtain through informal formative assessment is through classroom dialogue (Ruiz-Primo 2011). Hence, introducing a dialogue-based assessment method in the classroom will to a large extent align with an already existing instruction and assessment approach. Such a dialogue-based assessment might be used for either formative purposes, summative purposes or – as we will argue – both formative and summative purposes.

We base our understanding of dialogue on the works of Bakhtin (1981, 1986). For Bakhtin, in a dialogue "every utterance must be regarded as primarily a *response* to preceding utterances [...]. Each utterance refutes affirms, supplements, and relies upon the others, presupposes them to be known, and somehow takes them into account" (Bakhtin 1986). In our conception, a dialogue consists not only of the words that are being uttered but also of the representations that are used in the dialogue (e.g. Roth 2000; Roth and Lawless 2002). This includes drawings on black-

J. Dolin (✉) • J. Bruun • S.S. Nielsen
Department of Science Education, University of Copenhagen, Copenhagen, Denmark
e-mail: dolin@ind.ku.dk

S.B. Jensen
King's College London, London, UK

P. Nieminen
University of Jyväskylä, Jyväskylä, Finland

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science
Education Research 4, DOI 10.1007/978-3-319-63248-3_5

boards, experimental equipment, a model on a computer screen and gestures, such as pointing to a blackboard or showing balance with both hands.

The importance of classroom dialogue as a necessary resource to learning is supported by socio-constructivist learning theories (Dysthe 1996; Leach and Scot 2002; Ogborn et al. 1996). Alexander (2006) points out that “talk is arguably the true foundation of learning”, and classroom talk can be enriched by feedback, which encourages students’ engagement in dialogue. The recent focus on feedback due to especially Hattie and Timperly (2007) is in agreement to these ideas, insofar as feedback can be seen as a response to student utterances that take them into account. Often, there is an emphasis on teacher feedback, but recent developments have emphasised the possible positive contributions of peer feedback. For example, Cho and Schunn (2007) have shown that the process of providing and receiving/interpreting peer feedback is a way for students to develop their science-writing competencies.

Classroom dialogues – like most formative assessments within the typical classroom – are quite informal in nature and are used differently by different teachers. A teacher may not make explicit – neither for him/herself nor for the students – the criteria to which a student should live up to. This will make it difficult for a student to recognise his/her own level, and there might not be time for the student to reflect on his/her next steps in learning. As a result the learning potential through formative assessment of a classroom dialogue may be limited (Ruiz-Primo 2011). Furthermore, in order to be effective and to make it possible to be used also for summative purposes, the assessment method must provide a relatively standardised approach to how it is administered (see Chap. 3). The SAD formalises the assessment process and forces the teacher to make explicit the goals and the criteria for their fulfilment.

However, the learning prospects in the classroom dialogue do not materialise by themselves. Students learning, including their ownership and motivation is largely dependent on the teachers’ strategic use of various forms of dialogue (Scott et al. 2006). A way to promote students’ ownership and value and their thoughts and ideas in the classroom is through teachers’ “uptake”. Uptake means that the teacher incorporates students’ previous answers into a subsequent question (Dysthe 1996). In addition, the teachers’ use of different question types open up for different types of student answers and knowledge. For example, inquiry-based learning is often dominated by open-ended (authentic) questions in order to recall and challenge students’ thinking and encourage debate and further studies.

From an assessment perspective, the dialogue can provide evidence on *what* and *how* students are thinking. From a formative perspective, the dialogue can make students’ thinking explicit. It can voice their understanding so that the teacher can recognise and act on it to promote learning (Harrison 2006; Ruiz-Primo 2011). This information can also be used for summative purposes to make the level of students’ understanding explicit. However, teachers’ way of orchestrating the dialogue influences *what* and *to what extent* students’ understanding is voiced.

The most common kind of classroom interaction is dominated by “one-way dialogue” where the teacher controls the content, speed and sequencing. This reduces students' opportunities to raise questions and to use their own previous knowledge, and it constrains students' ownership, motivation and learning. A challenge related to formative assessment is to create a dialogue between teacher and students and among the students to involve learners as active partners of their own learning. According to Dysthe (1996), this kind of dialogue is characterised by authentic questions, uptake and students' voices. However, this requires that teachers as well as students develop a kind of "dialogical assessment for learning culture".

We first describe the SAD in detail, its phases and the principles behind. We then pose the two research questions driving the chapter. The first research question is directed towards teachers' experiences using the SAD, and the second is linking a characteristic of the dialogue in the SAD with the quality of students' self-reflections. After a short description of the empirical data, we then answer the research questions one by one.

The first research question is answered straightforwardly using qualitative data, while the second involves the design of a new research method for analysing dialogues by transforming the dialogue into a network. We are then able to characterise dialogue maps and their relationship to student self-reflections and teacher preparation. The characterisation of dialogue groups leads to a typology of dialogues, and we link this typology to the qualitative findings in the discussion.

The Structured Assessment Dialogue

This section describes the aims, structure and principles behind the SAD. In designing the SAD, the aim was two pronged: on the one hand we wanted to develop a method resembling an already existing dialogue-based formative assessment practice, and, on the other hand, we wanted to develop a method that teachers could use for *both* formative and summative assessment.

The Aim of the SAD

As noted above, the prospects for learning through formative assessment of a classroom dialogue may vary and in the worst case be very limited. The limited prospects for learning may be caused by multiple factors. The SAD intends to address the following factors that are expected to severely limit the prospects for learning: (1) learning intentions and criteria, which are not planned, clarified and shared with students, (2) the role student engagement plays in their own and peers' learning which is minimal and (3) no or very little time for the students to reflect on his/her level and the next steps in learning.

Furthermore, in order to be effective and to make it possible to be used also for summative purposes, the assessment method must provide a relatively standardised approach to how the SAD is structured and administered (see Chap. 3).

To address the above factors, the structure of the SAD was designed with a defined content and time structure and with a clear division of roles among the participants in the classroom.

The Structure of the SAD

The structure in the SAD follows three well-defined time-limited phases with clear instructions for the teacher and the students of their role. See Fig. 5.1. The first phase (5-min) is a dialogue between the teacher and a preselected focus student about a specific aspect of a competence. The second phase (5-min) is a peer assessment process where a handful of students give feedback to the focus student. In the third phase (2–3 min), the focus student, the feedback students and the rest of the class engage in self-reflection on their own level of competence and areas for further learning. For summative purposes, the teacher then notes down the level of the focus student. Thus, one SAD session will lead to the formative and summative assessment of one student and possible formative assessment of many students combined with summative considerations.

The Principles Behind the SAD Model

The development of the SAD was based on the work of Christensen (2005), who developed the SAD to focus on eliciting student knowledge, student experience of acquiring knowledge, evidence of what has been taught, equal participation in the dialogue and evidence of reflection. The ASSIST-ME project mainly focuses on the assessment potentials associated with the SAD. Hence, we refined the method to

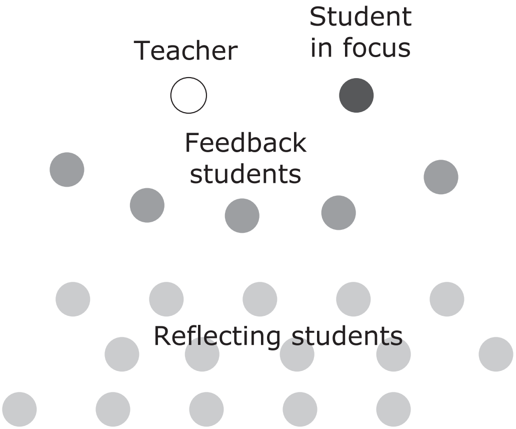


Fig. 5.1 The setting of the structured assessment dialogue. One student is in focus for a 5 min teacher-student dialogue, followed by a 5-min phase with five to six students giving feedback to the student in focus and ending with all students reflecting upon the dialogue and the feedback during a 2–3 min phase

take into account the following eight criteria that may impact how assessments affect students' learning:

(i) Effective assessment dialogues are learning goal guided (Ruiz-Primo 2011).
(ii) The learning intentions reflected in the goals are elaborated in terms of explicit assessment criteria (Hattie and Timperley 2007).
(iii) There must be consistency between goals and observable assessment criteria adapted to specific teaching sequences (Krajcik et al. 2008).
(iv) Likewise, there must be consistency between goals, teaching and assessment approaches (Bennett 2011).
(v) Student awareness of the criteria is important (Black and Wiliam 2009).
(vi) Student involvement, including self-assessment and peer assessment, is important (Black et al. 2004).
(vii) Feedback holds prospects for learning when delivered in the right manner, i.e. the feedback should be student and task targeted, timely in order to use the feedback, focused, specific and clear on how to improve learning (Shute 2008).
(viii) A short time interval between eliciting information on students' level of success and using this information in providing feedback to the student will benefit specific learning outcomes and students (Shute 2008).

In the following sections, we elaborate on how the eight criteria are reflected in the SAD. In each section headline, we provide reference to the eight criteria (i–viii) listed above. In the first section, we elaborate on how the assessment process in the SAD is guided by goals and their associated success criteria.

Goals and Assessment Criteria Constructed and Shared: (i), (ii), (v), (vi) and (vii)

Since learning goals describe the intended consequences of teaching and learning, they could form the basis for focusing and structuring assessment. Based on this assumption, the SAD is guided by specific learning goals identified and described by the teacher ahead of the teaching and assessment session. In our classroom implementations, the goals were related to two of the three general ASSIST-ME competencies (modelling or argumentation) contextualised in a specific subject area (e.g. electrical circuits, membrane permeability or geometry).

A teacher may not make explicit – neither for him/herself nor for the students – the criteria for assessing whether, or at which level, a learning goal is being achieved. This will make it difficult for a student to recognise his/her own level or provide feedback and engage actively with their own and others' learning. As a result the learning potential through formative assessment of a classroom dialogue may be limited (Ruiz-Primo 2011).

In the SAD, teachers are requested to subdivide the learning goals into a range of specific assessment criteria reflecting different aspects and levels of the competence being assessed. This is to avoid an unfocused assessment practice with a tendency to assess more general, trivial or managerial aspects. In addition, teachers are asked to share and clarify to students the range of criteria. The criteria are also used in the peer feedback as well as in the self-assessment phase. It gives students an opportunity to reflect on their current level of fulfilment of criteria aligned with the learning goals and on their next steps in learning. It is intended as a way of strengthening classroom learning cultures by having students engage actively with their own and others' learning.

Finally, providing the students with transparent assessment criteria would formalise the peer feedback, thus reducing the personal aspect among the students (e.g. friends, status).

Consistency and Adaption: (iii) and (iv)

In order to be effective, formative assessment has to be integrated into classroom practice (Wiliam 2011). Since teacher-led classroom dialogue is a very common instruction practice, a dialogue-based assessment, such as SAD, might feasibly facilitate consistency between instruction and assessment. A SAD session will typically be performed when a class has worked for some time with one or more competencies. The ASSIST-ME project focused on problem solving as it pertains to mathematics, investigation as it pertains to science and design as it pertains to technology and on the general competencies argumentation, innovation and modelling. Whatever the competency, the SAD will be situated within learning activities with which students have just engaged. SAD sessions can be implemented repeatedly at times where it makes sense for the teacher and students (e.g. when information is needed for adjusting instruction or providing feedback). In addition, students may be prompted to use artefacts (e.g. models, drawings, lab results) from the instruction during a SAD session to strengthen the consistency.

Planning and Timeframe: (i), (vii) and (viii)

In a dialogue-based assessment such as SAD, the timeframe for interpreting evidence of students' learning and using the evidence for the next dialogue turn is very short (typically, less than 10s). This short timeframe holds prospects for the teacher to continuously adapt response to student utterances and for allowing the next questions to move in different directions depending on the ways in which the students respond. However, the short timeframe may also be a real challenge for the classroom participants. The teacher and peers have to promptly interpret student's utterances and need to provide verbal feedback almost instantly. We believe this is only possible if the teacher carefully plans the strategy for the dialogue.

As stated previously, this means that the teacher must have formulated clear goals and assessment criteria in advance and considered how to use them subsequently in the dialogue.

In addition, the SAD allocates time for all students to reflect on his/her level and next steps in learning immediately after peer feedback.

Operationalization of the SAD

The teachers followed a particular schema, which had been constructed by the authors. Table 5.1 indicates this operationalization of the SAD in terms of materials and preparation.

Each phase characteristics were discussed with teachers and students before the first enactment of the SAD. For the first phase, the teachers were aware of choosing the student for the first enactment carefully. It should not be a too strong or too weak student, and it should be a well-formulating student. The students were told their roles in each phase. The teacher filled in a template with key questions covering the learning goals for the sequence to be assessed. The teachers were giving a short introduction to the Dysthe (1996) dialogical framework (invitation, uptake, valuing, etc.) – see later. Principles for good peer feedback were discussed in class as were the ideas with self-assessment.

Many teachers adapted this standard set-up to local needs and cultures. The first teacher-student dialogue was in some cases prolonged to 6–7 min, some teachers only used the SAD for formative purposes, some teachers had two to three students in focus, etc.

Table 5.1 Phases, teaching materials and preparation for the SAD

Phases	Teaching materials	Preparation
Focus student-teacher dialogue	Teacher tool filled in (first part)	Teacher has prepared questions that will help mapping student attainment of learning goals in terms of learning progression. Student in focus has prepared his/her presentation
	Criteria created and shared with the students	
Feedback from a group of students to the focus student	Discussion questions: What would be needed in order for the student-teacher dialogue to better cover the learning goal?	Feedback students have been given instructions on feedback guidelines
Teacher input to support feedback from students	Guidelines for peer feedback	
Self-reflection for all students	Reflection tool for the purpose of self-assessment in relation to how far they are in the learning process and how they proceed	Reflection questions need to be prepared and accessible to students, for example, online

On Combining Formative and Summative Assessment

The SAD gives evidence of student learning that is used mainly for formative purposes, but with a potential for summative use without distorting the formative aspects. The key reason for this is that the summative assessment happens at the very end of the ritual and is embedded in a formative process. We thus avoid the domination of the summative purpose; you tend to see when the two purposes are mixed (Butler 1987). As argued in Chap. 3, the summative assessment regime has produced educational environments, which are dominated by various high-stakes tests which demonstratively have led to backwash effects diminishing learning outcomes. Thus, teachers need methods for assessment that can provide summative data without being high stakes and which preserve validity and reliability. With this method of assessment, the teacher can summatively assess one student at a time. Students may also self-assess in a summative way (by self-grading). The viability of the method as a summative assessment tool for all students rests on the fact that it can be applied quickly throughout the year. Thus, not all students will be assessed in terms of all content, but a teacher might be able to assess particular competencies.

Research Questions

Given that the student-teacher dialogue is the central activity of the ritual, it is natural that the research questions centre on this dialogue. The purpose of the dialogue is for students and teachers to gain insight into the students' current level of attainment and the next steps in students' learning. As noted above we have two agendas. First, given the novelty of the method, how teachers perceive the method after having used it, and in particular it is interesting to investigate the challenges and affordances teachers see in implementing the method as part of their practice. Second, we want to investigate the structural and relational aspects of student-teacher dialogue, i.e. how specific types of dialogues lead to differently productive student reflections. Our framework for that is built around network analysis of the dialogues, producing what we call dialogue maps. Framed in these terms, our research questions are:

RQ1: What are the challenges and affordances perceived by teachers for using the structured assessment dialogue?

RQ2: How are groups of similar dialogues related to student self-reflections, teacher preparation and context?

RQ1 and RQ2 are in principle two independent questions, and they will be treated separately in the following. They can be seen as two different albeit-related ways of focusing on the dialogues. We expect the answers to the first question will benefit from being qualified by the answer to the second question. This is because we expect the answer to the second question to be in the form of a typology, which meaningfully characterises groups of dialogue maps. We then expect that we can relate affordances and challenges perceived by teachers to this typology.

Despite the RQs being different, they share the same essential background: The performing of the structured assessment dialogue. We describe this background next.

Background for Study

We analyse data from Finland and Denmark. Finland only gathered very little experience with the method, while the Danish implementations were more numerous.

In Finland the SAD was implemented in two lower secondary physics classes (same teacher) and in one upper secondary math class. In all classes SAD was conducted three times, but the first time was a practice without the data collection. Before SAD session students worked about 45 min with an inquiry task (electromagnetism or geometry) with the task-related levels of attainment. For the teacher-student dialogue, the teachers tended to select a student who had enough both social and subject competence.

In Denmark 8 teachers implemented a total of 20 SAD sessions: 11 dialogues in science (physics and biology), 5 dialogues in mathematics and 4 dialogues in technology. All sessions took place in upper secondary school (year 10–12) except for one session implemented in a lower secondary physics class. The student-teacher dialogues were video recorded, and the feedback sessions were either video or audio recorded. Furthermore, the teachers' plans for and evaluations of the sessions were collected together with students' self-reflections where this was possible: Some teachers didn't write down their plans for the sessions, some student reflections were lost and some teachers use reflection templates too different from the canonical version to be comparable to these.

The placement of the SAD sessions within the inquiry-teaching units varies from teacher to teacher, as does the number of dialogues implemented in each session. The dialogues lasted for approximately 5 min each, and the feedback sessions were typically somewhat shorter, 2–4 min for most of them. In the mathematics sessions, there wasn't a clear break between the dialogue and the feedback; rather, the dialogue gradually evolved into a feedback session when the teacher started asking the feedback students questions. Each of these combined sessions lasted 10–15 min.

Some of the Danish teachers were reluctant to take up the SAD with only one student participating in the dialogue with the teacher as they felt it deviated too much from their usual teaching which often relied on group work. Thus, in six of the sessions, the dialogue is between a teacher and a group of two to four students.

Table 5.2 Summary of data

Country	Completed teacher preparation forms (RQ1 + 2)	Student-teacher dialogue (RQ2)	Feedback sessions	Student self-reflection forms (RQ2)	Teacher focus group interviews (RQ1)
DK	11	20	20	314	1
FI	6	6	6	117	0
Total	17	26	26	431	1

Summary of the Available Data

The entire corpus of data consists of filled out teacher preparation forms, video recordings of the student-teacher dialogue, audio recordings of feedback sessions, filled out student self-reflection forms and teacher focus group interviews. Table 5.2 summarises the corpus.

Table 5.3 below is a summary of the contextual information we have available for the individual SADs. Most SADs were performed at the upper secondary school level. Some teachers chose to implement the SAD multiple times, while others only found time for one implementation. The teachers that implemented the SAD multiple times did so in different ways. Some teachers chose to implement the method multiple times with the same class for different topics. Some teachers tried the SAD in different classes but assessing the same topic. Finally, one teacher combined the two approaches.

The ASSIST-ME project conceptualised two sets of competences. The first set of competences was specific to the disciplines (science, mathematics or technology). The second set was a cross-disciplinary competence (*modelling*, *argumentation* or *innovation*). The subject uniquely determined the disciplinary competence (see Chap. 2). However, teachers chose the cross-disciplinary competency to focus on in the unit. 13 dialogues were part of a unit that focused on *argumentation*, while the remaining 13 dialogues were part of a unit that focused on modelling.

The SADs also differ in their durations, number of feedback students and number of focus students chosen. Ultimately, it was part of the research design to let teachers incorporate the method into their teaching as they saw best to fit the needs of their own classroom.

What Are the Benefits and Challenges Perceived by Teachers for Using the SAD?

To answer RQ 1, we used data from Denmark only. The data set consists of filled-in teacher preparation forms (n = 11), two semi-structured interviews of teachers, one focus group interview with teachers (Kvale 2007) and an open-ended questionnaire for teachers (n = 4).

Table 5.3 Summary of dialogues that were analysed in this study

ID	Teacher	Subject	Duration	#Students	Teacher form	Level
SAD 1	FI-1	Physics	3′15″	20	Yes	Lower
SAD 2	FI-1	Physics	4′15″	22	Yes	Lower
SAD 3	FI-1	Physics	2′40″	22	Yes	Lower
SAD 4	FI-1	Physics	4′20″	22	Yes	Lower
SAD 5	FI-2	Mathematics	6′30″	31	Yes	Upper
SAD 6	FI-2	Mathematics	6′35″	32	No	Upper
SAD 7	DK-1	Biotechnology	5′	32	Yes	Upper
SAD 8	DK-1	Biotechnology	5′	32	Yes	Upper
SAD 9	DK-2	General science	5′55″	33	Yes	Lower
SAD 10	DK-3	Physics	4′35″	35	No	Upper
SAD 11	DK-3	Physics	4′45″	35	No	Upper
SAD 12	DK-3	Physics	6′35″	29	No	Upper
SAD 13	DK-3	Philosophy of science	4′40″	33	No	Upper
SAD 14	DK-4	Technology	4′15″	26	Yes	Upper
SAD 15	DK-4	Technology	5′	24	Yes	Upper
SAD 16	DK-4	Technology	5′	25	Yes	Upper
SAD 17	DK-5	Technology	4′55″	27	Yes	Upper
SAD 18	DK-6	Physics	6′35″	9	Yes	Upper
SAD 19	DK-6	Physics	10′45”	9	Yes	Upper
SAD 20	DK-7	Physics	5′	33	Yes	Upper
SAD 21	DK-7	Physics	5′25″	33	Yes	Upper
SAD 22	DK-8	Mathematics	8′	21	No	Upper
SAD 23	DK-8	Mathematics	7′20″	21	No	Upper
SAD 24	DK-8	Mathematics	5′5″	17	No	Upper
SAD 25	DK-8	Mathematics	5′10″	21	No	Upper
SAD 26	DK-8	Mathematics	7′15″	21	No	Upper

The focus group interview and the semi-structured interviews were facilitated by two of the authors. The facilitator would ask the teachers about the key challenges and benefits of the SAD and of each of its different phases. The questions asked by the facilitator prompted teachers to explain how they had prepared, how they dealt with practical aspects of the SAD, how they managed the dialogue part and the feedback part, what they believed were important concerns and which opportunities they saw for the SAD. In all cases, the facilitator prompted teachers to provide the rationale for their ideas, comments and suggestions. In the questionnaire, teachers were asked to comment on (1) the main strengths and weaknesses of the SAD, (2) the opportunities and challenges identified in enacting the SAD, (3) suggestions to overcome the challenges encountered and (4) what they would do different if they had an opportunity to repeat the SAD session.

Both the interviews and the responses to the open-ended questionnaire were analysed using semantic thematic analysis (Braun and Clarke 2006). We analysed the interview by transcribing, reading, rereading and, in some cases, re-listening to the interviews using the RQ1 as an analytical lens. Specifically, we focused on the challenges and benefits perceived by teachers with respect to each of the eight principles behind the SAD. We used the same analytical approach for the questionnaire and the teacher preparation form.

We present our findings structured into two overall themes reflecting the RQ1: (1) Benefits of SAD as perceived by teachers and (2) challenges of SAD as perceived by teachers. Each theme is divided into subthemes based on the semantic thematic analysis.

Benefits of the SAD Perceived by Danish Teachers

The benefits can be divided into four subthemes.

Clearly Stated Learning Goals Subdivided into Explicit Assessment Criteria

The teachers saw several advantages in the formulation and utilisation of a clear statement of learning goals and explicit assessment criteria, such as enhancing students' involvement in the assessment and providing transparency in the assessment. Teachers highlighted the benefits of sharing the assessment criteria with the students. A teacher wrote, "The learning becomes explicit for the students". Adding to this another teacher said, "My students like it a lot and we will continue [to use the method]. Because they actually think that it worked - that the criteria were made clear and that they knew what to aim for. "The benefit of making criteria explicit to students is also illustrated in this quote, "I think that the students can really see what is required from them, when we speak 'language of physics'". In addition, teachers found the criteria very useful for students to provide peer feedback and self-reflection.

Moreover, teachers' utterance also indicates how sharing of learning goals with students in the SAD becomes coherent with, and benefits, upcoming teaching by activating students in their own learning, exemplified by this quotation, "I experienced that after a dialogue the students got better at setting up goals for themselves".

Another advantage in identifying a range of specific assessment criteria reflecting different aspects and levels of the learning goal is related to teacher use of questions to address different students' achievements. A teacher said, "During SAD, I mainly use the different assessment criteria for guiding my questions to address the differences between individual students' understanding and ability". This suggests that the teacher perceives the SAD as having the potentials for adapting the assessment criteria to the student in focus, even though the overall criteria are formulated in advance.

Interestingly, not only students seem to benefit from the teacher-generated assessment criteria. Teachers identified benefits relating to their own level of reflection when they formulated and used assessment criteria during the SAD. For example, a teacher noted "You have them [the assessment criteria] somewhere in your mind, but you still need to formulate them in a way that is suitable for students. It's a good exercise for the teacher, but something that requires preparation". In this self-reflection, the teacher acknowledges how critical and on the other hand also how difficult it is to formulate criteria explicitly for students even though the criteria are perceived as being already present in the mind.

A teacher used the learning goals for clarification and for documentation for parents' meetings and to provide transparency in grading during the semi-annual student conversation.

Timeframe, Structure and Discourse

In order for teachers to include SAD as part of the classroom repertoire, students need to be willing to participate. Some teachers stressed that the clear rules and short duration of the student-teacher dialogue helped to make the rather unfamiliar situation feel less intimidating for the focus student. The strict time limit of the SAD could then be seen as an advantage.

> The student tends to say, 'phew', that didn't take long. And the other students hear that, and then it may not be as dangerous to do. But it is a bit like an oral exam, the way we sit and the setup. *It is not as dangerous because it is done so quickly, so we need to be careful not to make it take too long. Then students may become afraid of it.*

The strict 5-min limit on the dialogue posed a challenge to most teachers. However, some teachers wanted to advance on the flexibility in the SAD and adapted the timeframe to the context based on their discretion,

> We needed to go into one of the last questions when the five minutes had passed. It would be a shame for the rest of the students, if we didn't get to the rest of the criteria. And there was something they had had a hard time understanding the last time, so if we had not gotten to that, the whole thing would have collapsed. So we took two more minutes.

In general, teachers acknowledge the SAD since it facilitates students to take part in the assessment process, as stated by a teacher, "short and time-bound" and "characterized by clear rules and roles". Adding to this, a teacher found the restricted timeframe made the SAD easy to undertake in the relatively short teaching units.

A part of the promise of the SAD is that it can facilitate coherence between learning goals, teaching and assessment approaches. Our findings indicate that students' reflections continue beyond the duration of the SAD. A teacher wrote, "In the following lesson, my experience was that all groups (although not all students in all groups) independently set goals for their further work. It was really a pleasure to be a teacher in that lesson". The coherence between teaching and assessment could be supported by appropriate teaching activities. For example, one teacher used SADs explicitly as a reference point to other activities: "After the SAD-sessions, our students were to write a report based on the unit, and they could use the SAD there. They were motivated to use that as a shortcut to understanding how to present material". In addition, several of the teachers described how the SAD facilitates teacher and student to take along their observations and reflections into the next teaching unit, e.g. students wanted to discuss particular questions afterwards, a pair of students asked for additional feedback to revise their electrical circuit diagrams, a student made an unsolicited follow-up on how to evaluate his own scientific model used in the SAD, a teacher made a follow-up on a students' misconception, and another teacher expanded on the assessment criteria.

In general, the teachers appreciated that SAD was dialogue-based and not written. This is because it resembles existing classroom practice, reduces the work load compared to provision of written feedback and provides better prospect in understanding the basic ideas behind student's response. A teacher highlighted the value of the SAD with respect to the field of technology teaching units with a change between theory and practice and with the demands that students integrate and at the same time master both competences and content. In this context, the SAD "Seems to work very well […..] and it is possible to get a nuanced picture of the students' understanding – something that a written text would not be able to capture to the same extent".

Formative Assessment

Teachers recognised the SAD as "appropriate", "refreshing" and "motivating" and as a "valuable" alternative assessment method to their existing practice. Despite the challenges encountered related to the SAD, none of the teachers found that SAD was difficult to enact in practice. This was regardless of teachers' different degrees of alignment between the assessment approaches represented in the SAD and teachers' existing teaching and assessment practice. The fact that SAD was relatively easy to implement might be related to its flexibility. One part of the flexibility is integrated in the SAD, e.g. choice of assessment criteria and questions as well as the opportunity to use artefacts from the teaching during the SAD session. The other part of the flexibility emerges during the implementation. The Danish teachers adapted the assessment method to the context, e.g. changed the physical set-up, prolonged the timeframe, selected groups of students instead of individuals, facilitated and supplemented the peer feedback, avoided grading and only partly used the students' self-reflection tools.

In general, the teachers acknowledged the SAD as a formative assessment method as illustrated in the following quotes: "It captures the essence of formative assessment". "The main strength is the focus on formative assessment". "It's very clear to the students that it is a part of a process". This point is also reflected in the fact that in many utterances, teachers did not only describe the SAD as an assessment method but as a "teaching and assessment tool".

Combining Summative and Formative Assessment

Given that there is a final oral exam, the SAD is perceived as formative in terms of preparing for the practice of the exam. This point is illustrated in the following utterance: “I find SAD useful for formative assessment since it resembles many aspects in the final examination”.

A teacher used the learning goals to provide transparency in grading during the semi-annual student conversation.

The benefits are summarised in Table 5.4.

Challenges of the SAD as Perceived by the Danish Teachers

The challenges can be divided into six subthemes.

Handling Multiple Purposes and Different Students

As stated above, teachers perceive the SAD as having the potentials for adapting assessments to the specific student in focus through the means of preformulated assessment criteria and questions. However, it was a challenge to simultaneously fulfil the needs of the focus student and the whole class. One teacher said, “One of my challenges was that I accidentally had chosen a student who performed too well and my question didn’t challenge her at all […]. Consequently, the feedback was just flattering praises […] we didn’t even use the time allocated for feedback since none (of the students) had anything to add-on”. Adding to this point, another teacher reported that it was “lucky” that a high performance student made a mistake during the SAD, because that made it easier to put precise value on what the student did. The same teacher explained that since the focus student was too strong, it was difficult for the other students to place themselves in relation to that student. Another concern related to selecting focus students voiced by the teachers was to display weak students’ level of achievements in front of their peers. Based on this, some teacher selected a group instead of one student or decided not to enact the SAD in specific classes. The latter was mainly based of a “rough” class culture or if they only had spent limited time with the class.

Another challenge raised by teachers relates to striking an appropriate balance between knowing what the focus student is capable of and which questions to include in the dialogue and, at the same time, clarifying a realistic level of learning expectations to the rest of the class. One teacher described, “It (the task and questions) must resemble the appropriate complexity required in a teaching situation and for the final exam. It should not be too easy […]. You have to find the right student to deal with that (the complexity). But it’s not an easy task to strike the balance”.

This confirms that an important part of planning is for the teacher to choose the focus student and to tailor the questions to that student while still making realistic

Table 5.4 Benefits of SAD as perceived by the Danish teachers related to four subthemes

Benefits related to clearly stated learning goals subdivided into assessment criteria	The students get better at setting up goals for themselves
	The students know what to aim for when the criteria are explicit from the beginning
	Helps guiding questions targeted different content aspects and student achievers
	Sharing of learning goals with students encourages coherence in the upcoming teaching
	It activates students in their own and peers' learning when used for peer feedback and self-reflection
	It facilitates consistency between learning goals, teaching and assessment approaches
	It enhances students' involvement in the assessment
	It is used for clarification and documentation for parents
	It is useful for providing transparency in grading
	Explicit formulations facilitate teacher reflection and development
Benefits related to timeframe, structure and discourse	Short and delimited in time and content, and this facilitates enacting in teaching units and student willingness to participate
	It is possible to adapt to the local needs and cultures
	Clear rules and roles facilitate student to take part in the assessment process
	Teacher develops and reflects on how to make the assessment criteria explicit to student
	Possible to get a nuanced picture of students' understanding and rational through dialogue
	It facilitates coherence between teaching units and assessment
Benefits related to formative assessment	Refreshing and motivating alternative assessment method to teachers' existing practice
	Relatively easy to enact and integrate into existing teaching practice
	Captures the essence of formative assessment
Benefits in combining summative and formative assessment	The SAD is useful when it resembles aspects in the final examination
	The learning goals can provide transparency in grading during the semi-annual student conversation

assessment criteria clear to other students. This involves thoughtful preparation on the part of the teacher. Adding to this, the teacher must ensure the questioning of the focus student provides other students with sufficient information enabling them to provide sound peer feedback.

Finally, teachers encountered challenges related to activating all students. Teachers described that students' (but not all and with variation in their effort) took an active involvement in the process. However, teachers also reported SAD sessions where it was hard to activate all students throughout the session: “the listening students may have a hard time keeping up” and “It is a challenge to keep the drive-over-time in the SAD so that the feedback group is serious about their own learning (self-assessment)”.

Teachers’ Preparation and Planning.

In general teachers found it time-consuming to prepare the learning goals, assessment criteria and questions: “I think it has been time-consuming to formulate different levels of assessment criteria”. “I think there has been a lot of preparation; to sit down and really think through with assessment criteria and questions”.

Another teacher was also challenged by formulating assessment criteria. However, she expressed that collaboration made this part of the SAD easier: “I had a colleague with whom I could prepare, and that makes it so much easier, because you can discuss assessment criteria and questions. You can talk about how to write them up. It is not easy to make it clear for students what the criteria are”.

Another teacher mentioned practical challenges related to planning, “We need to refurnish the classroom when we use this assessment method”.

Timeframe and Structure

The strict 5-min timeframe on the dialogue posed a challenge to most teachers. The timeframe limits the amount and complexity of the content to be assessed. As one teacher made explicit, “It can be difficult to make as limited an assessment that it is possible to keep the timeframe”. A different teacher states, “After five minutes we were just started. I was not able to address modelling appropriately”. Another teacher relates the content limitations to a SAD session on electrical circuits, “Including more components means that the student will need more time to explain. […..]. We did not have time for calculations and practical construction”. Based on experiences from an implemented SAD, the same teacher expressed how he adjusted the next SAD to the restricted timeframe, “The more complex questions were toned down”. Another teacher was also planning to repeat the SAD but adjusted the timeframe instead of the complexity: “I probably would not obey the five minutes, but use the time that is needed on the dialogue”.

Some teachers chose a group of students to be in focus rather than just one focus student. This was done in order to resemble their current classroom practice (e.g. group work), moderate the feeling of high-stakes assessment and avoid exposing a single student. However, this made the 5-min timeframe even more challenging, as stated by a teacher, “The timeframe was also problematic since I had selected four focus students. In retrospect I should only have selected one student”.

A teacher, experienced in written peer feedback, observed a lower quality in the SAD peer feedback compared to the written one. He believes this was because the

SAD didn't allocate enough time for student to reflect on the quality in the dialogue to provide useful feedback.

Peer Feedback

In general, teachers perceived the peer feedback session as the most challenging part of the SAD due to students' inadequate "assessment literacy", such as low assessment value with respect to both feedback quantity and quality. For example, a teacher wrote, "In the peer-feedback session I missed content depth and more comments".

Teachers mainly addressed challenges related to assessment literacy with respect to students' limited content knowledge and praised the students instead of providing guidance for the next step in the learning. As one teacher stressed, "I think the biggest challenge was to get feedback students to give some real response. It was very flattering and they think it was good although I thought it was pretty bad. It may indicate that they don't have the necessary level *(*of content knowledge) to assess if something is wrong".

Another teacher elaborated on this challenge by highlighting that assessment literacy with respect to providing feedback is not only a matter of being able to judge whether a student's answer is right or wrong but also to notice where students are in their learning and where to go next:

> I think that the student and the whole class need final feedback from the teacher after the feedback session. The first time we tried the SAD, one of the best students noted that she could assess if something was right or wrong but it was difficult to know if the focus student answers the questions adequately. I think this is a good remark. Therefore, I provided a final feedback, where I made clear what could have been elaborated more.

This quote also illustrates that the teachers generally see themselves as a gatekeeper for quality. Consequently, the teachers often perceived a need to add to or facilitate the peer feedback session. This might also be a reaction to the issue illustrated in this quote "It's a challenge for some students to lead the (feedback) dialogue". Despite the challenges encountered, a teacher was planning to repeat the SAD but nuance it in the following way "Allocating different roles to the students in the feedback group, so that each student gets his/her own assignment such as "focus on definitions, units and sizes" or "focus on the use of specific content knowledge etc."

Summative Assessment

With respect to SAD as a method for summative assessment – and especially for summative self-assessment with grading teachers – this raised some concerns. As preparation for the SADs, teachers had formulated concrete assessment criteria, aimed at helping them assess student competencies in the dialogue. Even with this operationalisation of the progression steps, teachers found it challenging to judge the level of the focus student: “It is not easy to work with learning progressions and planning for giving summative assessment at the end. How can I in 5 min of dialogue and 5 min of feedback be sure that someone asks questions, which will allow me to place the student on one of the progression steps?”.

Regarding self-assessment, teachers in general, but with exception, believed that the students assess themselves on too high a level of achievement. A teacher wrote “No demands for giving feedback to students self-assessment as I doubt the function, validity and seriousness”.

Combining Summative and Formative Assessment

During the peer feedback, there was a tendency only to comment on positive aspects: “When girlfriends were feedback students, they only provide each other positive feedback”. A teacher believed that the “overgrading” in self-assessment and the insufficient feedback were part of a “performance culture”. This point is illustrated in the following quotes: “Students’ didn’t believe me when I told them that it (the SAD) was a kind of a play and that it would not influence the grading at all”. “They think I will look at the (self) grade and base my grading on it”. Another challenge highlighted by teachers is related to the SAD’s physical set-up: “There is a tendency that students may experience the SAD as an interrogation. When that is the case, the students will not see the process as being useful with a view to the future”. The performance culture and the tendency to perceive the SAD as a kind of exam might hinder the formative prospects in the SAD.

Table 5.5 displays the challenges of the SAD as perceived by the Danish teachers.

Characterising and Grouping Dialogues

The previous section elaborated on teachers’ experiences with conducting SADs. The analysis showed that teachers see both affordances and limitations with the method but does not show how the dialogues unfold. The purpose of this section is to analyse the dialogues as they play out in the classroom from a dialogical perspective (Dysthe 1996; Nystrand et al. 1997). We do this by employing network analysis as a methodological tool.

Network analysis has previously been used to analyse interviews (Bodin 2012) and student actions when learning (Shaffer et al. 2009; Lindahl et al. 2016). These are both examples where the relational nature of discussions is brought to the forefront. In the same way, a dialogue may be seen as having a relational nature. Bruun (2016) characterises network analysis in general as it may be used in science education research as a way of bringing relational aspects to the fore. In general, a network consists of a set of entities, which in network literature are called nodes, and a set of relationships, which are called links (Newman 2010).

Table 5.5 Challenges of SAD as perceived by Danish teachers related to six subthemes

Challenges related to handling multiple purposes and different students	Selecting appropriate criteria and questions to challenge the focus student
	Ensure teachers' questioning of focus student provides other students with sufficient information to enable peer feedback
	Ensure sufficient information through questioning during the focus dialogue phase for students to provide peer feedback
	Ensure that the questions, the focus student's answers and the peer feedback together reflect a realistic level of learning
	Avoid exposing weak students' level of achievements in front of their peers
	Challenging to activate all students throughout the session
Challenges related to preparation and planning	Time-consuming to identify and formulate learning goals subdivided into assessment criteria suitable to students
	Identifying and formulating assessment criteria alone without collaborative discussions
	It takes time to refurnish the classroom
Challenges related to timeframe and structure	The restricted timeframe limits the amount and complexity of the content to be assessed
	The short timeframe between the focus dialogue and the peer feedback phase does not allow students to reflect on the quality of the dialogue and provide useful feedback
Challenges associated with peer feedback	Praising instead of providing guidance on the next step in the learning
	Low assessment and feedback value with respect to both quantity and quality
	Students limited content knowledge and experience limit their "assessment literacy"
	It is a challenge for students to lead the peer feedback phase
Challenges associated with summative assessment and grading	Students (but not all) assess themselves on too high a level as part of a performance culture
	The SAD is too short to assess and grade students
Challenges associated with combining summative and formative assessment	Students' tendency only to comment on positive aspects during peer feedback as part of performance culture
	Students believe that their performance during the SAD session would influence the teacher's annual grading
	A performance culture and the tendency to perceive the SAD as a kind of exam might hinder the formative prospects in the SAD.

For the purposes of this chapter, we develop a coding scheme to capture individual speech and gesture acts. We will refer to these as dialogical acts, and dialogical acts will serve as nodes. One could then capture the frequency of particular dialogical acts and try to characterise the dialogue in this way. However, a dialogue is not only characterised by the prevalence of dialogical acts, but it is also characterised by how these acts follow from each other and how they are connected.

The main idea behind network analysis is this: By linking different dialogical acts as identified by our coding scheme, we will create networks of dialogical acts. These networks will be different in terms of the structure of connections between dialogical acts. The idea behind these networks is much like the idea behind geographical maps, to highlight important features and relationships by reducing the amount of information presented to the reader. Thus, we call these particular networks *dialogical maps.*

Networks have both a visual and a mathematical aspect (Bruun 2016), and it is possible to use these aspects to classify dialogues into groups of similar dialogues. The purpose here is to use network analysis to classify dialogue structures and then later to relate these structures to students' self-reflection and how teachers prepare.

Methodological Approach

This section presents an overview of our methodological approach. Readers who are interested in the details are referred to the ASSIST-ME website: http://assistme.ku.dk/researchers/research-design-for-the-structured-assessment-dialogue/. The approach involved the following steps:

1. Making video recordings
2. Collecting student self-reflections and teacher reflections
3. Developing coding categories and coding video recordings
4. Converting codes to dialogue maps
5. Selecting and using a measure for comparing dialogue maps
6. Comparing dialogue maps and finding groups of related maps
7. Relating groups of maps to student self-reflections, teacher reflections and contextual data

Making Video Recordings In this step we made sure that recordings had both teacher and focus student clearly visible in the frame and that both student and focus student could be clearly heard on the audio track.

Collecting Student Self-Reflections and Teacher Reflections The last 2 min of the ritual in which the dialogues were embedded involved students reflecting on how the dialogue and peer feedback had helped students. Following the model for formative assessment presented in Chap. 3 (Fig. 3.1), we prepared five quantitative questions for students to answer. The questions were (in parenthesis how we have abbreviated the answers to these questions in our analysis section):

1. How many of the teacher's questions was I able to answer? (QA)
2. How well did the dialogue help me determine my own level? (Det.Dia)
3. How well did the peer feedback help me determine my own level? (Det.FB)
4. How certain am I about the next steps in my learning? (Next.Step)
5. What grade would I award myself? (Grade)

Students' answers to questions 1–4 were recorded on a 7-point Likert scale.

To help teachers prepare for SADs and to document this preparation in a systematic manner, we developed teacher preparation forms. The forms had two pages, a pre-dialogue preparation page and a post-dialogue reflection page. The pre-dialogue preparation page prompted the teacher for contextual data (date, class, subject) and qualitative reflections about the competencies, topic, questions and student. The post-dialogue reflection page included 4-point Likert scale questions accompanied with prompts for comments. The questions were:

1. Did you manage to get through all of your questions?
2. Did the student and the rest of the class get a good sense of the criteria you had for the given competency?
3. In your judgement did the student receive relevant formative feedback during the ritual?

Finally, teachers were prompted to grade student performance and to comment on that grade.

Developing Coding Categories and Coding Video Recordings We developed categories to describe the dialogue, to coarsely describe gesture and to describe the criteria for the dialogue. The criteria were dependent on the dialogue, and gesture was coded as none, pointing or other. We do not discuss these further in this chapter. The dialogical codes were based on an operationalisation of the Dysthe (1996) dialogical framework. The codes were developed iteratively; coders applied a set of preliminary codes on one Danish dialogue, the codes were discussed, changed and reapplied until a consensus on a particular set of codes and their use was reached. Then two coders split the remaining Danish dialogues between them, while a third coder coded the Finnish dialogues using the codes. Table 5.6 shows the final dialogical codes.

Coders first watched the entire dialogue and then divided the dialogue into strict 5-seconds intervals. Each interval was coded with a set of codes to describe who were active (teacher, student or both) in the dialogical action, any gesture and the addressed criteria. For each dialogue, this yielded a series of codes such as

1:30T_TUptake_NoTG_TCriterion1a_S_SLowerorderstatements_SGesture_SCriterionA.
1:35 NT_NTD_NoTG_NTC_S_SLowerorderstatements_SGesture_SCriterionA.

This code shows that in the time interval from 1:30 to 1:35, the teacher was active, was engaged in an uptake dialogical action, did not use gesture and addressed a specific criterion, *Criterion A*. At the same time, the student was also active, producing a lower-order statement, gesturing and also addressing *Criterion A*. In the interval from 1:35 to 1:40, the teacher is silent and not gesturing, while the student continues the statement addressing *Criterion A* while gesturing. To make the judgement of when different dialogical actions begin and end, it was necessary for coders to review the dialogue as a whole before commencing with the detailed 5-sec interval coding.

Table 5.6 Codes for describing dialogical aspects of the student-teacher dialogue

Code	Abbreviation	Description
Invitation	Inv	Broad invitations from the teacher for the student to say something, often to open the dialogue, e.g. "Could you tell me something about the experiment you did?"
Uptake	Upt	"[I]ncorporating students' responses into the next question, thus getting the students to reflect further about what they said, and integrating the answer into the dialogue[...]"
Focus	Foc	Meant as an opposite to uptake. Focus can be seen as an emphasis on the set teaching goals, where uptake can go out on a tangent.
Precise valuing	PreV	Analogous to high valuing (Dysthe 1996) but might not be strictly positive. The point is that it is precise and puts value to what is said.
Precise correction	PreC	A possible counterpart to precise valuing – but this code is for an explicit correction
General evaluation	GE	General evaluation. Mean as a possible counterpart to precise valuing – but this code is for general praise/ criticism
Higher-order question	HoQ	Questions that aim at the higher levels of Bloom's taxonomy (application, analysis, synthesis and evaluation)
Lower-order question	LoQ	Questions that aim at the lower levels of Bloom's taxonomy (knowledge and comprehension)
Summarising	Summ	Meant for instances when the teacher repeats or sums up what was said or done by the student without evaluating or correcting this.
Higher-order answer	HoA	Analogous to the higher-order question
Lower-order answer	LoA	Analogous to the lower-order question.
Higher-order statement	HoS	Analogous to the higher-order answer, but initiated by the student
Lower-order statement	LoS	Analogous to the lower-order answer, but initiated by the student
Student question/ non-understanding	QNU	Used when the student explicitly asks the teacher to repeat a question and says that he or she is unsure or unable to answer

Converting Codes to Dialogue Maps In order to convert the codes to networks (*dialogical maps*), each dialogue was represented as a timeline of codes that depicted how the dialogue progressed in time in terms of the dialogical framework. From this timeline, drawing an arrow from code A to code B, if B followed A in the timeline, could create *dialogue maps*. Thus, the nodes in dialogue maps represent codes as given above. Figure 5.2 shows two examples of dialogue maps. The maps seem to have very different structures. For example, the map on the left seems simpler and more linear than the one on the right. In this representation, we have used hatching to represent different aspects of the codes. For example, nodes that represent student speech actions are white, while nodes representing teacher actions are shaded. Nodes that represent speech actions of both student and teacher are black. The sizes

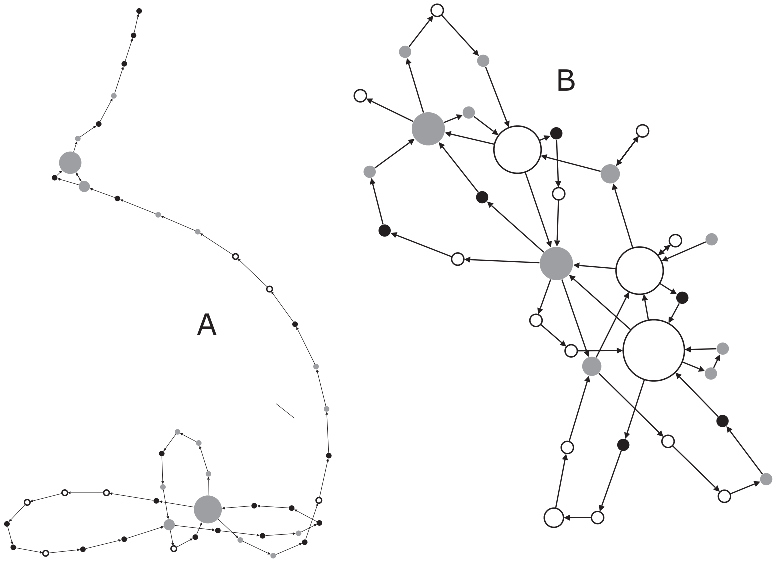


Fig. 5.2 Two examples of dialogue maps. Node sizes are not normalised between maps but can be used as a first visualisation of differences in structures and distribution of node sizes. The two maps, A and B, are qualitatively different: the dominant nodes (*shaded*) in A each have few incoming connections, while the dominant nodes in B have three or more. Also B connects a large number of nodes, whereas A is more string-like. Finally, two large teacher action nodes (*shaded*) dominate A, while three student action nodes (*white*) dominate B

of nodes represent the time spent on that particular speech action. Comparing these two maps, it is clear that the teacher is most prominently represented in the map on the left, while the map on the right shows a more equal distribution of speech acts. The next paragraph describes how these apparent differences can be quantified systematically across all dialogical maps by using the PageRank algorithm (Brin and Page 1998) for quantifying node importance.

Selecting and Using a Measure for Finding Groups of Related Dialogue Maps Network analysis offers many measures for characterising networks. These include structural measures and centrality measures (Costa et al. 2007). Centrality measures are employed to gauge the importance of single nodes in a network. Bruun and Brewe (2013) distinguished between local measures and global measures. Global measures incorporate network structure, the structure of dialogical maps in this study. This study uses PageRank (Brin and Page 1998) as a measure of prevalence. The PageRank is usually described from the perspective of a random walker on the network. If such a walker traverses the network for a long period of time, it will visit each node a certain fraction of the time. This fraction of time is dependent on the structure of the network. To characterise a dialogue map, PageRank was added for each code and each node throughout the network. For example, PageRank for all nodes that signify that only the teacher is active are added to gauge the prevalence of the teacher in each dialogue, likewise for when the student is active, when both are active and when neither is active. The procedure continues through all codes, except for criterion codes. Since criteria vary between dialogues, they are not suited for a comparison, although they still provide structure to each dialogue map. The result for each dialogue map is the accumulated PageRank (in per cent) for each code.

Comparing Dialogue Maps and Finding Groups of Related Maps The strategy from here was to use the characterisation found above as a basis for similarity between dialogue maps. If they shared characteristics, they were seen as similar. Based on that similarity, clusters of similar maps could be identified using community detection (see Fortunato (2010) for an extensive overview of community detection). Since each code was a variable, there were a large number of variables, which might result in typologies with meaningless groups. Inspired by Yeung and Ruzzo (2001), we used principal component analysis (PCA) as a way of reducing the number of variables in clustering analyses. We used the psych package (Revelle 2015) in R (R Core Team 2015) to perform the PCA. Each dialogue map was then represented as a point in a Euclidean space spanned by the principal components. This allows for the distance to be calculated as the square root of the sum of squared differences in coordinates between coordinates. A high distance signifies low similarity, while a distance of zero would signify exact similarity in the principal component space. To convert distance into similarity, we used a simple exponential function with negative exponent. For distances, which are positive or zero, this yields similarities between zero and one. The result of this procedure is a matrix of similarities between each dialogue.

Starting from the distance matrix, clustering was based on the method of Brewe, Bruun, and Bearden (2016) yielding (1) visual and mathematical representations of significant connections between each pair of dialogues and (2) a network of dialogue maps with an associate group structure that was determined empirically from the data. This method makes use of the Infomap algorithm (Rosvall and Bergstrom 2008) on a version of the similarity network where we have removed weak connections (see Brewe et al. (2016) for a thorough description of the procedure).

Relating Groups of Maps to Student Self-Reflections, Teacher Preparation and Contextual Data This final step is needed to put the dialogical maps back into the context from which they were derived. We make use of students' quantitative Likert scale self-reflections and of the categorical contextual data. We used t-test, ANOVAs and Z-values of the tendency of categories to be over-represented in a group (Bruun and Bearden 2014) as appropriate. The details of our analysis are given at the ASSIST-ME website: http://assistme.ku.dk/researchers/research-design-for-the-structured-assessment-dialogue/.

Findings

We first present the groups of similar dialogical maps along with a description of the identified principal components (PC). Then we relate the similarity groups to student self-reflections, teacher reflections and contextual data. Finally, we describe a typology, which draws on these results.

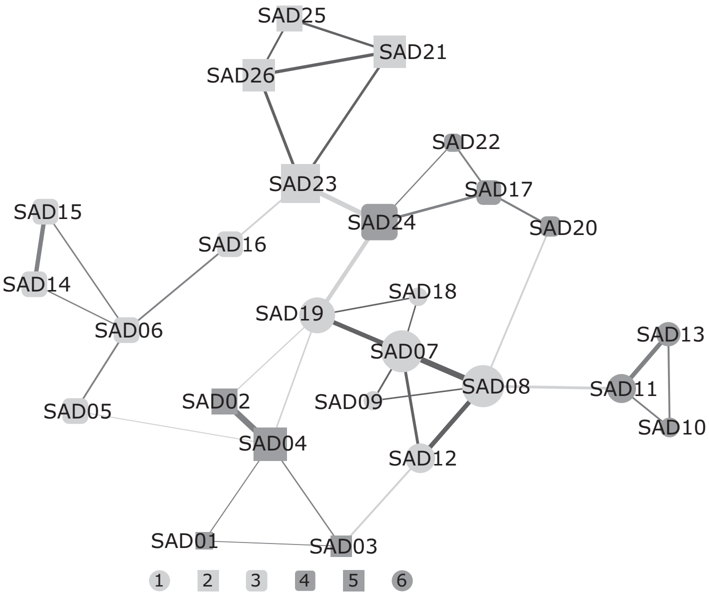


Fig. 5.3 Similarity networks based on similarities of code PageRanks. Visualisation was made with software Gephi (Bastian et al. 2009) using the Force Atlas layout algorithm. *Hatching and shape* represent groups, and *node* sizes represent weighted degree

Table 5.7 PageRank-based grouping and group scores on each principal component (PC). PC numbers are group means, while the numbers in parentheses are standard deviations

	Group 1	Group 2	Group 3	Group 4	Group 5	Group 6
SAD no	7 8 9 12 18 19	21 23 25 26	5 6 14 15 16	17 20 22 24	1 2 3 4	10 11 13
N	6	4	5	4	4	3
PC 1	0.5 (0.7)	−0.7 (0.6)	−1.3 (0.4)	−0.1 (0.4)	1.4 (0.2)	0.4 (0.6)
PC 2	−0.1 (0.7)	−0.2 (0.6)	−0.4 (1.0)	1.7 (0.8)	−0.4 (0.4)	−0.6 (0.2)
PC 3	−0.6 (0.5)	−0.8 (0.5)	0.8 (0.4)	0.2 (0.5)	1.3 (1.0)	−1.0 (0.5)
PC 4	0.7 (0.5)	−0.5 (0.6)	0.3 (1.5)	0.1 (0.7)	−0.4 (1.0)	−1.0 (0.7)
PC 5	−0.3 (1)	1.2 (0.5)	−0.1 (0.4)	−0.1 (0.8)	0.5 (0.8)	−1.5 (0.5)

Groups of Dialogue Maps and Description of Principal Components Figure 5.3 shows visual representations of the PageRank-based similarity network. Each node (circle) represents a SAD while links (lines) represent similarity links. Table 5.7 gives an overview of the PageRank-based grouping. Based on the mean scores and their standard deviation, each component seems to separate particular groups. For example, PC 1 seems to separate Groups 2, 3 and 4 (negative mean scores) from 1, 5 and 6 (positive mean scores). The principal component analysis yielded five components, and we describe them below along with group differences found by t-tests and ANOVAs. We have left the detailed results of t-tests and ANOVAs at the ASSIST-ME website: http://assistme.ku.dk/researchers/research-design-for-the-structured-assessment-dialogue/.

PC 1: No Gesture and Confusion vs. Pointing and Lower-Order Answers This factor is dominated by gestural codes: Prevalence of the no gesture code will count towards a high score, while pointing will count towards lower scores. Also, student not understanding/questioning will count towards a high score, while lower-order statements will count towards a low score.

Groups 1, 5 and 6 seem to be characterised by teachers and students not gesturing, some degree of student confusion about teacher questions, few lower-order statements and, to some degree, silence. On the other hand, we can expect the SADs of Groups 2, 3 and 4 to on average to display both teachers and students pointing more and for students to utter more lower-order statements.

PC 2: Active Student and Higher-Order Statements This component is characterised by the absence of the *teacher* and an active student. Moreover, student actions such as *other gestures* and *higher-order statements* count towards a high score. High scorers in this type of dialogue must have an active and probably quite autonomous student, which is guided minimally by the teacher's precise corrections.

The SADs in Group 4 seem to be characterised by a more autonomous student than the other groups. And a teacher, who takes the background and lets the student speak, accompanies this. This in turn seems to give the student room to make higher-order statements using gesture and manipulating artefacts. An example of the latter is SAD 17, where the student actively builds a circuit as part of the dialogue.

PC 3: Few Question and Answers, Teacher Talking and Silence This component is characterised by a lack of *lower-order questions* and *answers*. *No student dialogue* and *no-action* both count towards a high score. It seems that high scores on this component signify that the student is less active. The codes *uptake* and *summarising* both count towards high scores on the component, which may signify a teacher trying to help the student to get started.

There might not be much of a dialogue in Groups 3 and 5, whereas for Groups 1, 2 and 6, the teacher asks lower-order questions and gets lower-order answers.

PC 4: Active Student, Teacher Gives Space This component is characterised by an active student and a teacher, who does not speak when the student speaks. Thus, the *both student and teacher active* code counts towards a low score, while the *active student* and *teacher no dialogue* codes count towards a positive score. Groups 1, 3 and 4 tend two have more active students and teachers that provide space, while the teacher would be more active in Groups 2, 5 and 6.

PC 5: Higher-Order Elements in Dialogue This component is characterised by *higher-order questions* and *higher-order answers* counting towards a positive score, while *invitation* counts towards a lower score. Thus, the teacher spends less time inviting students to speak in dialogues that score relatively high in this dimension, while the focus is on higher-order dialogue. Group 2 might be seen as an example of high scoring, while Group 6 may be seen as an example of low scoring.

Relation to Student Self-Reflections The different groups of dialogical maps were associated with significant differences in student self-reflections. Again the details can be found in the Online Appendix. Group 3 scored lower on how many questions students thought they were able to answer (QA) than all other groups.

Groups 3–5 scored lower than the other groups on students' experiences of determining their own level from the dialogue. The same two groups scored significantly lower than the other groups with regard to students' self-reported ability to use the SAD to determine their own next steps.

Finally, students in Groups 3–6 self-graded themselves significantly lower than students in the rest of the groups.

Relation to Teacher Reflections We tested both groupings against various elements related to the teacher and the context. Using ANOVA tests as before, we tested for group differences on duration of dialogue, number of criteria addressed and on the teacher's perception of the quality of the dialogue. None of the tests yield significant differences at the 0.05 level according to the Tukey HSDs.

Relation to Contextual Data We tested four group differences in subject, ASSIST-ME area, ASSIST-ME competencies and educational level, whether the teacher form indicated that the focus student had been chosen beforehand or not, country, class ID and teacher ID. We tested this using the segregation measure developed by Bruun and Bearden (2014). The segregation measures the tendency for particular attributes to be over-represented in a group.

We found significant segregation for the PageRank-based grouping with regard to teacher ID; teachers have a tendency to conduct SADs that are in one of these groups. For example, Group 1 consists of six SADs. Teachers DK-1 and DK-6 both conducted two SADs, and they are located in Group 1. The dialogues of DK-4 and FI-2 are located in Group 3, while Groups 5 and 6 consist of FI-1 and DK-3, respectively. DK-7 and DK-8 are the only teachers, which have both conducted more than two dialogues and are seen in two groups (they are both found in Groups 2 and 4).

We also found other significant segregation patterns. However, these could mostly be explained by the teacher segregation. For example, Group 1 seems segregated according to subject, biotechnology. However, on closer scrutiny, this coincides with the fact that the only teacher who taught biotechnology is in this group. One possible exception to this is Group 2, which does not show significant segregation with regard to teacher, but according to ASSIST-ME competencies, all four SADs in this were intended to focus on argumentation, which suggests that SADs about argumentation may potentially be different than other dialogues. However, given the small numbers in each group, even small deviations might change the results drastically.

Typology We answer research question 2 by providing a typology. We do this in Table 5.8 below by summing up the interpretations and findings given so far for each group. The typology answers the research question by (1) providing groups of dialogical maps, which are different based on our network analytical approach and (2) finding significant differences between these groups in terms of contextual information and student self-reflections.

Table 5.8 The typology of structured assessment dialogues

Group 1: Simple talk SAD: 7 8 9 12 18 19	Group 2: Teacher driven SAD: 21 23 25 26	Group 3: Difficult content SAD: 5 6 14 15 16
Few gestures (some pointing), some student confusion, lower-order questions and answer The student is active, and the teacher gives space Enacted by four different teachers in both upper and lower secondary science	Pointing and lower-order statements, lower-order questions and answers Also, higher-order questions and answers, and a tendency for the teacher drives the dialogue Perhaps driven by a focus on argumentation	Pointing and lower-order statements, few lower-order questions and answers, teacher summarises and tries to use student utterances. Active student and teacher give space Reflecting students feel that they could not answer many of the questions and that it was difficult to use the dialogue to determine own level and next steps in learning. Also, student self-grades were low
Group 4: Competent student SAD 17 20 22 24	Group 5: The struggle SAD: 1 2 3 4	Group 6: Trying to invite SAD 10 11 13
Pointing and lower-order statements but also an autonomous student that acts at a high level. Active student with the teacher giving space Enacted by three different teachers in upper secondary school in science, technology and mathematics	No gesture, some student confusion, few lower-order questions and answers, teacher summarises and tries to use student utterances. Inactive student, teacher does not give space Reflecting students feel that they could not answer many of the questions and that it was difficult to use the dialogue to determine own level and next steps in learning	No gesture, some student confusion, lower-order questions and answer Lack of higher-order questions and answers, but a teacher that tries to invite students. The teacher is active, and the student does not take or is not given much space Student self-grades were low

Conclusion

In this study we have implemented a novel method for analysing dialogues and applied it to a particular method of assessment: the structured assessment dialogue (SAD), with the ability to productively combining formative and summative use of assessment. The method of assessment was tested with 10 different teachers resulting in 26 different SADs in two countries, Finland and Denmark. We have analysed these dialogues by coding them using a dialogical framework, converting the codes to dialogical maps, finding similarities between this maps using the PageRank centrality measure and finding groups based on these similarities. The groups showed significant differences in terms of student self-reflections and contextual data but not on teacher reflections. The network analysis resulted in a typology of dialogues, consisting of six types: *Simple talk*, *teacher driven*, *difficult content*, *competent student*, *the struggle* and *trying to invite*. Each type has different characteristics based on relationships to student self-reflections and to contextual data in particular.

Overall Conclusions

In this study we have implemented and analysed a novel method for assessment: The Structured Assessment Dialogue (SAD). The method was tested with 10 different teachers resulting in 26 different SAD enactments in Denmark and Finland.

We have analysed these implementations qualitatively, and the findings showed that teachers found the method challenging but potentially useful to implement in their own practice and with ability to productively combine the formative and summative use of assessment. The short timeframe made it especially tractable for implementation in teacher practice but also constrains the dialogue in terms of coverage. To be successful the method requires teacher planning in terms of, e.g. defining assessment criteria and particular questions to ask.

In order to be able to link characteristics of a dialogue with the effect of the dialogue, we analysed these dialogues by coding them using a dialogical framework. The codes were converted to dialogical maps, and we found similarities between these maps using the PageRank centrality measure and were able to group the maps based on these similarities. The groups showed significant differences in terms of student self-reflections and contextual data but not on teacher reflections. The network analysis resulted in a typology of dialogues, consisting of six types: simple talk, teacher driven, difficult content, competent student, the struggle and trying to invite. Each type has different characteristics based on relationships to student self-reflections and to contextual data in particular.

In sum, we find that the SAD is a promising method for combining formative and summative use of assessment, easily adaptable to local educational cultures. We also find the dialogical mapping a potential useful method for analysing various effects of dialogues.

Appendix

All technical results and analyses can be found at the ASSIST-ME project's website: http://assistme.ku.dk/researchers/research-design-for-the-structured-assessment-dialogue/

References

Alexander, R. (2006). *Towards dialogic thinking: Rethinking classroom talk*. York: Dialog.

Bakhtin, M. M. (1981). *The dialogic imagination*. Austin: University of Texas Press.

Bakhtin, M. M. (1986). The Problem of Speech Genres (Vern W McGee, övers.) I Caryl Emerson & Michael Holquist (Red.), Speech Genres & Other Late Essays (ss. 60–102).

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education, 18*(1), 5–26.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*, 5.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2004). Working inside the Black box: Assessment for learning in the classroom. (Cover story). *Phi Delta Kappan, 86*(1), 9–21.

Bodin, M. (2012). Mapping university students' epistemic framing of computational physics using network analysis. *Physical Review Special Topics-Physics Education Research, 8*(1), 010115.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101.

Brewe, E., Bruun, J., & Bearden, I. G. (2016). Using module analysis for multiple choice responses: A new method applied to Force concept inventory data. *Physical Review Physics Education Research, 12*(2), 020131.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems, 30*(1), 107–117.

Bruun, J. (2016). Networks as integrated in research methodologies in PER. In D. Jones, L. Ding, & A. Traxler (Eds.), 2016 PERC Proceedings. (pp. 11–17). Chapter Plenary Manuscripts. [2] Sacramentro, CA: American Association of Physics Teachers. (PERC Proceedings). DOI:10.1119/perc.2016.plenary.002.

Bruun, J., & Bearden, I. G. (2014). Time development in the early history of social networks: Link stabilization, group dynamics, and segregation. *PloS One, 9*(11), e112775. doi:10.1371/journal.pone.0112775.

Bruun, J., & Brewe, E. (2013). Talking and learning physics: Predicting future grades from network measures and Force concept inventory pretest scores. *Physical Review Special Topics-Physics Education Research, 9*(2), 020109.

Butler, R. (1987). Task-involving and ego-involving properties of evaluation: Effects of different feedback conditions on motivational perceptions, interest, and performance. *Journal of Educational Psychology, 79*(4), 474.

Cho, K., & Schunn, C. D. (2007). Scaffolded writing and rewriting in the discipline: A web-based reciprocal peer review system. *Computers & Education, 48*(3), 409–426.

Christensen, T. S. (2005). *Integreret evaluering: en undersøgelse af den fagligt evaluerende lærer-elevsamtale.* Syddansk Universitet. Det Humanistiske Fakultet. 328 s. (doctoral dissertation, in Danish). Retrieved from http://findresearcher.sdu.dk:8080/portal/da/person/tosc

Costa, L. D. F., Rodrigues, F. A., Travieso, G., & Villas Boas, P. R. (2007). Characterization of complex networks: A survey of measurements. *Advances in Physics, 56*(1), 167–242.

Dysthe, O. (1996). The Multivoiced classroom interactions of writing and classroom discourse. *Written Communication, 13*(3), 385–425.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports, 486*(3), 75–174.

Harrison, C. (2006). Banishing the quiet classroom. *Education Review, 19*(2), 67–77.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Krajcik, J., McNeill, K. L., & Reiser, B. J. (2008). Learning-goals-driven design model: Developing curriculum materials that align with national standards and incorporate project-based pedagogy. *Science Education, 92*, 1–32.

Kvale, S. (2007). *Doing interviews* (Book 2 of The SAGE Qualitative Research Kit).

Leach, J., & Scott, P. (2002). Designing and evaluating science teaching sequences: An approach drawing upon the concept of learning demand and a social constructivist perspective on learning. *Studies in Science Education, 38*(1), 115–142.

Lemke, J. L. (1990). *Talking science: Language, learning and values*. Norwood: Ablex.

Lindahl, M., Bruun, J., & Linder, C. (2016). *Integrating text-mining, network analysis and thematic discourse analysis to produce maps of student discussions about sustainability*. In *PERC 2016*.

Newman, M. (2010). *Networks: An introduction* (pp. 1–2). Oxford: Oxford University Press.

Nystrand, M., Gamoran, A., Kachur, R., & Prendergast, C. (1997). *Opening dialogue: Understanding the dynamics of language and learning in the English classroom*. New York: Teachers College Press.

Ogborn, J., Kress, G., Martins, I., & McGillicuddy, K. (1996). *Explaining science in the classroom*. Buckingham: Open University Press.

R Core Team. (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/

Revelle, W. (2015). *Psych: Procedures for personality and psychological research*. Evanston: Northwestern University. http://CRAN.R-project.org/package=psych Version = 1.5.8.

Rosvall, M. & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences of the United States of America (PNAS)*, *105*(4), 1118–1123. Retrieved April 1, 2017, from http://www.pnas.org/content/105/4/1118.short

Roth, W. M. (2000). From gesture to scientific language. *Journal of Pragmatics, 32*(11), 1683–1714.

Roth, W. M., & Lawless, D. V. (2002). How does the body get into the mind? *Human Studies, 25*(3), 333–358.

Ruiz-Primo, M. A. (2011). Informal formative assessment: The role of instructional dialogues in assessing students' learning. *Studies in Educational Evaluation, 37*(1), 15–24.

Scott, P. H., Mortimer, E. F., & Aguiar, O. G. (2006). The tension between authoritative and dialogic discourse: A fundamental characteristic of meaning making interactions in high school science lessons. *Science Education, 90*, 605–631.

Shaffer, D. W., Hatfield, D., Svarovsky, G. N., Nash, P., Nulty, A., Bagley, E., Frank, K., Rupp, A. A., & Mislevy, R. (2009). Epistemic network analysis: A prototype for 21st-century assessment of learning. *International Journal of Learning and Media, 1*(2), 33–53. doi:10.1162/ijlm.2009.0013.

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153–189.

Wiliam, D. (2011). *Embedded formative assessment, study guide*. Bloomington: Solution Tree Press.

Wiliam, D., & Leahy, S. (2015). *Embedding formative assessment: Practical techniques for F-12 classrooms*. West Palm Beach: Hawker Brownlow Education.

Yeung, K. Y., & Ruzzo, W. L. (2001). Principal component analysis for clustering gene expression data. *Bioinformatics, 17*(9), 763–774.

Chapter 6
Students' Perspectives on Peer Assessment

Florence Le Hebel, Costas P. Constantinou, Alena Hospesova, Regula Grob, Monika Holmeier, Pascale Montpied, Marianne Moulin, Jan Petr, Lukáš Rokos, Iva Stuchlíková, Andrée Tiberghien, Olia Tsivitanidou, and Iva Žlábková

Introduction

The role of feedback on student performance is central in formative assessment (Black and Williams 1998) and, obviously, in peer assessment. Peer feedback is expected to support the learning process by providing an intermediate check of the performance according to the criteria and adapted to the individual student, accompanied by comments on strengths, weaknesses, and/or tips for improvement (Falchikov 1996). Learning benefits may arise for students while enacting both the role of peer assessor and peer assessee. Thus, peer assessment can be perceived as a learning tool, since assessing their peers can develop students' judgment-making skills about what constitutes high-quality work and a self-reflection about their own understanding (Topping 2013).

This chapter reports the results of three research studies on peer assessment made in different countries where such practice is unfrequently implemented in as

F. Le Hebel (✉)
Université de Lyon. Laboratoire ICAR, UMR 5191, LLE UMS 3773, Ecole Normale Supérieure de Lyon, Lyon, France
e-mail: florence.le-hebel@ens-lyon.fr

C.P. Constantinou • O. Tsivitanidou
Department of Education, University of Cyprus, Nicosia, Cyprus

A. Hospesova • J. Petr • L. Rokos • I. Stuchlíková • I. Žlábková
Jihočeská univerzita, University of South Bohemia, České Budějovice, Czech Republic

R. Grob • M. Holmeier
School of Education, University of Applied Sciences and Arts Northwestern Switzerland, Basel, Switzerland

P. Montpied • M. Moulin • A. Tiberghien
CNRS, ICAR, UMR 5191, Ecole Normale Supérieure de Lyon, Lyon, France

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_6

a steady classroom organization (France, Switzerland, and the Czech Republic). The three countries are all participating in the EU-funded project ASSIST-ME aiming to develop formative assessment (http://assistme.ku.dk/). The three research studies focus on different competences and different disciplines, but they all involve inquiry-based approaches at primary and secondary school level. The research design is quite similar in the three studies, involving students in a class situation with a teacher who volunteered to participate along with his/her class in the practical implementations. Peer assessment is included in the inquiry-based teaching sequence and students are guaranteed that the research experience would not contribute to their final mark at the end of the semester. In the first step, the students work on tasks individually or in pairs (randomly matched), then their written artefacts are given to another pair (or individual), and they are asked to use rubrics with prespecified criteria for providing feedback to their peers. Once the students complete the peer assessment, they exchange their peer feedback and review it in pairs or individually. Students are allowed to use the peer feedback for revising their artefacts. The students were assured that the study would not contribute to their final mark.

However, as the research questions differ, the data analysis in the three studies varies. In the French study, the data reported in this chapter explore relationships between the success in task processing and the ability to mark a peer's written artefact about the same task. It corresponds in a part of a study investigating to what extent peer assessment helps students to develop understanding and competences involved in the teaching sequence on science. Peer feedback implementation is conducted in physics and geosciences at upper secondary level and focuses on students' investigational competence. In Switzerland, peer feedback is implemented in physics at upper secondary level focusing on modeling competence. Based on a fine-grained analysis of peer feedback comments, the research study examines the type of peer feedback students offer to their peers while assessing their models. In the Czech Republic, the study focuses on students' reflection on peer assessment in inquiry lessons. Peer assessment is conducted in mathematics and biology, at primary and lower secondary level, focusing on problem-solving and investigational competences and is followed by semi-structured interviews with the students. The results of these three studies show some convergent and divergent points of views, and some perspectives on implementing peer feedback as part of formative assessment raised are discussed in the last part of this chapter.

Theoretical Background

The main aim of formative assessment is seen as helping learning. As developed in Chap. 3, the concept of formative assessment is defined differently according to two main conflicting views (Bennett 2011). Some authors consider formative assessment as referring to an instrument and others conceive it as a process. In this project, formative assessment is conceived between these two views. In this chapter, we focus on the process view involving students and teachers (see Fig. 3.1 in Chap. 3, based on Harlen 2013 and modified by Dolin et al.). In particular, we focus on the collection and interpretation of evidences in terms of what it indicates about existing ideas and competences required to meet the lesson goals. In our work, feedback is from student to student. This is consistent with the social constructivist view of learning, which emphasizes the role of interaction in students' understanding construction (see Chap. 3).

Peer assessment is generally an educational arrangement for classmates to judge the level, value, or worth of the products or learning outcomes of their equal-status peers by offering written and/or oral feedback (Topping 2013). The students' picture of themselves or of their "equal-status peers" is one of the crucial points in peer assessment since most often students do not feel fully confident in their own or their peers' knowledge as they are not expert in a subject area. They doubt their peers' ability to assess (Hanrahan and Isaacs 2001; Strijbos et al. 2010; van Gennip et al. 2009; Walker 2001). The peer assessor is usually not regarded as a knowledge authority by an assessee. Students frequently claim that it is the role of the teacher to be the assessor (Brown et al. 2009). Students lack confidence in both their and their peers'abilities as assessors (Ballantyne et al. 2002). Nevertheless, even if the accuracy of peer feedback can vary as students are not experts in the subject area, it can be helpful for learning while students peer assess and when they review their received peer feedback, as they engage in self-reflection processes. In their study, Yang et al. (2006) show that revision following the teacher's feedback is less beneficial for students' understanding than peers' feedback. The authors argue that teacher's feedback was accepted as such, often misinterpreted, and students often considered that no further corrections were expected. whereas peer feedback leads to more discussions and checking for confirmation and consequently a deeper understanding. Moreover, for Strijbos et al. (2010), the qualification "equal-status students" in Topping's definition (1998) might be retained in the sense of class level of students, but there are individual differences that affect perceived status and may impact peer feedback perceptions. Through these research studies, it appears that a student's representation of knowledge authority plays a central role in the dynamics of peer assessment.

Another crucial aspect of peer assessment is the quality of peer feedback. Peer feedback may be delivered either as qualitative (oral or written comments), quantitative (mark), or both. There are various perspectives on peer feedback quality (see Gielen et al. 2010 for an overview). A first perspective defines peer feedback quality as the degree to which a peer's quantitative feedback (mark) matches that assigned by an expert assessor, where scoring validity is the leading concept (Cho et al. 2006; Falchikov and Goldfinch 2000; Van Steendam et al. 2010). A second perspective defines peer feedback quality in terms of content and/or style characteristics. Written comments on a specific piece of work/artefact could vary among peer assessors, because they might focus on different aspects of an assessee's work (Topping 1998). This variability makes the determination of the quality of written comments through specific measures/indices, such as a reliability index, extremely difficult or even impossible in some cases. Some studies try to build a framework for assessing the quality of feedback (Gielen et al. 2010; Prins et al. 2006; Sluijsmans et al. 2002). Different characteristics (summed up in Gielen et al. 2010; Table 1, p. 307) are identified as, for instance, the extent of the peer feedback comments (if the comments are elaborate or superficial) or the justification of peer feedback comments (whether the reasoning underlying the assessors' judgments is revealed to assessees or not).

In this chapter, the peer feedback quality (quantitative and qualitative) is analyzed in the French and the Swiss studies, whereas the Czech study focuses on the students' peer feedback perception.

Research Aims

Different main research questions are addressed:

The French study investigates if a student can actually assess a classmate's work prior to receiving any formal teacher feedback on his own artefact and therefore being in the process of learning from the task through possible exchange with the teacher.

The Swiss study examines the type of peer feedback students generate for their peers while assessing their models in a physics course at upper secondary level.

The Czech research relates to the students' reflection on peer assessment, and more specifically they investigate how students perceive peer feedback they offer and they receive in the context of inquiry lessons.

Relations Between Peer Assessment and Students' Artefact (France)

In France, traditional summative assessment is emphasized; therefore, peer assessment in elementary and secondary school is not a usual practice in most classrooms. Teachers are usually the only ones that provide feedback. However, some French universities initiate and develop students' peer assessment at university (Le Monde 2016).

French official instructions encourage inquiry-based teaching. But this implies to be aware of scientists and students functioning in inquiry. According to Etkina et al. (2010), scientific abilities include but are not limited to collecting and analyzing data from experiments; devising hypotheses and explanations and building theories, assessing, testing, and validating hypotheses and theories, and using specialized ways of representing phenomena and of communicating ideas (Duschl et al. 2007). For scientists, scientific abilities are internalized and become habits of mind to approach new problems. For the students who have not internalized these processes and procedures, scientific abilities are processed that they need to use reflectively and critically (Etkina et al. 2010). Teaching has the necessity to explain and verbalize with students processes and procedures which are internalized by scientists. Inquiry-based activities should allow to make some students' cognitive process visible, not only to the student but also to the whole class in order for them to be a shared knowledge (Tiberghien 2011). Moreover, it leads the teacher to a better understanding of the students and a better adaptation of his/her teaching. In our study, we focus on these three main scientific procedures (from Etkina et al. 2010) that we relate to assessment criteria:

- Make hypotheses/speculations and explanations.
- Collect and analyze data based on experimentations.
- Assess, test, and validate hypotheses.

Peer assessment could help to make these procedures visible to students.

Objectives of the Study

This study aims at investigating to what extent peer assessment helps students to develop understanding and competences involved in the teaching sequence on science. To that end, we use students' written artefacts and videos in the classroom during the activities and peer assessment.

A part of the study, presented in this chapter, investigates if a student can assess a classmate's work successfully even before receiving any formal teacher feedback on his own artefact. The aim here is to explore the relationships between the success in task processing and the ability to assess quantitatively (with a mark) a peer's written artefact about the same task.

More specifically, we worked on two research questions:

1. To what extent are students able to give a mark to each question of their peers' work that is consistent with the mark that an "expert" (teacher/researcher) would have given?
2. To what extent is high achievement of student assessor in the assessed task a mandatory condition to proper marking (consistent with an expert's mark)?

Methodology

We engaged secondary level students (grade 10, 15–16 years old) in a teaching sequence for experimental science. In physics, the sequence was about periodic phenomena (like a heartbeat or beats of the membrane of a loudspeaker, a pendulum, the Earth's movement around the Sun, etc.). In geology, the sequence was about fossil fuels (Table 6.1). In both teaching sequences, two activities were combined with peer assessment (G1 and G2 in geosciences; P1 and P2 in physics).

For each of these activities (G1, G2, P1, and P2), students' pairs had to commit to an investigation-based activity and had to produce a written report of their work. The same pairs of students had to examine afterward the written artefacts of some

Table 6.1 Implementation plan

Activities	Ed. level	Subject	Topic	Number of students
G1 and G2	Grade 10	Geosciences	Fossil energies	168 (6 classes)
P1 and P2	Grade 10	Physics	Periodic phenomena	172 (6 classes)

other group thanks to a structured template that we had previously built. Then, based on the comments they received, they had to revise their initial written work. In all activities, the frame of the investigation-based activity was the same. Students had to:

- Make hypotheses related to a scientific problem
- Use and apply a protocol giving the guidance to design experiments in order to test their hypotheses
- Proceed through the experiment
- Present and analyze the results of their experiment
- Conclude and solve the problem

A major difficulty for students in the case of peer assessment is to determine criteria (OCDE 2005), to see "which goals, which achievements are hidden behind the work they did" (p. 253). To avoid this difficulty, the protocol developed in this experiment does not involve a *collaborative peer assessment scheme* (Stefani 1994, p. 69) where students themselves have to define their own assessment criteria. We chose to provide students with a template giving them several criteria based on the competences they had to use (Examples: Tables 6.2 and 6.3). The templates presented in Tables 6.2 and 6.3 correspond to two activities. In activity G2, the students had to build an analogical model of a real outcrop and a model of an exploitable oil deposit in order to explain why the real outcrop does not contain any oil; in activity P1, the students were supposed to make previsions if three different materials reflect the ultrasound and propose a protocol of an experiment which can check the predictions, carry out the experiment (constrained by the available devices), and compare and discuss the previsions and the results. All templates were built thanks to collaborative work between teachers and researchers.

We observe that when the criteria are close to the question, students understand them better. For instance, in activity P1, the question is "On which objects among three available, do the ultasounds reflect?" The criterion "the predictions are justified" (fully, incomplete, totally unrelevant) is better understood than "the justification is drawn on elements of knowledge other than ultrasounds."

For each criterion, peer assessors had to give a mark (four-level scale: one is the lowest level of achievement and four is the highest) and a written comment (justification) in order to justify and help assessees to improve their artefact. In geosciences, criteria relate to complex competences (as presented in the French curricula) such as the ability to make empirical observations, interpret them, and conclude (criterion 3). In physics, teachers and researchers worked on the elicitation of these complex competences in order to provide students with criteria which assess rather specific components of the answer (Table 6.3). For instance, in assessing if the pro-

Table 6.2 Assessment template (Geosciences: Activity G2)

Criteria	Description	Mark (1–4)	Justification
1	Schematic drawings are well accomplished and fit with the instructions of the methodology sheet		
2	The model allows us to understand the situation described in the document: The links between elements of reality and elements of the model are effectively made; the oil movement is shown (e.g.,with an arrow)		
3	The observations made on the model are well expressed and interpreted		

Table 6.3 Assessment template (Physics—Activity P1)

Criteria	Description	Mark (1–4)	Justification
1	A prediction is made and justified		
2	The protocol is relevant to test the prediction		
3	The results of the experiment are clearly presented and appear coherent		
4	The experimental results are interpreted in relation to the predictions, and conclusions are drawn with regard to the questions on the reflection of ultrasounds by different materials		

tocol is relevant for testing the hypothesis (criterion 2), peer assessors have to check the coherence of the hypotheses and the protocol and not the quality of the hypotheses or of the protocol.

We focus here on the students' ability to offer the quantitative feedback (i.e., providing a mark) consistent with the peer assesses' level of achievement as determined by an expert. We gathered data in 12 10-grade classes (6 in physics, 6 in geosciences, Table 6.4). We collected written artefacts of each group. Our methodology of analysis consists of cross-examining, with the help of the ASSIST-ME Project coding tool (ASSIST-ME report D5.11 http://assistme.ku.dk) written assessment with the work previously done by the students. We first compared the peer assessors' marking with an expert's marking (Q1) to see if students are able to give a proper mark to their classmates. Then we combined this information with the peer assessors' level to determine if a high level of achievement during the activity is a precondition for peer assessment (Q2).

Data Analysis

The aim of our analysis is to study the students' ability to assess other students' artefacts by providing a mark on a four-level scale. Peer assessors had to provide a mark for each assessed criterion. Table 6.4 gives the number of marks expected for

Table 6.4 Data collected for each activity (all criteria)

Activity	Topic	Number of expected marks	Number of marks collected
G1	The origin of fossil fuels	2 per group Total: 180 (90 groups)	170 (94%)
G2	The formation of a fuel deposit	3 per group Total: 228 (76 groups)	228 (100%)
P1	Periodic phenomena 1	4 per group (4 classes) 5 per group (2 classes) Total: 358 (83 groups)	358 (100%)
P2	Periodic phenomena 2	5 per group (4 classes) 6 per group (2 classes) Total: 389 (72 groups)	360 (93%)

each group (which is also the number of criteria in the template), the number of marks effectively collected for all those criteria (column 4).

Our analysis relies on a comparison between the marks given by an expert (teacher and/or researcher) with the marks given by the students. We studied for all activities and classes the gap between the expert's mark and the student's mark for the same criterion. With a four-level scale for grading, we can get seven values for this gap (from −3 to 3):

- When the gap is null, expert and assessors gave the same mark.
- When the gap is positive, the expert gave a higher mark than the assessors.
- When the gap is negative, the assessors gave a higher mark than the expert.

Regarding our first research question (Q1), we considered that the mark is consistent with the level of the assessees when the gap has a value of {−1; 0; 1} and is inconsistent when the gap has a value of {−3; −2; 2; 3}. Then, we cross-examine this information with the level of achievement of the assessors wondering if a high level of achievement is a precondition for peer assessment (Q2). In order to do this, we have counted how many students:

- Succeeded in the initial task (Mark of their own work given by the expert ≥3) and succeeded in the marking (Gap ≤1)
- Succeeded in the initial task (Mark ≥3) and failed in the marking (Gap ≥2)
- Failed in the initial task (Mark ≤2) and succeeded in the marking (Gap ≤1)
- Failed in the initial task (Mark ≤2) and failed in the marking (Gap ≥2)

Results

In most of the 12 classes, students were committed to the peer assessment task even if initially some of them expressed doubts or concerns in their abilities to be good and fair assessors. For example, students sometimes asked their teachers how they could possibly evaluate the work of their peers without knowing if their own answer was correct or without having successfully accomplished the activity, which shows their willingness to do well by their peers.

Table 6.5 Gaps between assessors' and experts' marks (Note that the number of marks can vary slightly from those in Table 6.3 because we did not take into account the anonymous artefacts)

	G1		**G2**		**GeoS**		**P1**		**P2**		**Physics**		**Total**	
	#	%	#	%	#	%	#	%	#	%	#	%	#	%
Number of marks	161		221		382		326		286		612		994	
Gap = 0	87	54	108	49	195	51	143	44	132	46	275	45	470	47
Gap = 1	10	6	14	6	24	6	56	17	29	10	85	14	109	11
Gap = -1	48	30	60	27	108	28	93	29	70	24	163	27	271	27
Total Consistent						85 %						86 %	850	85 %
Gap = 2	1	1	8	4	9	2	6	2	14	5	20	3	29	3
Gap = -2	8	5	15	7	23	6	25	8	31	11	56	9	79	8
Gap = 3	2	1	9	4	11	3	1	0	3	1	4	1	15	2
Gap = -3	5	3	7	3	12	3	2	1	7	2	9	1	21	2
Total Inconsistent						15 %						14 %	144	15 %

The analyses of the gap between the experts' and the students' marks (Table 6.5) show that, for the most part (85%), students are able to provide a mark consistent with the assessor' levels:

- More than 47% of the students' marks exactly equal the expert mark showing that the assessors gave the proper mark to their classmates (470 out of 994).
- More than 85% of the gaps (850 out of 994) valued between {−1, 0, 1} showing a marking consistent with the assessor level.
- Only 15% of the gaps showed inconsistency (108 with {−2; 2}, 36 with {−3; 3}).

When not equal to 0, the gaps are for the most part negative, meaning that students gave a better mark than the experts. This tendency (which occurred in all classes except one in geosciences for G2[1] and one in physics for P1[2]) could be explained in two ways:

- The assessors wanted to be nice to their classmates and friends (stated by some students in the interviews).
- The assessors didn't recognize a mistake or a lack in their peers' written artefact.

[1] In this specific class, 21 gaps were positive and 21 negative.

[2] In this specific class, 32 gaps were positive and 21 negative.

Table 6.6 Relations between success/failure on a criterion in the activity and the success/failure in the assessment (for all criteria listed by activity)

Initial activity	Marking	G1 N= 156	G2 N=207	P1 N=333	P2 N=327	Total N=1023
Success **N = 719**	Success	126 (90%)	142 (92%)	192 (93%)	194 (87%)	654 (91%)
	Failure	13 (10%)	12 (8%)	13 (7%)	27 (13%)	65 (9%)
Failure **N = 304**	Success	16 (94%)	38 (71%)	107 (83%)	77 (72%)	238 (78%)
	Failure	1 (6%)	15 (29%)	21 (17%)	29 (28%)	66 (22%)

As a preliminary work regarding our second research question, we have numbered, for all criteria listed by activvity, the number of times that a two-student pair have succeeded/failed in the activity and the number of times that a student pair have succeeded in giving/failing to give a proper mark. In total, 752 expert's marks show a success regarding a criterion during the initial activity and 898 students' marks were consistent with the assessors' level. This means that some students who didn't reach a high level of achievement during the activity on a specific criterion (mark ≤2) were able to provide a proper mark on this specific criterion (gap ≤1 with the expert mark).

More specifically, each time we could, we have numbered the times where a two-student pair succeeded/failed in the activity and succeeded/failed in the marking on the same criterion (Table 6.6). We can see (column7) that there are 65 times where students who succeeded in the activity (on a specific criterion) didn't give a proper mark on the same criterion. Conversely, there are 238 times where students who didn't reach a high level of achievement managed to give a proper mark to their peers on the same criterion.

The percentages in the last column of Table 6.6 show that a majority of students were able to invalidate/validate the work of their peers (Q1) even if they did not succeed in the task (Q2). Indeed, 78% of the students with a low achievement on a criterion manage to give a proper mark on the same criterion. This can seem surprising, but the fact that we provided students with meaningful criteria with some details during peer assessment may have helped some students to reconsider and maybe to understand the objectives of the activity. In the first activity in geosciences, we can see that this rate is even better. In 16 cases out of 17 (column 1), students were able to give a proper mark even if they didn't manage to reach the expectation for the same criterion during the activity. This can be explained by the fact that this activity was easier, in terms of knowledge and competences involved, than G2. Students had to make an observation, draw a sample containing information, and use documents in order to discuss the origin of fossil fuels. The criteria were related to the ability to communicate their observation scientifically through the drawing (which students are used to doing) and to write an argumentative paragraph explaining the origin of fossil fuels based on written documents. In G2, they had to build an analogical model of a real situation (complex task) in order to understand how fuel deposits are formed and explain why the situation under study does not contain oil anymore. Criteria (Table 6.1) are related to the ability to establish those links and use those links in an argumentative paragraph. The same goes in physics; P2's activity and assessment were more difficult than P1's.

Conclusion

In this work, we investigate student's abilities to mark a peer's written artefact in relation to the given criteria and in regard to their own artefact in the same activity. We show that most of the students, without any teacher correction, are able to give a proper mark, consistent with the one that an expert could give. Even part of the students with a low achievement rate manage to mark their peers. But one of our major concerns is to know if the fact that most students have the capacity to give a proper mark also implies that those students understand the tasks and their own possible mistakes. The analysis of the comments students gave each other does not show a deeper student understanding. The following study investigates further student comments. A first analysis of the geosciences activity, based on video data of students doing their activity and assessing their peers, tends to show that assessing their peers does not help students to be aware of their own misunderstanding (Le Hebel et al. 2016). The authors conclude that there is the necessity of a crucial phase of discussion after assessment between peers and with the teacher during the phase of correction.

Another concern is to characterize the type of tasks which are most suitable for peer assessment. It seems that one of the main features for this suitability is that students are asked to assess the process (even if they do not assess all the steps of the process) rather than factual knowledge (they may lack). As stated above, it seems that peer assessment, despite the information provided in the criteria, does not always help students to recognize their own mistakes and improve their understanding and knowledge. When assessors and assessees made the same mistake, there is no possibility for them to recognize their misunderstanding even with the help of the criteria. Sometimes, even students who have succeeded in the task are not aware of a misunderstanding in their peers' artefact. These observations show the importance of the role that teachers need to play during peer assessment. We think that a time of common correction by the teacher with the whole class is necessary at the end of the peer assessment. Moreover, peer assessment templates are a way for teachers to share explicit assessment criteria and give students guidance on what is expected from them. Collaborative work and discussion of these criteria (during the peer assessment time and/or during the correction of the activity or possibly during collaborative construction of peer assessment templates) can enhance the awareness of their students' needs in order to improve their knowledge and work processes. By the way, we built new templates for peer assessment in geosciences including an answer along with the associated assessment criteria, each one being located on the answer and a grid with formulated justifications making a scale to assess the level of achievement for each criterion. This grid is a way to permit student discussion and reflection on the answers of their peers. It is a source for students who felt unconfident with the previous template to be certain of their assessment. The fact that the answer is given is maybe not in alignment with the idea that peer assessment is a way for students to improve by themselves. In our case, students have to correct a functional schema which necessitates the whole process understanding. The interpretation cannot be done element by element. The qualitative study that we are currently running, based on comments and video analysis, shows that the improvement is not that valuable even if students seem to be able to give a mark. We think that peer assessment must be a time for metacognition and a reflection on the expectations. This type of grid, considered as support for this metacognitive work, helps students to establish links between the answer, the criteria, the objectives, and the knowledge involved in the activity, even if in some cases it will not be sufficient.

As a perspective, we also want to emphasize the fact that it seems necessary that the responsibility of assessing other students' artefacts given by the teacher to the students should be in accordance with the usual responsibilities given by the teacher. Taking more responsibility in the validation of their artefacts is a main way for students to develop their competences and autonomy but not only for this type of activity, which is, as Allal (1999) underlines, in the spirit of investigation activities. In consideration, teachers have to work on the formulation and moreover the transmission of assessment criteria (OCDE, 2005) in coherence with the classroom practices that they instigate. Students must be aware of what is expected of them.

Examining the Peer Feedback that Secondary School Students Generate for Their Peers' Models in Science (Switzerland)

The educational system in Switzerland allows for much individuality at the level of school units (Husfeld 2009) as well as at the level of the individual teacher at all school levels (Kronig 2009). This is particularly true for the culture of assessment (Vögeli-Mantovani 1999). The use of formative assessment strategies including peer assessment therefore depends largely on the initiative of the school or the initiative of the individual teacher (Smit 2009). In particular, at upper secondary school level, the external stimuli from curricula and from textbooks play a minor role, and the interest in formative assessment is almost exclusively triggered by ongoing activities at the compulsory school levels.

Educational research indicates that the development of modeling competence is facilitating students learning of science, about science, and of how to do science (Saari and Viiri 2003; Schwarz and White 2005). The modeling competence could be fostered in the context of modeling-based learning, which refers to "learning through construction and refinement of scientific models by students" (Nicolaou and Constantinou 2014; p. 55). In consideration of the complexity of acquiring and mastering the model construction competence itself, it might be even more demanding to request secondary school students to offer feedback to their peers, after having evaluated their peers' models. Model critiquing involves engaging students in discussing the quality of models for evaluation and revision (Chang et al. 2010; Schwarz and White 2005; Schwarz et al. 2009). Critiquing is an important scientific practice that needs to be addressed in science classrooms (Duschl et al. 2007), and it could be practiced through peer assessment activities. A few studies (e.g., Chang and Chang 2013; Pluta et al. 2011) have provided evidence specific to the educational value of teaching-learning activities that involve the critiquing of models. Chang and Chang (2013) stressed the need for more research in this direction, with the ultimate goal being the identification of what students can do when assessing peers' models (Chang and Chang 2013). Considering the lack of previous research on what student critiquing of peers' models entails, we sought to further examine this issue in this study.

Objectives and Research Questions

The present study focuses on peer assessment of the model construction component. It aims at investigating the peer feedback that secondary school students generate for their peers' model in science.

More specifically, two research questions are addressed:

1. What are the characteristics of the qualitative peer feedback that secondary school students generate while assessing their peers' models in the context of light and color in a physics course?
2. What is the relation of qualitative (written comments) and quantitative (ratings in the rubric) peer feedback provided by secondary school students?

Methodology

Participants

The physics teacher involved in this study volunteered to participate with his class in a classroom implementation of formative assessment methods developed by the ASSIST-ME project. The typical instructional format of the classroom was characterized by student group activities in the laboratory due to the nature of the course (physics) in combination with lectures given by the teacher. The learning goals of the course entailed the development of conceptual understanding of physical phenomena and the development of experimental and problem-solving skills. In the meetings that took place among the teacher and the researchers involved in this study, for organizational purposes, it was clarified by the teacher that the peer assessment method was not a commonly practiced method in his class, except the cases in which students exchange oral peer feedback among them in a nonformal setting.

The class comprised 22 students of the 11th grade at a Gymnasium (i.e., the highest track at upper secondary school) in Switzerland. The students were assured that the study would not contribute to their final mark. They worked in randomized pairs throughout most parts of the intervention, and the pairs did not change during the intervention. There were 11 groups of 2 students. However, the students worked individually when enacting the peer assessor role.

Teaching-Learning Sequence

The sequence was grounded in collaborative modeling-based learning. The students worked through learning material on the topic colors and light. The curriculum material required the students to work with a list of hands-on experiments on additive and subtractive color mixing (McDermott et al. 1996). After completing the experiments, students were instructed to draw inferences relying on their observations and the gathered data. Their inferences were explicitly expected to lead to a scientific model which represents, interprets, and predicts the additive and subtractive color mixing of light. In order to do this, students were provided with a sheet of paper, color pencils, and a list of specifications that they were asked to consider when developing their model. The list of specifications constituted three benchmarks that students should consider while developing their model which are the following: the model should (i) represent, (ii) interpret, and (iii) predict the additive and subtractive color mixing of light. The list of specifications was in line with the assessment criteria that were given later on during the peer assessment activity to students. Overall, it took the student groups five lessons of 45 minutes to complete this sequence.

The Process of Peer Assessment

As soon as the students had finalized their models in their home groups, they exchanged them with models of other groups; i.e., two groups assessed their models mutually. The exchange pairs were randomly defined by the teacher. Peer assessors used a rating scale with eight prespecified assessment criteria (e.g., Does the model appropriately represent what subtractive color mixing is? Is it explained and justified in the model, which are the primary colors? Can the formation of white or black be derived from the model?), which were in line with the list of specifications that was given to students while constructing their models. Assessors rated their peers' models on all criteria according to a four-point Likert scale (i.e., (1) unsatisfactory; (2) moderately satisfactory; (3) good; (4) (fully) satisfactory/excellent). A fifth column was provided next to the rating scale for each criterion for the provision of written comments. Along with ratings, assessors were prompted by the teacher to provide written feedback (for each criterion separately) to assessee groups, in which they were to explain the reasoning behind their ratings and provide judgments and suggestions for revisions (Table 6.7).

The students were instructed to individually assess the model of the peer groups assigned to them. On average, it took each peer assessor 15 minutes to complete the

Table 6.7 The four-point rating scale with eight prespecified assessment criteria. This is an example of a fulfilled rating scale provided by student coded as 8A from the assessor group 8 to the assessee group 9. The model of group 9 is presented in Fig. 6.1. Note: The text was translated from German to English

Assessment criteria for the model for color mixing: Assess your peers' model according to the following criteria. Provide a meaningful comment in the space provided.

Assessee Group: 9		Your code: 8A				
		1 = unsatisfactory, 2 = moderately satisfactory, 3 = good, 4 = (fully) satisfactory/excellent				
	Assessment criteria	1	2	3	4	Your comments
1	Does the model appropriately represent the primary colors?			x		Only with drawing, without explanation. Bottom right and the light circles you can recognize them. But it is not explained
2	Does the model appropriately represent what additive color mixing is?			x		Very good model, maybe a small text to explain … again no description of what is what
3	Does the model appropriately represent what subtractive color mixing is?			x	x	Illustration provided
4	Does the model appropriately represent how color filters work?				x	Illustration provided
5	Does the model explain how color filters work?		x			Description missing
6	Is it explained and justified in the model, which are the primary colors?		x			Only with drawing, without explanation. Cyan and magenta were not different from yellow without blue
7	Can the formation of white or black color be derived from the model?		x			White is to be recognized in the color circle on the bottom right. Black is illustrated at the top as white. However, it is not explained how black is formed, and it is not said that black is formed
8	Can you predict what is formed in the additive mixture of yellow and cyan, relying on your peers' model?			x		It does not say what color is formed. However, good representation. The color circle shows a lot

assessment. Once the students had completed the assessment of their fellow students' models, they exchanged their peer feedback and reviewed it in collaboration with their groupmate, deciding whether to make any revisions to their model.

Data Collection

At the beginning of the intervention, a consent form was signed by the students' parents, allowing us to use the collected data anonymously and for research purposes. The filled-out rating scales with peer feedback comments produced by students, along with their schoolwork, own constructed models (e.g., in Table 6.7 and Fig. 6.1) were collected to allow us to address our research objective.

Data Analysis

We used a mixed-method approach that involved both qualitative and quantitative analyses of the data. In particular, data were firstly analyzed qualitatively and then also treated quantitatively with the use of the SPSS™.

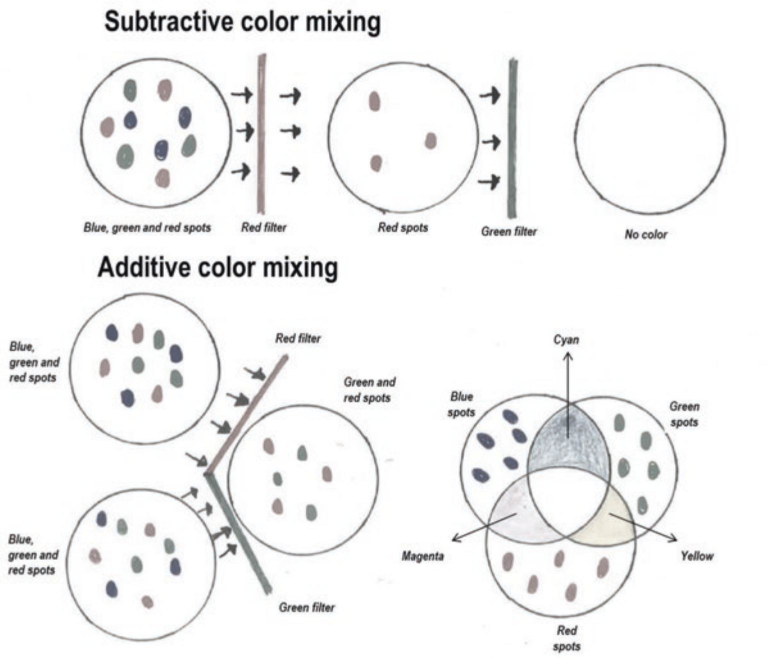
Determining the level of quality of peer feedback requires examining the quality of both the quantitative and qualitative feedback. In this study, we mainly focus on the qualitative part of peer feedback, that is, to say the written comments provided in the four-Likert scale rubric along with the ratings, because it has been emphasized in prior research that qualitative feedback is more important than quantitative (e.g., Topping et al. 2000).

To analyze peer feedback written comments, we developed and further used a coding scheme including the following dimensions:

1. Comprehensiveness (whether the assessors drew on the intended assessment criteria and to what extent).
2. Validity of peer feedback comments (their scientific accuracy and their correspondence to the models assessed).
3. Verification of comments (peer feedback comments were perceived as "positive" when including references to what the assessees had already achieved with respect to the list of specifications; likewise peer feedback comments were perceived as "negative" when including references to what the assessees had yet to achieve with respect to the list of specifications).
4. Justification of positive and negative comments (i.e., justification offered by the assessor(s) on what the assessees had achieved or not yet achieved with respect to the list of specifications related to the modeling competence).
5. guidance provided by the assessor(s) to the assessee(s) on how to proceed with possible revisions. We perceived "guidance" as statements which could potentially help the assessee(s) to improve their models.

The peer feedback comments of each student were coded separately. Each complete sentence included in the peer feedback comments, from each student, was analyzed with respect to all the aforementioned categories (comprehensiveness, validity, verification, justification, guidance). The resulting codes were further used quantitatively, for running nonparametric correlations (Kendall's T_b) between the coded written comments and the ratings assigned by students to each criterion of the rating scale.

A possible internal consistency between the qualitative (comments) and quantitative (ratings) peer feedback provided can be used as an indicator of the quality of peer feedback (Hovardas et al. 2014). For that reason, we further examined whether there is a statistically significant correlation between the quantitative score assigned and the number of references to what the assessees have already achieved and what



-All three primary colors → white light
-No color → black
-How a filter functions: a blue filter filters everything but blue
-Likewise for a cyan filter → green, blue

- ~~B~~ 3 Grundfarben → Weiss
- Ohne Farbe → schwarz
- ~~Filter Fi~~ ~~I~~ Funktion Filter:
  Blau filtert alles ausser Bla
  Cyan → Grün Blau

Subtractive color mix:
The first filter filters the white light and allows only some colours to pass. Because the green filter only allows green light to pass, while only red light falls on the filter, nothing comes through.
→<u>This is how black is formed</u>

This can be adopted for all colors.

- subtraktive Farbmischung
  Der erste Filter filtert das weisse Licht
  und lässt nur rotes Licht durch.
  Da der Grünfilter nur grünes
  Licht durchlässt und nur rotes
  Licht auf den Filter trifft, ~~bleibt~~
  ~~keine~~ kann nichts durchdringen.
  => Schwarz

Dies kann man auf alle Farben anwenden.

Fig. 6.1 This is an example of the type of models that students constructed and the kind of revisions applied by assessees after receiving peer feedback. The initial model of Group 9 included three representations (*on the top*). The students added in their revised model an explanatory text for their representations (*on the bottom*). Note: The text was translated from German to English (*bold font letters*). Also, labels have been added for the each color for readability purposes (*letters in italics*)

the assessees have not yet achieved in respect to the modeling competence (i.e., if quantitative scores are positively correlated with number of positive judgments references to what the assessees have already achieved in respect to the modeling competences and negatively correlated with number of references to what the assessees have not yet achieved in respect to the modeling competence).To estimate the interrater reliability of the data coding, a second coder who had not participated in the first round of coding repeated the coding process for 40% of the peer feedback data and the students' models. Each peer feedback comment (one complete sentence) provided by each assessor for each assessment criterion of the rating scale that students used while giving feedback was rated for 16 items of the coding scheme, addressing the aforementioned dimensions. In all cases, the two raters involved in the data analysis process were also asked to justify their reasoning for their ratings and/or provide illustrative examples. Krippendorff's alpha was calculated above 0.79 for the coding of all data. The differences in the assigned codes were resolved through discussion.

Results

Type of Qualitative Peer Feedback Provided

Assessors provided feedback comments which were found to carry affective connotations. In particular, peer assessors provided on average 3 and in total 75 comments which carried affective connotations (illustrative quotes: "Well done, well described," "Very nice illustration, I like the models"). Assessors tended to provide a balance of statements which are likely to serve as discouraging and encouraging feedback for the assessees.

In terms of the comprehensiveness of peer feedback comments, the analysis revealed that assessors took into account most of the assessment criteria which were given to them while peer assessing their peers' models (i.e., they rated and commented on average on 6.54 criteria out of the total of 8 criteria of the given rubric) and they drew on them in a rather thorough manner. In particular, in the criteria where the assessors did not provide full marks in their rating with the four-point Likert scale, they justified the awarded low marks by suggesting what was missing and what could be added in their peers' models; in other words, they were specifying what their peers did not manage to achieve in respect to their modeling competence. With respect to the validity of peer feedback comments, the findings have shown that assessors' judgments with respect to the criteria which they attended to were mostly valid. In a few cases, students provided invalid comments to their peers, and those comments were found to be related to misconceptions identified in their own models.

Peer feedback comments were found to be critical enough, as the assessors tended to include in their feedback comments more negative and fewer positive comments. In particular, the data analysis revealed that assessors provided on average 3.71 positive comments (M = 3.71; total = 89; SD = 2.87; illustrative quote:

"Graphically very well recognizable and easily comprehensible") and on average 5.58 negative comments (M = 5.58; total = 134; SD = 2.28; illustrative quote: "...but it could be added why it is called additive color mixing."). Positive and negative peer feedback comments were mostly justified. Assessors provided on average 3.0 and in total 73 justified positive feedback comments (SD =2.76) (illustrative quote: "Good, because it is explained and graphically illustrated.") and on average 0.67 and in total 16 positive comments (SD =1.09) which were not justified (illustrative quote: "Your model looks great."). Similarly, the assessors provided on average 5.33 and in total 128 negative comments (SD =2.28) that were justified (e.g., "...the thing with white containing all colors not enough. Below maybe use red, blue and green dots instead of white ones.") and on average 0.25 and in total 6 negative comments (SD =0.53) which were not justified ("Wrong.").

Moreover, assessors provided the assessees with guidance statements (average = five comments per assessor) on what the assessees needed to further achieve to improve the quality of their models. Overall, all the assessors provided 115 comments which could be perceived as guidance statements. Five statements were not related to the competence of interest but with superficial aspects related with the appearance of the model (e.g., legibility of writing; illustrative quote*:* "The writing is not always legible."), whereas the rest were related (e.g., "An interpretational text is missing, so you should add this in your model."). In addition, specific guidance about next concrete steps, provided by the assessors, was found to be present to a very small extent (e.g., "I understand from your model how white is formed but not black.") up to a great extent (e.g., "The primary colors are green, red and blue. Orange & violet are mixed colors as well as magenta, cyan. Therefore, you should revise your model according to this."). The guidance statements provided by the assessors were found to be mostly valid; this means that the students, as assessors, were able not only to identify most of the weaknesses in their peers' models but also to guide them on the next steps that were to be taken to improve their models.

Relation of Qualitative (Comments) and Quantitative (Ratings in the Rubric) Peer Feedback Provided

As part of the quality check of peer feedback, we ran Kendall's T_b correlations between the written feedback given per criterion on what the assessees have achieved or not in respect to the modeling competence and the score assigned to the corresponding criterion, to check whether there was internal consistency between the quantitative and qualitative aspects of the provided peer feedback. The analysis revealed that the mean scores, which were assigned by the assessors in the four-point Likert scale rubric along with their comments in the criteria which they attended, were negatively correlated with the number of negative comments (i.e., references to what the assessees had not yet achieved in respect to the modeling competence) (Kendall's $T_b = -0.749$; $p < 0.01$). This means that the lower the score an assessor was assigning to a certain criterion, the more the possibilities were for providing written comments to what the assessees had not yet achieved. No statistically significant correlation was found between the mean of scores and the number of references to what the assessees had already achieved in respect to the modeling competence.

Discussion and Conclusion

This study focused on examining the type of peer feedback that secondary school students generate while assessing their peers' models in the topic of light and color in a physics course. The findings of this study have shown that the participants were committed to their assessor role in a satisfactory way. They offered justified positive and negative comments, as well as guidance statements to their peers, acknowledging what assessees had already achieved or not yet achieved in respect to the list of specifications which was given to them and was related to the modeling competence. Assessees were also capable of offering guidelines to their peer for the next steps to be taken for improving their models. To follow on from this, it was found that assessors tended to provide a balance of statements which were likely to serve as discouraging and encouraging feedback for the assessees. This indicates that assessors could identify flaws and shortcomings in assessees' models (thus offering negative comments), justify their comments, and provide suggestions for improvements (offering guidance). Suggestions and recommendations for possible ways for improvement (Hovardas et al. 2014; Strijbos and Sluijsmans 2010), as well as justified comments (Gielen et al. 2010; Narciss 2008; Narciss and Huth 2006), have been identified by researchers as essential characteristics of constructive peer feedback. This commitment to the assessor role was also confirmed by the study's consistency checks, which revealed that the peer assessors were attentive while assessing. For instance, it was found that scores given by peer assessors were negatively correlated with number of references to what the assessees had not yet achieved in respect to the modeling competence, which indicates that quantitative aspects of peer feedback (i.e., negative ratings) were consistent with qualitative aspects of peer feedback (i.e., references to what the assessees had not yet achieved in respect to the modeling competence). Also the findings of this study have revealed that peer assessors drew on almost all available criteria of the rubric while providing feedback. Lastly, peer feedback comments were found to be mostly scientifically accurate and consistent throughout, even though in a few cases assessors offered invalid peer feedback comments which were related to misconceptions identified in their own models. Overall, the findings suggest that students have the beginnings of providing peer feedback of good quality in modeling-based learning.

The findings of this study have some practical implications. The fact that secondary school students were committed to the peer assessor role renders peer assessment as a promising formative assessment method and learning tool in modeling-based learning. Students, as assessors, were found to be capable of providing their peers with constructive feedback in a satisfactory manner, which could be ultimately used to foster the learning progress of both. Teachers could use the peer feedback as a probe to students' model construction practices and understanding of scientific concepts.

However, considering that few students—who addressed some of the intended specifications in an invalid way in their own model—provided their peers with invalid feedback comments with respect to those specifications, it is implied that the validity of peer feedback comments is related to students' understanding of the topic under emphasis. Previous studies have shown that the quality of peer feedback is associated with students' understanding of the subject studied (Ballantyne et al. 2002). Peer assessment requires students to use not only their assessment skills (Sluijsmans 2002) but also their knowledge on the content/topic in order to review, clarify, and correct peers' work. Teachers who engage their students in peer assessment activities in modeling-based teaching should safeguard that students, as assessees, do not receive wrong signals from their peers through the peer feedback comments, especially when the assessors have not completely comprehended what constitutes a good scientific model or how the phenomenon operates. This requires support mechanisms for the teachers themselves, via tools or guidelines on how to filter peer feedback content, before exchanging it among students or on how to easily identify students' models of low validity and henceforth filter or closely moderate the feedback comments that those students deliver to their peers.

This study was carried out in the context of a regular physics course, thus offering ecological validity to the aforementioned findings. On the other hand, this study reveals limitations related to the sample size, which was small (22 students). Future studies should replicate the findings of the present study with bigger sample sizes. The conclusions of this study should be limited to the particular characteristics and affordances of the students involved.

Overall peer assessment in modeling-based learning constitutes an area that calls for future research to further explore the potential benefits that peer assessment may entail in this context.

Student Perspective on Introduction of Formative Peer Assessment (Czech Republic)

The idea of formative assessment which emerged in the 1990s was only slowly introduced in the Czech educational context (c. f. Žlábkova and Rokos 2013). The Czech Framework Educational Programme for Basic Education implemented in 2007 aims at gradually accomplishing changes in the assessment of pupils toward diagnostics on an ongoing basis, assessment of pupils' achievement history, and a wider use of verbal assessment (compared to marks). The practice of assessment ought to be driven by guidelines embedded in each school rules document. This document should describe principles and methods of assessment and self-assessment of learning outcomes and conduct of students, including the acquisition of data for evaluation and criteria for the evaluation. Self-assessment is thus explicitly stated as a necessary part of the school assessment practice, but explicit peer assessment is not. Peer assessment is seen rather as a particular method in cooperative teaching (Novotná and Krabsová 2013).

In recent years, there has been great interest in upscaling the formative assessment as there is a need for change in learning culture (Education at a Glance 2015: OECD Indicators Czech Republic 2015; Santiago et al. 2012). The problem is that though there are some examples of good practice (c. f. Košťálová et al. 2008; Kratochvílová 2011; Slavík 2003), they are not empirically studied and focus mostly on selected subjects and educational levels. Empirical research which would provide some evidence on the effectiveness of using various formative assessment methods (FAM) is quite scarce (Novotná and Krabsová 2013).

Some types of formative assessment are seen as more or less embedded in common Czech teaching culture, like teacher provided on-the-fly assessment or to a lesser degree written feedback (Laufková and Novotná 2014; Lukášová 2012; Košťálová et al. 2008; Novotná and Krabsová 2013), and methodological literature for teachers (e.g., Kratochvílová 2011; Starý 2006) pays more attention to these forms of formative assessment. In these materials, formative peer assessment is mentioned only as a supplementary option (e.g., Košťálová et al. 2008), probably also due to the fact that peer assessment as a form of classroom communication is not very frequent (Šeďová et al. 2012) and formative peer assessment was not yet empirically studied in Czech schools. The aim of this study is to investigate students' views on introduction of formative peer assessment in inquiry-based lessons of primary mathematics and biology and secondary biology.

Objectives of the Study and Research Questions

The main research questions relate to the students' reflection of peer assessment in inquiry lessons aiming at development of problem-solving and empirical investigation competencies. The central issues under investigation are:

- How do students perceive formative and summative peer assessment?
- Do they prefer peer or teacher assessment?
- How do they reason out their preferences?
- Which difficulties do students experience in providing their peers with assessment?
- How did they perceive the peer assessment that they received and its value for their learning?

Methodology

Formative peer assessment was investigated in three samples of students in different subjects (primary mathematics, primary biology, and secondary biology), who tried to provide their classmates with oral (second graders) or written (third to ninth

Table 6.8 Implementation plan

Subject	Education level	Topic	Organization of work and peer assessment	Number of students
Primary mathematics	Grades 2, 4, 5	Basic geometrical shapes and their area, big numbers	Group discussion (2nd grade), pairs or small groups	113 (6 classes)
Primary biology	Grades 3, 4	Germination	Individual	79 (experimental groups in 6 classes)
Secondary biology	Grades 6,7 8, 9	Germination human physiology	Individual	76 (experimental groups in 6 classes)

graders) peer assessment while solving inquiry tasks oriented toward problem-solving (PS) and empirical investigation (EI) competence development (Table 6.8). Formative peer assessment was structured (by forms in written form and by teachers questions in oral form) and was, as such, a new method of assessment for both students and teachers. Also the inquiry tasks were developed ad hoc for this study (see description below).

Experimental teaching units consisted of one to six inquiry tasks of different complexity, which usually took 2 months of teaching. In biology classes, we also use bogus peer feedback (the written feedback was in fact provided by a teacher, but the students did not know this) to see whether the students react more to the authorship or the content of the assessment. Inquiry task in mathematics usually presented natural life problem, and the students were asked to search for possible solution, describe the procedure, and use mathematical notation of it. The task in biology asked for finding factors which influence particular phenomena (germination, breathing frequency) and develop and experiment which could validate hypothesized factors.

Whereas students participating in the primary mathematics group worked mostly in small groups and provided peer feedback also for groups of mates (in group discussion in the second grade, in rubrics with oral comments in the fourth grade), the students participating in both primary and lower secondary biology groups worked on tasks individually and used only rubrics for providing their peers with the feedback (Table 6.9). In these rubrics the assessors provided formative feedback (e.g., assessment of experiment design, of the possibility to collect the necessary evidence and use this evidence for testing the stated hypothesis) and also summative feedback, using a mark (1–5) summarizing the overall assessment (hence, the name summative feedback) of the task solution and a justification of the mark.

Structured interviews contained a set of questions to prompt students' reflection of their experience (questions relevant to the peer assessment are in Table 6.9). The interviews were transcribed. Where appropriate (questions 4–7), we also used data from the students who only received peer feedback but did not provide it (e.g., due to taking longer on their own solution, or as a member of the control group; primary biology classes N = 61; lower secondary biology classes N = 67).

Table 6.9 Structured interview questions

Questions	Template codes
1. Have you been assessing the work of some of your peers?	
2. Do you think that you did well in assessing the peer's work?	(1) Yes or mostly yes, (2) no or mostly no, (3) I do not know or other answer
3. Did you have any difficulties when assessing the work of your peers? What kind and why?	(1) Lack of knowledge or skills (related to task, criteria, proper formulation)[a], (2) social regards (positive and negative)[b], (3) would also rather know other solutions[c], (4) bad handwriting—not sure whether understood correctly, (5) no difficulties reported
4. Would you rather get the feedback on your work from your peers or from the teacher? Why so?	(1) From teachers, (2) from peer, (3) does not matter
5. Do you think that there are any differences between the teachers' and peers' assessment of your work? What sort?	
6. When you received the feedback from the peer, did it help you to improve your solution?	(1) Yes or mostly yes, (2) no or mostly no, (3) I do not know or other answer
7. When you received the feedback, what information interested you the most?	(1) Mark, (2) comments to your work, (3) both

Examples of coded utterances:

[a] Firstly, we need to know what it is about and what should be done, and I am not sure what is correct; I did not know how to describe what the flower needs, the assessment itself, whether it is correct or not; I did not understand the picture, it was difficult

[b] It is difficult if you know that it is your friend; I did not want to assess him badly

[c] I did not know whether it was correct, I may have solved it in another way, I would rather see more solutions

The answers to questions were coded by three coders according to the same template. The differences in codes were discussed until consensus was reached.

Data Analysis

We used deductive thematic coding with a template approach (Crabtree and Miller 1999). An a priori template of coding categories in the form of a codebook was applied as a means of organizing the data for subsequent interpretation. When using a template, a researcher defines the template (or codebook) before commencing an in-depth analysis of the data. The codebook is sometimes based on a preliminary scan of the text, but for this study, the template was developed a priori, based on the research question. The codebook presented coding categories (template codes in Table 6.8) and examples of codes from utterances of different age cohorts. A

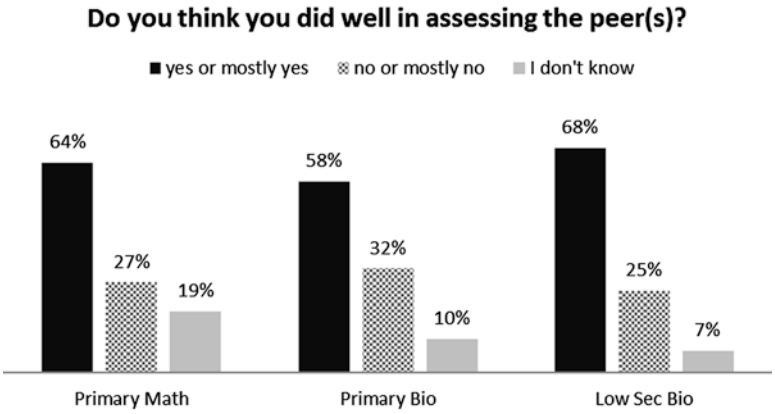


Fig. 6.2 Subjective perception of doing well in assessing the peers

quantitative survey of received codes was then compiled and compared between the age and subject cohorts.

Results

The participating students did not have any previous experience with peer assessment, and therefore we wanted to know how well they did in providing the feedback on peers' work.

The students felt relatively competent in assessing their peers, regardless of their age and regardless of the subject or form of feedback (see Fig. 6.2).

Nonetheless, the students reported some difficulties they encountered when providing the feedback. Figure 6.3 shows the most frequent areas of difficulty that the students mentioned.

The most frequent and most consistently reported difficulty relates to the lack of knowledge or skills necessary for correct assessment, which was associated either with uncertainty about the solution of the inquiry task or the criteria for the assessment. For example, the task on breathing frequency asked the students to develop an experiment which could decide whether there is a relation between intensity of motion and breathing frequency. The assessor was prompted to assess design of the experiment, whether it will provide data on investigated relation and whether the experiment could be realized and how many relevant factors are included. The students who experienced uncertainty about the solution or criteria mentioned that, e.g., "it was difficult to decide whether it is correct or not," "I did it differently and I am not sure that this could be realized," etc. Dealing with this uncertainty in the classroom is crucial for implementation of peer feedback in a broader context. What is further evident is that the primary mathematics groups, which mostly worked in small teams, was more sensitive to social regards when providing the feedback (e.g., worries of negative emotions of peers, bad own feelings when assessing the bad

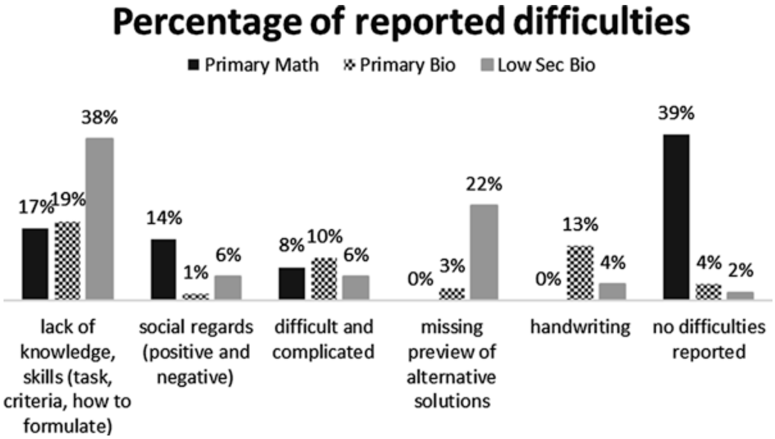


Fig. 6.3 Percentage of reported difficulties when providing peers with feedback

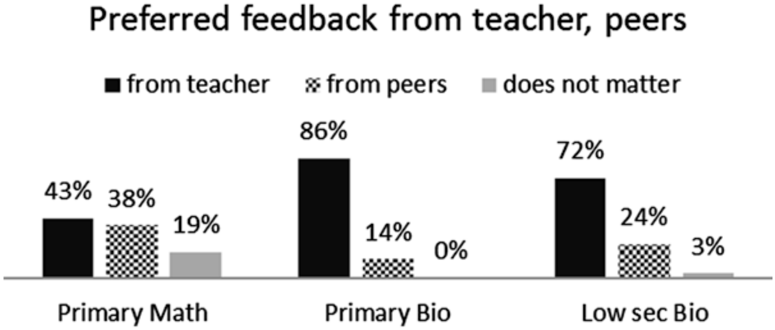


Fig. 6.4 Preferences of teacher feedback and peer feedback

artefact of a good friend, etc.). On the contrary, students from the primary mathematics group did not mention the need to see alternative solutions or trouble with handwriting, and more than one third of this group did not report any difficulties.

An important issue emerging from the interviews with students (and teachers) was the preference of feedback either from the teacher or from peers. Before the start of experimental teaching, the teachers frequently expected that the fact that peer feedback is expressed in more accessible peer language could be a potential advantage of implementation of formative peer assessment. We therefore asked the students about their preferences and reasons for them.

Again, there is a difference between biology and mathematics groups, students in the mathematics group do not show such a pronounced prevalence of teacher assessment preference as the other two cohorts do (Fig. 6.4). The mathematics group students valued the ideas of their peer assessors as easily accessible and did not worry that much about their own mistakes when these are mentioned by peers, unlike when it is the teacher. Biology group students significantly preferred the

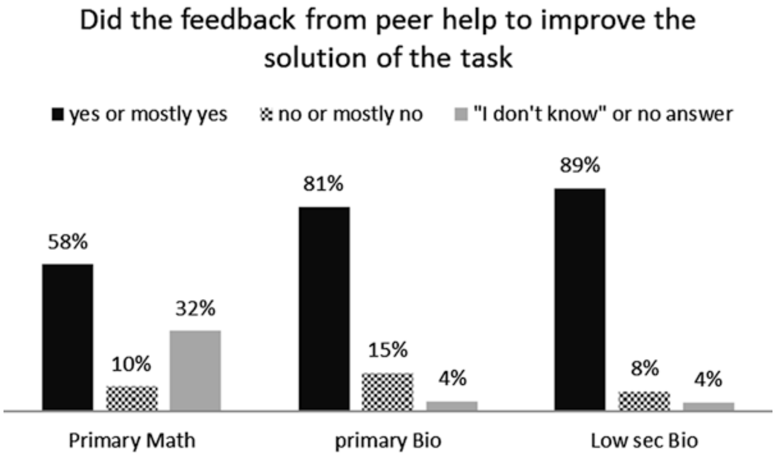


Fig. 6.5 Students' perception of formative impact of peer feedback

feedback to come from the teacher and came to the conclusion it was more reliable and effective (pointing out weak or bad parts of the solution more directly).

What is considered as essential for formative assessment is the informative part of feedback which can foster further learning. For the question whether the feedback from peers was seen as valuable for further improvement of the artefact/solution, the codes were summarized into three categories: positive statement, negative statement, and uncertain standpoint (see Fig. 6.5). In the secondary biology group, we used also bogus peer feedback where the feedback written in a rubric was in fact provided by a teacher. We wanted to know whether the quality and particularity of feedback when thought to be provided by peers is used more for the improvement of their own work. No differences were found; some students did not pay attention to the utilization of any feedback, regardless of its quality. What is even more important is the fact that students reported that the feedback was valuable for improving their solution, but did not actually use it, as could be seen from the working sheets.

We asked all the students, who worked in biology groups where comments as well as marks were provided by assessors in the feedback rubric, what they were more interested in.

In spite of the fact that the students first searched for the mark, which was also frequently mentioned by teachers, students reported that they were more interested in comments that they received on their work (see Fig. 6.6).

Discussion

A student's perspective on their own learning is an important resource of valuable information for teachers, especially when they are trying to implement new approaches to teaching. To understand the ways in which their practice influences student learning, they need to listen to students' accounts of their learning

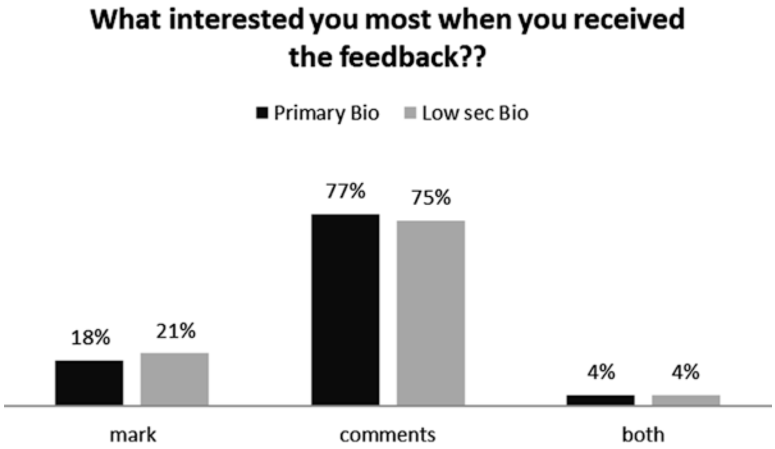


Fig. 6.6 Marks and/or verbal assessment preference

experiences (Kane and Chimwayange 2014). Students participating in this study experienced peer assessment in inquiry lessons for the first time. The results show that the students considered themselves as rather capable of providing peer assessment. Nonetheless, they also experienced some difficulties. The most pronounced category of these was a subjectively perceived lack of domain knowledge needed for evaluating correctness of the peer's solution or procedural knowledge related to providing proper hints or advice for peers on how to proceed further. Students also expressed that they lacked the opportunity to see other alternative solutions which made them uncertain in assessing the peer's work. It seems that specification of proper criteria in rubrics is not enough and that students need some support in applying them, at least when they are not yet used to it.

Results also showed that students also take into account the social context of peer assessment. It was more visible in our primary mathematics sample, where the students worked in small groups. Working in groups can make the students more aware of the socio-emotional context of assessment. Working in small teams led, on the contrary, to more opportunities to discuss the solution and the formulation of proper feedback before giving it to the other team. These students therefore did not report interest in other solutions. The mathematics group students also reported a very small proportion of trouble with handwriting and a high percentage of no difficulty. This could also be influenced by the fact that students in these groups frequently used the possibility to add a verbal explanation to what is written when problems in understanding arise or provided feedback within group discussion (second grade) instead of in writing.

Students in biology groups, who were working and providing peer assessment individually, preferred a teacher's assessment over that of their peers significantly more than the mathematics group. It could be related to the fact that biology groups worked on more convergent inquiry tasks than the mathematics groups where the tasks were more divergent and sharing them in small groups usually led to broader discussion in the whole class. The use of rubrics in our biology groups may also have reduced the advantage of "peer language" as it was more structured, and the proportion of open expression in the feedback was smaller. It seems that a lack of opportunity to broadly discuss the solution and the criteria for its assessment increases students' reservations about plausibility, correctness, and completeness of the feedback from peers.

Conclusions

The students' interviews provided important information on the main challenges and hindrances the students faced in providing peer assessment. As this method of formative assessment is not common in Czech learning culture, students experienced peer assessment for the first time. The students seemed to prefer feedback from a teacher over the assessment of their peers because they see it as more reliable and relevant. But this preference is dependent on the organization of students' work and peer assessment. When students worked in small groups, where they had more opportunities to discuss the solution, this prevalence disappeared. A similar finding is that the work in small groups was accompanied by fewer difficulties encountered in providing the feedback. Discussions in small groups may alleviate the uncertainty associated with a lack of factual knowledge and with any socio-emotional consequences of assessment. Though the students reported that the feedback helped them to improve the solution of the task, they carried out these improvements only in small numbers (which was found in the analyses of revised protocols). This could be due to a lack of time, motivation, or fatigue of students (mathematics lessons videodata provide evidence that it was the case in two inquiry tasks). It is an important message that teachers must pay attention to.

Students also commented on the difficulties they had in assessing inquiry of their peers (uncertainty about correctness, about a considerate way of reporting mistakes, etc.) and advantages they saw in peer feedback over feedback from the teacher, although these issues are probably dependent on the organizational forms of instructions (e.g., individual vs. small group work). Students mostly appreciated the inquiry tasks and would like to extend the proportion of such learning to their everyday lessons. They perceived peer feedback as a new alternative to assessment, but having experienced it for the first time, they remained more in favor of feedback provided by a teacher.

Synthesis of Results and Perspectives

In the French study, results show most of the students are able to give a quantitative feedback with a proper mark, consistent with one that an expert (researcher/teacher) could give without any teacher correction. In this study related to investigation competence, even students with a low achievement rate manage to mark their peers. However, the fact that students with a low achievement rate manage to mark their peers may indicate at least a surface understanding (e.g., if the question and the answer respect drawing codes, the wording of the answer uses similar words to the question), even if the link between the quality of students' peer feedback and the quality of their own artefacts may mean that justification and comments reveal a deeper conceptual understanding. The findings of the Swiss study show that students are able to offer their peers justified negative and positive comments as well as guidelines for the next steps to be taken, acknowledging what the assessees have already achieved or not in respect to the modeling competence. Quantitative aspects of peer feedback (i.e., negative rating) are consistent with qualitative aspects of peer feedback (i.e., references to what the assessees have not yet achieved in respect to the modeling competence). In the Czech study, one of their conclusions is that students perceived peer feedback as a new alternative to assessment; however, following their first experience of it, they express their preference for the feedback provided by a teacher. But this preference is dependent on the organization of students' work and peer assessment. When students worked in small groups, this prevalence disappeared. The authors suggest that the kind of task (e.g., in biology or mathematics) could also have an impact on students' feedback perception.

The three studies conclude the necessity of allowing the sharing of "knowledge authority" in the classroom to evolve. It needs to be integrated in usual classroom practice. However, researchers have a divergent view on the sharing of responsibility for validation of knowledge between the student and the teacher. For instance, the authors of the Swiss study propose that to prevent assessees from receiving wrong signals from their peers through the peer feedback comments, especially when the assessors have not completely comprehended what constitutes a good scientific model or how the phenomenon operates, the teachers themselves might be able via appropriate tools and guidelines to filter feedback content before exchanging it among students. The French authors have a different position on this point and propose to give assessors more autonomy in assessing the other students' artefact. That is in line with previous studies showing that peer feedback leads to more discussions and checking for confirmation and consequently a deeper understanding than teacher feedback, accepted as such and often misinterpreted (e.g., Yang et al. 2006). For the French authors, the phase of discussion after assessment between peers is crucial, and the phase of institutionalization by the teacher during the phase of correction is also essential. Students need to be aware of what the teacher and their peers expect from them as assessees and assessors. Moreover, peer assessment templates could be a way for teachers to share explicit assessment criteria and give students guidance on what is expected from them. Collaborative work and discussion of these criteria (during the peer assessment time and/or during the correction of the activity or possibly during collaborative construction of peer assessment templates) could enhance the awareness of their students' needs in order to improve their knowledge. Peer assessment could be a way to trigger metacognitive work on knowledge and competences in science.

Moreover, in the countries participating in ASSIST-ME, most of the teachers agree on the usefulness and effectiveness of formative assessment, but they all express lack of time to implement it in classroom (see Chap. 3). In these three studies, included in this chapter, teachers all express that peer assessment is time-consuming. As it is already pointed out in Chap. 3, we think it is crucial that teachers do not conceive peer assessment as an add-on to the usual teaching but perceive it as central and integrated part of teaching. It implies that teachers implement peer assessment in a continuity with the other assessments (formative and summative). For instance, they have to use the same criteria in their different formative and summative assessments.

Further research from these three studies could focus on to what extent peer assessment helps students to develop understanding and competences involved in the teaching sequence for science and how to characterize the type of tasks that are most suitable for peer assessment.

References

Allal, L. (1999). *Impliquer l'apprenant dans le processus d'évaluation: promesses et pièges de l'autoévaluation*. In Depover, C., & Noël, B. (1999). L'évaluation des compétences et des processus cognitifs. Modèles, pratiques et contextes (pp. 35–56).

Ballantyne, R., Hughes, K., & Mylonas, A. (2002). Developing procedures for implementing peer assessment in large classes using an action research process. *Assessment & Evaluation in Higher Education, 27*, 427–441.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice*, *5*, 7e74.

Brown, G. T., Irving, S. E., Peterson, E. R., & Hirschfeld, G. H. (2009). Use of interactive–informal assessment practices: New Zealand secondary students' conceptions of assessment. *Learning and Instruction, 19*(2), 97–111.

Chang, H. Y., & Chang, H. C. (2013). Scaffolding students' online critiquing of expert-and peer-generated molecular models of chemical reactions. *International Journal of Science Education, 35*(12), 2028–2056.

Chang, H. Y., Quintana, C., & Krajcik, J. S. (2010). The impact of designing and evaluating molecular animations on how well middle school students understand the particulate nature of matter. *Science Education, 94*(1), 73–94.

Cho, K., Schunn, C. D., & Charney, D. (2006). Commenting on writing: Typology and perceived helpfulness of comments from novice peer reviewers and subject matter experts. *Written Communication, 23*(3), 260–294.

Crabtree, B. F., & Miller, W. L. (Eds.). (1999). *Doing qualitative research* (pp. 163–177). Sage Publications.

Duschl, R., Schweingruber, H. A., & Shouse, A. W. (Eds.). (2007). *Taking science to school: Learning and teaching science in grades K-8*. Washington, DC: National Academies Press.

Education at a Glance. (2015). OECD Indicators Czech Republic. DOI:10.1787/eag-2015-51

Etkina, E., Karelina, A., Ruibal-Villasenor, M., Rosengrant, D., Jordan, R., & Hmelo-Silver, C. E. (2010). Design and reflection help students develop scientific abilities: Learning in introductory physics laboratories. *The Journal of the Learning Sciences, 19*(1), 54–98.

Falchikov, N. (1996). Improving learning through critical peer feedback and reflection. In *Different Approaches: Theory and Practice in Higher Education. Proceedings of HERDSA Conference*.

Falchikov, N., & Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of Educational Research, 70*(3), 287–322.

Gielen, S., Peeters, E., Dochy, F., Onghena, P., & Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. *Learning and Instruction, 20*(4), 304–315.

Hanrahan, S. J., & Isaacs, G. (2001). Assessing self-and peer-assessment: The students' views. *Higher Education Research & Development, 20*(1), 53–70.

Harlen, W. (2013). *Assessment & inquiry-based science education: Issues in policy and practice*. Trieste: Global Network of Science Academies.

Hovardas, T., Tsivitanidou, O. E., & Zacharias, C. Z. (2014). Peer versus Expert feedback: Investigating the quality of peer feedback among secondary school students assessing each other's science web-portfolios. *Computers & Education, 71*, 133–152.

Husfeld, V. (2009). Aus der Praxis der Leistungsbeurteilung. In D. Fischer, A. Strittmatter, & U. Vögeli-Mantovani (Eds.), *Noten, was denn sonst? Leistungsbeurteilung und -bewertung* (pp. 33–40). Zürich: Verlag LCH.

Kane, R. G., & Chimwayange, C. (2014). Teacher action research and student voice: Making sense of learning in secondary school. *Action Research, 12*(1), 52–77.

Košťálová, H., Miková, Š., & Stang, J. (2008). *Školní hodnocení žáků a studentů se zaměřením na slovní hodnocení* [School assessment of students]. Praha: Portál.

Kratochvílová, J. (2011). *Systém hodnocení a sebehodnocení žáků.* [System of assessment and self-assessment of pupils]. Brno: MSD.

Kronig, W. (2009). Schulnoten–Glasperlen des Bildungssystems. *Noten, was denn sonst*, 27–31.

Laufková, V., & Novotná, K. (2014). *Školní hodnocení z pohledu žáků* [School assessment from the students' perspective] *Orbis scholae*, (*1)*, 111–127.

Le Hebel, F., Montpied, P., & Moulin, M. (2016). Does peer assessment help students' understanding? Paper presented at the International Conference of East-Asian Association for Science Education, Tokyo, Japan.

Le Monde. (2016). *Quand les étudiants se forment entre eux.* http://www.lemonde.fr/campus/article/2016/03/09/quand-les-etudiants-se-forment-entre-eux_4878964_4401467.html

Lukášová, H. (2012). *Proměny pojetí vzdělávání a školního hodnocení* [Changes in conceptualisation of the education and school assessment]. Praha: Asociace waldorfských škol ČR.

McDermott, L. C., Shaffer, P. S., & Rosenquist, M. L. The Physics Education Group.(1996). *Physics by inquiry*.

Narciss, S. (2008). Feedback strategies for interactive learning tasks. In J. M. Spector, M. D. Merrill, J. J. G. Van Merrie¨nboer, & M. P. Driscoll (Eds.), *Handbook of research on educational communications and technology* (3rd ed., pp. 125–143). Mahwah: Erlbaum.

Narciss, S., & Huth, K. (2006). Fostering achievement and motivation with bug–related tutoring feedback in a computer–based training for written subtraction. *Learning and Instruction, 16*, 310–322.

Nicolaou, C. T., & Constantinou, C. P. (2014). Assessment of the modeling competence: A systematic review and synthesis of empirical research. *Educational Research Review, 13*, 52–73.

Novotná, K., & Krabsová, V. (2013). Formativní hodnocení: Případová studie [Formative assessment: case study]. *Pedagogika, 63*(3), 355–371.

Organisation de coopération et de développement économique. (2005). *L'évaluation formative: pour un meilleur apprentissage dans les classes secondaires*. Paris: OCDE.

Pluta, W. J., Chinn, C. A., & Duncan, R. G. (2011). Learners' epistemic criteria for good scientific models. *Journal of Research in Science Teaching, 48*(5), 486–511.

Prins, F. J., Sluijsmans, D. M., & Kirschner, P. A. (2006). Feedback for general practitioners in training: Quality, styles, and preferences. *Advances in Health Sciences Education, 11*(3), 289–303.

Saari, H., & Viiri, J. (2003). A research-based teaching sequence for teaching the concept of modelling to seventh-grade students. *International Journal of Science Education, 25*(11), 1333–1352.

Santiago, P., Gilmore, A., Nusche, D., & Sammons, P. (2012). *OECD Reviews of Evaluation and Assessment in Education: Czech Republic 2012*. OECD.

Schwarz, C. V., & White, B. Y. (2005). Metamodeling knowledge: Developing students' understanding of scientific modeling. *Cognition and Instruction, 23*(2), 165–205.

Schwarz, C. V., Reiser, B. J., Davis, E. A., Kenyon, L., Acher, A., Fortus, D., et al. (2009). Developing a learning progression for scientific modeling: Making scientific modeling accessible and meaningful for learners. *Journal of Research in Science Teaching, 46*(6), 632–654.

Sedová, K., Svarícek, R., & Salamounová, Z. (2012). Komunikace ve skolní tríde [Communication in the classroom]. *Praha: Portál.*

Slavík, J. (2003). Autonomní a heteronomní pojetí školního hodnocení – aktuální problem pedagogické teorie a praxe [Autonomous and heteronomous assessment – current issue of educational theory and practice]. *Pedagogika*, *53*(1), 5–25.

Sluijsmans, D. M. A. (2002). *Student involvement in assessment, the training of peer-assessment skills.* Interuniversity Centre for Educational Research.

Sluijsmans, D. M., Brand-Gruwel, S., & van Merriënboer, J. J. (2002). Peer assessment training in teacher education: Effects on performance and perceptions. *Assessment & Evaluation in Higher Education, 27*(5), 443–454.

Smit, R. (2009). Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz. In *Eine empirische Studie in der Sekundarstufe 1*. Baltmannsweiler: Schneider Verlag Hohengehren GmbH.

Starý, K. (2006). *Sumativní a formativní hodnocení* [Summative and formative assessment]. Portál RVP, www.rvp.cz

Stefani, L. A. (1994). Peer, self and tutor assessment: Relative reliabilities. *Studies in Higher Education, 19*(1), 69–75.

Strijbos, J. W., Narciss, S., & Dünnebier, K. (2010). Peer feedback content and sender's competence level in academic writing revision tasks: Are they critical for feedback perceptions and efficiency? *Learning and Instruction, 20*(4), 291–303.

Strijbos, J. W., & Sluijsmans, D. (2010). Unravelling peer assessment: Methodological, functional, and conceptual developments. *Learning and Instruction, 20*(4), 265–269.

Tiberghien, A. (2011). Conception et analyse de ressources d'enseignement : le cas des démarches d'investigation. In M. Grangeat (Ed.), *Les démarches d'investigation dans l'enseignement scientifique Pratiques de classe, travail collectif enseignant, acquisitions des élèves* (pp. 185–212). Lyon: INRP.

Topping, K. (1998). Peer assessment between students in colleges and universities. *Review of Educational Research, 68*(3), 249–276.

Topping, K. (2013). Peers as a source of formative and summative assessment. In J. H. Mac Millan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 395–412). London: Sage.

Topping, K. J., Smith, E. F., Swanson, I., & Elliot, A. (2000). Formative peer assessment of academic writing between postgraduate students. *Assessment & Evaluation in Higher Education, 25*(2), 149–169.

van Gennip, N. A., Segers, M. S., & Tillema, H. H. (2009). Peer assessment for learning from a social perspective: The influence of interpersonal variables and structural features. *Educational Research Review, 4*(1), 41–54.

Van Steendam, E., Rijlaarsdam, G., Sercu, L., & Van den Berg, H. (2010). The effect of instruction type and dyadic or individual emulation on the quality of higher-order peer feedback in EFL. *Learning and Instruction, 20*(4), 316–327.

Vögeli-Mantovani, U. (1999). SKBF Trendbericht Nr. 3: Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz. Aarau: Schweizerische Koordinationsstelle für Bildungsforschung.

Walker, A. (2001). British psychology students' perceptions of group-work and peer assessment. *Psychology Learning & Teaching, 1*(1), 28–36.

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200.

Žlábkova, I., & Rokos, L. (2013). Pohledy na formativní a sumativní hodnocení žáka v českých publikacích [Formative and summative assessment in Czech publications]. *Pedagogika, 58*(3), 328–354.

Chapter 7
Written Teacher Feedback: Aspects of Quality, Benefits and Challenges

Monika Holmeier, Regula Grob, Jan Alexis Nielsen, Silke Rönnebeck, and Mathias Ropohl

Introduction

Formative assessment is reported to have positive effects on student competences (e.g. Black and Wiliam 1998; Hattie 2009; Kingston and Nash 2011; Kluger and DeNisi 1996). It has therefore been suggested as a means of improving teaching practice (e.g. OECD 2005, 2013). Formative assessment has led to changes in a number of educational systems, at national or federal state level, with the expectation of being incorporated into everyday teaching (e.g. D-EDK 2014; Ministerium für Schule und Berufsbildung 2016; Nielsen and Dolin 2016; Vögeli-Mantovani 1999).

A crucial element of formative assessment is to provide the students with feedback on their next steps in learning so that they can reach a specific learning goal (Hattie and Timperley 2007). Feedback can be given in oral form (e.g. in the case of on-the-fly feedback; see Chap. 4 in this book) or in written form (e.g. comments in a science notebook). It can be provided by the teacher or by student peers. The study presented in this chapter concentrates on written feedback provided by the teacher.

In the three countries that were part of this study, there has been very little to no research conducted on written feedback in the context of science education and

M. Holmeier (✉) • R. Grob
Centre for Science and Technology Education, University of Applied Sciences and Arts Northwestern Switzerland, Basel, Switzerland
e-mail: monika.holmeier@fhnw.ch

J.A. Nielsen
Department of Science Education, University of Copenhagen, Copenhagen, Denmark

S. Rönnebeck
Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

Kiel University, Kiel, Germany

M. Ropohl
Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_7

inquiry-based learning in general or, more specifically, on its quality (e.g. Köller 2005). Descriptive results from PISA 2015 indicate that students from Germany and Switzerland generally receive a low extent of feedback from their teachers (Schiepe-Tiska et al. 2016). If they do receive feedback, it is often in the form of simple and short phrases like 'yes', 'no' or 'that's right' (Kobarg and Seidel 2007). Furthermore, science teachers often miss opportunities to give feedback (Ruiz-Primo et al. 2004). Possible reasons for this are that teachers do not value the importance of feedback for student learning, that they do not notice situations where feedback would be beneficial or that they do not know effective ways of giving feedback.

Against this background, this study examines to what extent teachers in science classrooms in Germany, Switzerland and Denmark can provide written feedback of high quality based on specific tools. The quality of written teacher feedback was analysed in the context of inquiry-based science teaching and learning. Usually, inquiry is conceived as various activities conducted by students (Abd El Khalick et al. 2004; Linn et al. 2004; Minner et al. 2010; Rönnebeck et al. 2016). In this exploratory study, a number of these activities were assessed: in Germany, the planning of experiments with a specific focus on the control-of-variables strategy; in Denmark and Switzerland, the planning and carrying out of experiments and the interpretation and evaluation of data and results; and in two further Danish cases, the activities included design and modelling in the discipline of technology.

The first part of this chapter describes, from a theoretical perspective, written feedback as one approach of giving feedback within formative assessment. The second part of the chapter presents case studies from Germany, Switzerland and Denmark in which the quality of written teacher feedback in science teaching will be explored. Further, the chapter provides results of interviews with the participating teachers on the perceived benefits and challenges of giving written feedback. Based on these findings, conclusions will be derived for further research and practical implications.

Written Feedback

In science teaching, written feedback is generally based on a student's written artefacts such as science lab journals, reports or written answers to given tasks. Written feedback can be provided by peers or by the teacher, with the latter being the focus of this chapter. The present work concentrates on the quality of written teacher feedback.

Teacher Prerequisites for Giving Feedback

Formative assessment has been identified as a promising way to support student learning (Black and Wiliam 1998). It is based on the analysis of a student's level of attainment with the aim of providing feedback about his or her learning and of planning and implementing activities that improve student learning (Bennett 2011). Formative assessment is a challenging part of teacher practice. It requires deep content knowledge, targeted attention, noticing student learning progress and a rich repertoire of teaching strategies for effective action (Furtak et al. 2008). Indeed, the impact of formative assessment on student learning depends on the teacher's ability to focus on the basic ideas behind a student's response, to analyse possible misconceptions and to generate productive interpretations about the major challenges that students face (Bennett 2011; Sadler 1989). Thus, teachers should not only be able to judge whether a student's answer is right or wrong but also notice ideas expressed by a student that may hinder learning or foster it (Furtak 2012; Levin et al. 2009; Russ et al. 2009). The inferences that teachers make are critical in the assessment process as they lay the foundation for his or her subsequent decisions and actions.

Empirical results indicate that science teachers struggle to pay attention to the substance of student thinking (Hammer and van Zee 2006; Roth et al. 2011; Russ et al. 2009). Furthermore, teachers frequently evaluate student ideas based on how well a student's answers meet predefined learning goals in curricula, undervaluing other productive ways of reasoning (Russ et al. 2009). One reason for these problems could be that the development of science teachers' competence in noticing has not been traditionally the focus of teacher education and professional development programmes (Coffey et al. 2011; Stiggins 2002). However, programmes that focused on the analysis and interpretation of student work have proven to be successful in fostering knowledge and knowing how to respond to a student's learning needs (Ball and Forzani 2011; Gearhart et al. 2006; Kazemi and Franke 2004; Love et al. 2008).

Quality of Written Feedback

The feedback model from Hattie and Timperley (2007) suggests that the purpose of feedback is 'to reduce discrepancies between current understandings/performance and a desired goal' (p. 87). In other words, the purpose of feedback is to close the gap between the student's current level and the learning objectives. Therefore, feedback should provide the students with information concerning three major questions: 'Where am I going? (What are the goals?), How am I going? (What progress is being made toward the goal?), and Where to next? (What activities need to be undertaken to make better progress?)' (Hattie and Timperley 2007, p. 86). Other authors have identified four criteria for what constitutes *effective* feedback (e.g. Arts et al. 2016 based on Nicol and Macfarlane-Dick 2006 and Gibbs and Simpson 2004; Brookhart 2008; Harks et al. 2014):

1. Effective feedback is linked to predefined assessment criteria which are an operationalisation of the goal(s) and to an individual reference norm.
2. Effective feedback makes clear what the expected learning goals are, informs the recipient on his or her current level of achievement and includes information on how to reach the expected learning goals. In this context, it is important that the student is not only told whether the assignment was solved correctly or not but also why. The teacher has to justify his feedback (Arts et al. 2016; Glover and Brown 2006; Orsmond and Merry 2011).
3. Effective feedback reassures motivational beliefs and self-confidence and aids the recipient in adequate self-assessment (Nicol and Macfarlane-Dick 2006).

4. Effective feedback is complete, contains enough detail and is clearly formulated (Brookhart 2008; Gibbs and Simpson 2004). Moreover, it is given timely, i.e. during the learning process when students are still mindful of the task. The optimal timing, however, depends on the task and on the students' level of proficiency (Brown et al. 2012).

Written feedback can be provided based on rubrics (Andrade 2005; Arter and McTighe 2001; Burke 2006; Moskal 2003; Panadero and Jonsson 2013) or on open comments (e.g. Black and Harrison 2004). Typically, rubrics are a closed format that encompasses descriptions of competence levels. Through ticking checkboxes, the teacher indicates an individual student's competences. Open comments are also based on competence levels, but the teacher gives detailed feedback in his or her own words especially focusing on hints for learning and justifications. In many cases, the two formats are combined so that the teacher gives feedback with rubrics (which includes the learning goals) as well as with open comments.

Rubrics as a Means of Feedback

Rubrics articulate the expectations for a learning goal or an assignment, i.e. the criteria that define what competences students should demonstrate at various levels of competence (Andrade 2000; Smit and Birri 2014). These levels can be defined using either quantitative (i.e. numerical, 1–4; beginning, developing, accomplished, exemplary) or qualitative (i.e. one or two whole sentences) information. In this sense, rubrics are a tool for the teacher to diagnose and assess a student's competence and level of attainment. Additionally, they can be a tool for the teacher to communicate the learning goals and the current state of learning to the student. The teacher can hand out the completed rubric to the students or use it as a basis for written feedback. The advantage for students is that they get a detailed overview of where they are related to the learning goals which is a prerequisite in deciding about the next steps in learning (Andrade 2000).

A number of studies focus on effective feedback in the form of filled-out rubrics and the mechanisms through which rubrics support learning. Rubrics provide students with both, information on their current level of learning (level of competence that is marked on the instructional rubric) and information on where they are going (highest level of competence on the rubric). It has been documented that the use of rubrics is meaningful only if a number of prerequisites are met: there must be clear, tangible goals that are available to the students, and there must be a student activity which is suitable for assessing those goals (Jonsson 2014; Luft 1999; Moskal 2003). Rubrics are found to be particularly suitable for providing feedback on student performance in authentic tasks and highly contextualised activities such as inquiry (Darling-Hammond et al. 1995; Ni 1997; Wiggins 1998). Furthermore, the content and use of the instrument has to be explained to students in order to make it a useful tool for learning (Andrade and Du 2005; Moni and Moni 2008).

Regarding the mechanisms through which rubrics support learning, Panadero and Jonsson (2013) claim in their meta-analysis that there are several ways: (1)

increasing transparency, (2) reducing anxiety, (3) aiding the feedback process, (4) improving student self-efficacy and (5) supporting student self-regulation. Students use the feedback provided by rubrics on a task-based level and, to some extent, to structure the progress of their work.

Another focus in the research literature is the perspective of the teachers. From their point of view, rubrics help to clarify their expectations (Andrade 2005). This benefit is twofold: on the one hand, it helps the teacher to plan the instruction; on the other hand, it helps the students to be clear about the expectations they should meet. Luft summarises this twofold benefit as 'reflective practice among students and instructor' (Luft 1999, p. 114). However, this perspective does not seem to be true for all teachers: both So and Lee (2011) and Bharuthram (2015) found that many teachers and lecturers from different school levels perceive rubrics as mostly useful for students but not for themselves. A possible reason for this could be that the rubrics are used in a rather unconscious manner and without using the full range of possibilities and potential of the tool, which would include using the feedback in order to plan subsequent instruction activities. From the perspective of teachers and lecturers, there can also be disadvantages using rubrics. For example, rubrics need time before the criteria become clear. Just handing out the rubric is also not appropriate. Rather, an exchange between the class and the teacher is needed. Therefore, reported disadvantages typically include issues related to time and the difficulty to formulate the criteria in a rubric in a way that is understandable to students (e.g. Bharuthram 2015; Luft 1999).

Open Comments as a Means of Feedback

A rubric-based diagnosis and assessment by competence level descriptions can be accompanied by writing open comments. By using open comments, the teacher can precisely describe individual problems and specific strengths of a piece of work (Black et al. 2003). Compared to rubrics, it is easier to give guidance on how to make improvements towards the learning goals and to show concrete steps in learning (Nunes 2004; Santos and Dias 2006; Stracke and Kumar 2010). Therefore, open comments can be used for formative assessment if they answer the three already mentioned questions by Hattie and Timperley (2007) based on an underlying description of competence levels in a running text.

In a study in language education, Parr and Timperley (2010) showed that the effect of open comments on student learning depends on the characteristics of the open comments. The problem is that feedback could be ignored, misunderstood or misinterpreted by the students due to its poor quality (Marzano and Arthur 1977; Searle and Dillon 1980). This is true for simple and short feedback like 'ok' or 'not complete'. Therefore, it is important to keep general guidelines for high-quality feedback in mind. Brookhart (2008) gives suggestions of how to formulate effective written feedback that could be given in open comments:

- Clarity: effective written feedback should use simple vocabulary and sentence structure. Furthermore, teachers should write according to the student's developmental level and also check that the student understands the feedback.
- Specificity: effective written feedback should give the students guidance without giving the solution away. Neither should it be too narrow nor too broad. This should be done by providing the students with suggestions which are specific enough (e.g. describing useful learning strategies) so that the students can take concrete next steps.
- Tone: effective written feedback should be based on words that motivate students and show them that they are active learners. The right word choice helps to communicate respect for the students and can also inspire the students.

Bailey and Garner (2010) investigated open comments as one type of written feedback and found that teachers have different beliefs about its purpose. Furthermore, the teachers seem to have difficulties in combining their two roles (supporting and assessing) when providing written feedback. Similar results are published by Tuck (2012). For science education, Bruno and Santos (2010) focus on teacher difficulties when providing open comments. The teachers participating in their study mentioned two challenges, namely, how to select what to comment on from the many issues that could be addressed and how to avoid giving away part of the answer but still provide useful guidance on the next steps. In addition, other findings on the nature of teacher feedback suggest that teachers often focus their assessment efforts on deciding whether a student's work is right or wrong, whereas describing how and why the work is correct or not and identifying what should be done to improve it both rarely occur (Hattie and Timperley 2007; Ruiz-Primo and Li 2013).

Research Gap and Research Questions

Based on the preceding theory, the following conclusions in view of the quality of written feedback can be drawn if the feedback is given by means of rubrics and open comments:

- Rubrics are a valuable tool for diagnosing and assessing student levels of competence. Additionally, they can be used for providing written feedback because they allow the teacher to communicate the learning goals and the current state of learning to the student. This is a prerequisite so that the student can decide about the next steps in learning.
- Open comments can also be used to communicate feedback to students. They are particularly effective when focusing on competences rather than superficial aspects (such as spelling or layout), when they include information to students on how to proceed and when it is formulated in an easily understandable language. The difficulties related to open comments are how to select what to comment on and how to avoid giving away part of the answer in the feedback.
- A combination of both can be seen as the most desirable way of generating written feedback because it fulfils a basic requirement: the assessment of student

competences is based on objective criteria, communicated to the students with comments showing where to go and how to go there.

However, empirical research in science education on the use and quality of written feedback in general or, more specifically, on the use and quality of rubrics and open comments is still scarce. Existing studies investigate the effects of feedback on student learning without focusing on teacher competences (e.g. Wollenschläger et al. 2011) or teacher beliefs about assessment (e.g. Brown et al. 2012). In a recent study, Furtak et al. (2016) investigated the changes in the quality of oral teacher feedback over 3 years. In total, the quality and quantity of the oral feedback improved over time. Teachers more often gave feedback which promoted student thinking, e.g. by asking students to elaborate their responses or by asking for more information instead of just saying ‘right’ or ‘good job’ (Furtak et al. 2016). Similar investigations of written feedback are still missing in the field of science education. In order to fill this gap, the present study explores the quality of written feedback based on rubrics and open comments. The exploratory study will focus on teachers from three European countries – Germany, Switzerland and Denmark. The results will give insights into science teacher competences of giving feedback as well as into the benefits and challenges seen by teachers when implementing written feedback in teaching practice. In addition, from the findings it will be possible to deduce hypotheses for further research projects. The following research questions are posed: (1) What is the quality of written feedback that is provided based on rubrics and templates for open comments? (2) Which benefits and challenges do teachers see in using rubrics and open comments for written feedback?

Research Design

In order to answer the two research questions, an exploratory study was designed to identify the quality of written feedback that was given based on rubrics and templates for open comments. The aim of the study was to facilitate the effective implementation of written teacher feedback by supporting teacher attempts to interpret student data, to diagnose difficulties and needs and to provide students with effective written feedback. In all three countries, case studies of written feedback in the context of scientific inquiry were analysed using a mixed methods approach: the quality of written feedback was analysed quantitatively by using the teachers’ open comments, whereas the benefits and challenges were analysed qualitatively using data from interviews. The cases were part of an intervention without comparison or control group.

In the intervention, the teachers were provided with theoretical descriptions of written feedback and with examples of inquiry-based units with integrated formative assessment. The teachers were also provided with rubrics. During the course of eight regular meetings across a period of 1.5 years (between October 2014 and December 2015), the teachers were introduced to the concept of formative assessment and to the relevance of feedback. Furthermore, the teachers were familiarised with different approaches on how to incorporate the provision of feedback in their teaching units. Afterwards, the teachers were asked to implement written feedback in their regular inquiry-based science lessons in order to formatively assess the competences of their students. At a specific point during the inquiry-based unit, students submitted certain artefacts to the teacher. The teacher provided written feedback to each student based on a rubric (examples for all three countries are given in the results section). For this, the teacher used specially designed tools provided to facilitate his or her attempt to diagnose a student's needs or difficulties, with regard to the competence under emphasis. Detailed implementations are described in country-specific sections below, as they vary slightly between the different countries due to different study conditions, such as science subject, topic, grade level and type of written feedback (see Table 7.1).

Table 7.1 Overview of the sample

Case	RI	EL	S	Topic	Type	Stud. [*N*]
Germany						
G1	1	Lower sec.	Ch	Metals	Open comments[a]	23
G2	1	Lower sec.	Ch	Metals	Open comments[a]	6
G3	2	Lower sec.	Ch	Salts	Open comments[a]	22
G4	2	Upper sec.	Ch	Washing detergents	Open comments[a]	16
G5	3	Lower sec.	Ch	Salts	Rubrics and open comments	24
Switzerland						
S1	2	Primary	InS	Growth of chicks	Open comments	21
S2	2	Upper sec.	Bio	Ecology	Rubrics and open comments	15
S3	3	Upper sec.	Phy	Electric circuits	Open comments	21
S4	3	Upper sec.	Phy	Science in the city	Rubrics and open comments	23
Denmark						
D1	1	Lower sec.	InS	Indoor climate	Open comments	47 in 29 groups
D2	1	Upper sec.	Tec	Electrical circuits	Open comments	21 in 6 groups
D3	1	Upper sec.	Tec	Electrical circuits	Open comments	26 in 7 groups
D4	1	Lower sec.	InS	Indoor climate	Open comments	15 in 6 groups
D5	1	Upper sec.	Bio	Respiration	Open comments	21 in 5 groups
D6	3	Upper sec.	Bio	Blood sugar regulation	Open comments	6 groups[b]
D7	3	Lower sec.	InS	Human nutrition	Rubrics and open comments	17
D8	3	Lower sec.	InS	Human nutrition	Open comments	19
D9	3	Upper sec.	Bio	Physical fitness rating	Rubrics and open comments	28 in 7 groups
D10	3	Upper sec.	Bio	Blood sugar regulation	Rubrics and open comments	22

RI round of implementation, *EL* educational level, *S* subject, *Stud.* [*N*] number of students, *Sec.* secondary, *Ch* chemistry, *Bio* biology, *Phy* physics, *InS* integrated science, *Tec* technology

[a] Open comments were offered using a template that also listed the learning goals for the students

[b] The number of students could not be exactly calculated since the teacher and the students only wrote the group number in the raw data that was accessible to the researchers

The sample was constituted on a voluntary basis. In order to find teachers participating in the study, different methods were applied (e.g. via personal contacts, email lists and advertisement in journals for science teachers). The teachers received financial compensation for the additional workload. In the end, the sample encompassed three chemistry teachers from Germany (lower and upper secondary school), four science teachers from Switzerland (primary and upper secondary school) and seven science and technology teachers from Denmark (lower and upper secondary school). The *N* = 14 teachers produced 19 implementations of written feedback which will be referred to as cases. The teachers' collaboration in the study lasted three semesters which will be referred to as rounds of implementation.

In order to explore the two research questions, two types of data were collected. For research question 1, the initial and the revised versions of the student artefacts as well as the written feedback provided by each teacher were collected. These data were analysed quantitatively using a specifically developed coding tool with 15 items. In each country, the coding was undertaken by two independent coders. Reliability was secured by applying the Wilcoxon test and/or Krippendorff's alpha on a subsample of each data set. As described below, while it was possible to achieve a satisfactory reliability rating in most of the cases, some items of the Danish implementation were not able to be coded reliably even after many iterations. Reasons for this are described in the country-specific results section for Denmark.

Five items of the coding tool were used for analysing the quality of written teacher feedback and for answering research question 1. These are:

1. To what extent did the feedback comments take into account the learning goals which had been provided by the teacher beforehand?
2. To what extent did the teacher justify his or her references to what the student had *already* achieved with regard to the targeted competence?
3. To what extent did the teacher justify his or her references to what the student had *not yet* achieved with regard to the targeted competence?
4. To what extent does the feedback provide specific guidance to the student about concrete next steps?
5. To what extent did the student address the feedback on the first artefact in his or her second artefact?

For research question 2, the teachers were asked to fill out a self-reporting questionnaire in which they reflected upon the usability of written feedback in their daily teaching practice. The self-reporting questionnaire included five open-ended questions; two of these are included in this study. These are:

1. What are the main strengths and weaknesses of written feedback within formative assessment?
2. What opportunities and challenges do you identify in enacting written feedback?

The responses by teachers were analysed using open coding and qualitative content analysis (Mayring 2004).

Analysis of Written Teacher Feedback

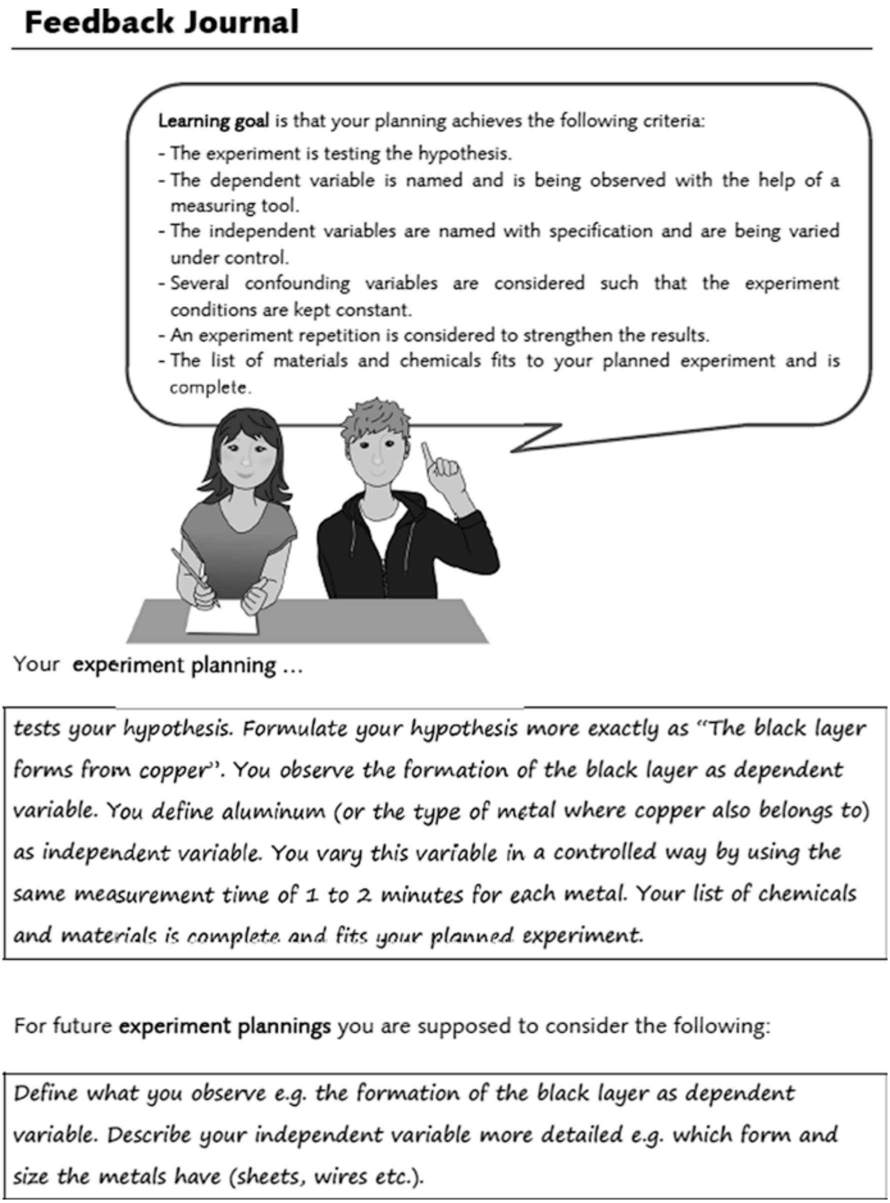
Germany

Description of the Implementations

In Germany, written feedback was implemented in inquiry units that the participating teachers developed in cooperation with researchers from IPN during regular project meetings. All implementations were related to the inquiry competence of planning an investigation with a specific focus on the control-of-variables strategy (CVS) in chemistry classrooms. The implementations were carried out in three rounds over 1.5 school years. In between the rounds, the teachers discussed their experiences and the suitability of the instruments and suggested improvements. In general, all implementations followed a similar set-up. In an introductory lesson, the teacher introduced the learning goals, i.e. the CVS strategy and criteria for good plans of scientific experiments to the students. In the following lesson, the students were asked to individually plan an experiment related to either a given or a self-developed hypothesis. For writing down their plans, they used a template called the 'scientist journal'. The teacher collected the written experimental plans and evaluated each student's performance using a rubric. For each of nine sub-competences of planning an investigation (such as 'the experiment tests the hypothesis' or 'the independent variable is varied under controlled conditions'), the rubric describes the student's expected performance at three potential levels (0 = criteria not addressed; 1 = criteria partly addressed; 2 = criteria fully addressed). The teacher gave feedback to the students using a second template, the 'feedback journal' (see Fig. 7.1). The feedback journal gives feedback to the students concerning their current state of learning (i.e. what they have already achieved or not achieved) and next steps (i.e. what they should change or consider when planning their next experiments). In addition, for all implementations, it displayed the learning goals as a reminder for the students. The students read the feedback and were then asked to plan another experiment.

The implementations differed with respect to the round of implementation, the school level and the chemistry topic (all described in Table 7.1). In the following, the five cases are described in detail.

The first and second implementations (G1 and G2) were carried out during the first implementation round, in an eighth and a ninth grade chemistry classroom, respectively. The topic was metals. In the scientist journal, the students were asked to plan an experiment related to a given hypothesis by using an open-answer format ('To test the hypothesis, I am planning the following experiment: …'; see Fig. 7.1). The teacher gave feedback in the feedback journal using open comments. The feedback journal asked specifically to give feedback related to the current state of learning and next steps ('In your experimental design you…' and 'Regarding future experimental designs you should consider the following…'; see Fig. 7.1). The procedure was repeated for two more experiments.



Feedback Journal

Learning goal is that your planning achieves the following criteria:

- The experiment is testing the hypothesis.
- The dependent variable is named and is being observed with the help of a measuring tool.
- The independent variables are named with specification and are being varied under control.
- Several confounding variables are considered such that the experiment conditions are kept constant.
- An experiment repetition is considered to strengthen the results.
- The list of materials and chemicals fits to your planned experiment and is complete.

Your **experiment planning** ...

tests your hypothesis. Formulate your hypothesis more exactly as "The black layer forms from copper". You observe the formation of the black layer as dependent variable. You define aluminum (or the type of metal where copper also belongs to) as independent variable. You vary this variable in a controlled way by using the same measurement time of 1 to 2 minutes for each metal. Your list of chemicals and materials is complete and fits your planned experiment.

For future **experiment plannings** you are supposed to consider the following:

Define what you observe e.g. the formation of the black layer as dependent variable. Describe your independent variable more detailed e.g. which form and size the metals have (sheets, wires etc.).

Fig. 7.1 Translated example of a teacher feedback journal (original language German). The feedback refers to the planning of an experiment to answer the question: what is the black layer that is formed if you heat a piece of copper foil in the flame of a Bunsen burner?

The third and fourth implementations (G3 and G4) took place in the second implementation round. Teacher G3 taught an inquiry unit on salts in an eighth grade and teacher G4 a unit on washing detergents in a twelfth grade chemistry classroom. Based on teacher feedback from the first implementation round, some scaffolding was introduced into the scientist journal to support students in addressing all relevant aspects of an experimental plan. Students were now asked to address the different aspects using specific questions, e.g. 'Name the phenomenon that you want to investigate in your experiment. This is your dependent variable'. The feedback journal was identical to the one used in G1 and G2. Having received feedback regarding their first experimental plan, the students planned another different experiment.

The fifth implementation (G5) was undertaken in the third implementation round, in an eighth grade chemistry classroom. The topic was again salts (the inquiry unit was the same as in G3). For the scientist journal, again the scaffolded version (used for G3 and G4 in the second implementation round) was used. However, in this implementation round, the structuring of the feedback journal was increased. Instead of using an open-answer text box to comment on the current state of learning, the teachers could now tick a checkbox for each of the sub-competences indicating which competence level the student has already achieved. Justifications for the assessment as well as feedback concerning next learning steps were still given as open comments. Again, students planned a second different experiment after having received and read the feedback concerning their first experimental plan.

Quality of Written Feedback Provided by Teachers

Table 7.2 displays the percentages of assessment criteria that were addressed in the feedback comments and the degree to which the teachers justified their references. The results show that all of the teachers were able to address the assessment criteria to a high extent. Only in the first implementation round, the overlap between learning goals and comments in the feedback was slightly worse, which may be due to the fact that the teachers first needed to become familiar with the concept of formative assessment and the instruments. Across the five cases, no preference for one of the two forms of the feedback journal could be observed.

With respect to the justifications offered by the teacher, it can be observed that teachers seem to be more likely to give justifications for learning goals that have not yet been achieved by the students. One reason for this could be that they see a greater need for justifications when the student still needs help for improvement. Consequently, teachers might have felt that a feedback comment such as 'You have correctly identified the dependent variable' needed no further clarification. Nevertheless, differences between cases as well as variation within cases can be observed. Since all of the teachers used very similar instruments, these differences might be seen as an indication that the ability to give written feedback depends not only on the individual teacher but is also related to the nature and quality of the student artefacts they are based upon.

Table 7.2 Overview on the quantitative results from Germany in terms of diagnosis

	Assessment of level of attainment					Justification offered by the teacher about student's level of attainment											
	Addressing assessment criteria [%][a]					Justification on what has been achieved [%][b]						Justification on what has not yet been achieved [%][b]					
	1	2	3	4	5	0	1	2	3	4	5	0	1	2	3	4	5
G1			4	96			35	48	13	4		4		4		43	52
G2		33	17		50		83	17				17	50			17	17
G3					100				32	68					18	36	45
G4				56	44		69	13	6	13			19	13	25	44	
G5					100	25	71	0	4				8	21	29	38	4

[a] 1 = none; 2 = some; 3 = half; 4 = most; 5 = all. Percentage calculated per class, see last column of Table 7.1 for exact number of students per class

[b] 0 = not applicable (no such comments); 1 = without; 2 = mostly without; 3 = balanced; 4 = mostly justified; 5 = all justified. Percentage calculated per class, see last column of Table 7.1 for exact number of students per class

Table 7.3 Overview on the quantitative results from Germany in terms of provision of guidance about next steps in learning

	Provision of guidance about concrete next steps [%][a]					Addressing feedback in second artefact [%][b]				
	0	1	2	3	4	0	1	2	3	4
G1			26	70	4		35	30	26	9
G2		17	50	33			100			
G3	9			23	68	9	9	14	41	27
G4	6			19	75	6	6	63	19	6
G5				33	67		4	50	33	13

[a] 0 = not applicable (no guidance necessary); 1 = non-specific guidance; 4 = very specific guidance. Percentage calculated per class, see last column of Table 7.1 for number of students per class

[b] 0 = not applicable (no improvement necessary); 1 = none; 4 = all. Percentage calculated per class, see last column of Table 7.1 for number of students per class

Table 7.3 displays the second part of the quantitative data from Germany: to what extent do teachers provide specific guidance to their students on concrete next learning steps and to what extent do students make use of the feedback in producing their second artefact?

The data shows that the teachers in all cases provided specific guidance about next steps in learning. In addition, except for G1, the variation within cases is small which indicates that the teachers provided guidance at the same level of concreteness regardless of the achievement of the students. The greater variance for G1 is probably again due to the teacher's unfamiliarity with the approach and the materials. The picture is much more heterogeneous for student usage of the feedback. Here, except for G2, no patterns can be observed. G2 is a special case since the sample size is much smaller than in the other implementations. Nevertheless, the low level of concreteness in the feedback comments might contribute to student problems in making use of the feedback comments. However, in general the results

Table 7.4 Benefits of written feedback as perceived by the German teachers

Advantages related to the learning of the students	The feedback can be individualised
	Students are supported in the individual development of their competences
	Students can work on their weaknesses
	Students can engage with their work outside the classroom
Advantages related to motivational aspects	Individual feedback increases student motivation
Advantages for the teacher	Has potential for summative assessment
	Allows for the assessment of misconceptions
	Gives the teacher a detailed picture of individual students' competences which supports the teacher in planning his or her instruction and allows for an adaptation of the instruction (by, e.g. differentiating or giving different tasks to different students)

Note. Teachers were asked to reflect on their experiences regarding the implementation of formative assessment and especially feedback. From the answers, it became obvious that the teachers conflated the two terms to some extent. When they reflected on the advantages for the teacher they mainly referred to formative assessment

show that although the teachers provide a lot of guidance on next steps, students do not seem to be able to make efficient use of this guidance.

Benefits and Challenges Seen by Teachers in Using Written Feedback

In general, the teachers regarded written feedback as helpful for both teachers and students. The specific benefits that they saw are displayed in Table 7.4. They can be allocated to three categories: (a) advantages related to student learning, (b) advantages related to motivational aspects and (c) advantages for the teacher. In general, the teachers regarded the individuality, the depth and the detailedness of the information that the written feedback provided as the major benefit. It supports students in the individual development of their competences. For the teachers, it provides information about each individual student and allows them to plan and adapt their instruction according to their students' needs. One teacher said: 'As a teacher I get information about each individual student that I would not get if students work on their experimental plans in groups right from the beginning'. Contrary to the results from Switzerland and Denmark (see next sections), the teachers made no reference to benefits related to the provision of clearly defined learning goals and assessment criteria. One reason for this could be that the teachers were provided with the assessment criteria, i.e. their development was not a specific focus of the intervention in Germany. One interesting result from the qualitative analysis, however, is that the teachers were confident that their students would be able to interpret and use the feedback that they got. This is to some extent in disagreement with the quantitative results presented in Table 7.3, which showed that students seemed not to be able to make efficient use of their teacher's guidance.

Table 7.5 Challenges of written feedback as perceived by the German teachers

Practical challenges	Managing the time requirements and the workload associated with written feedback
	Ensuring that the feedback is not too delayed (due to the organisation of science instruction in schools, students often will not receive their feedback until the following week)
	Integrating written feedback into regular teaching practice
	Finding suitable materials/teaching units that allow for the formative assessment of inquiry competences
	Increasing the acceptance of the method by students, parents and colleagues but also educational policy
Challenges related to teacher assessment literacy	Providing feedback of high quality over time and not changing the reference norm from student to student
	Providing effective feedback without giving away the correct answer

The teachers in Germany reported challenges related to practical issues (see Table 7.5). The most important practical challenge they saw was the amount of time and work that giving written feedback requires and that this impedes its integration in regular teaching practice. Moreover, the students often received a delayed feedback response. One teacher said:

> *The students get their feedback considerably after they have written their experimental plans, in my class for example one week later since students have only one double lesson [chemistry] per week. I think feedback that is given immediately after students have planned their experiments is more effective since at that moment they still know exactly what they have written and why.*

Eventually, teachers struggled with finding or developing adequate experiments. They realised that typical student activities in their instruction often do not require an application of the CVS strategy.

In contrast to the Swiss and Danish results (see next sections), however, the second category of challenges is more related to aspects of the teacher assessment literacy. For instance, teachers found it difficult to ensure a high reliability or to formulate the feedback in a way that it is helpful for students but does not give away the correct answer.

Switzerland

Description of the Implementations

In Switzerland, four teachers implemented written feedback in their regular inquiry-based science units. The first teacher (S1) implemented written feedback at a primary school. The topic of the unit in integrated science was the growth of chicks. For this, the teacher had living chicks in her classroom and let the students describe these chicks in the so-called chick-journal every morning over the course of a month. The students focused on a specific aspect of physical appearance or behaviour of the chicks. The observations were noted down individually, and the teacher collected the chick-journals after every science lesson in order to provide written feedback. She did not use a standardised feedback form for this but instead wrote prompts, questions and comments directly in the students' journals. The next day, the students would have time to read the teacher's comments and would afterwards continue observing and describing. The students knew the two assessment criteria from the very beginning, namely, (1) distinction between observations and assumptions and (2) precise descriptions and sketches. The teacher concentrated on these two criteria throughout the unit. These two criteria are considered part of the competence 'planning and carrying out investigations'.

The second teacher (S2) teaches biology at upper secondary school level. She implemented written feedback in a unit on ecology in which her students investigated the ecological quality of a small river taking into account different factors. The students had to present the results of their investigations and the conclusion on the ecological quality in a written report. They handed in a preliminary version of that report and received written feedback from the teacher. This feedback could be taken into account for the final version of the report. For the written feedback, the teacher used a standardised template known to the students from the very beginning of the unit, in which she listed different aspects of 'presentation of results' and 'interpretation of results' in the form of an assessment rubric, followed by an open field for recommendations and comments on each of these aspects.

The third teacher (S3) implemented written feedback in a unit on electrical circuits in physics at upper secondary level. The students wrote lab reports on a series of experiments. The teacher provided written feedback in the form of open comments on these lab reports so that this feedback could be taken into account for the next similar lab report. The feedback was focused on the precise description and sketches of the experimental setting which was made transparent to the students from the beginning of the unit.

The fourth teacher (S4) integrated written feedback in a physics unit on science in the city. The students had to come up with solutions on how to measure, such as the height of a building or the length of a bridge. As auxiliary tools, they had a measuring tape and an app to measure angles. For the formative assessment, the teacher also used a standardised feedback form with a rubric and open space for comments and recommendations. The students were familiar with the content of this form which was focused on the 'presentation' and 'interpretation of results' in the preliminary version of the report. The students then had time to consider the written feedback for the final version of their reports.

Quality of Written Teacher Feedback

Table 7.6 shows the results on each teacher's ability to use his or her respective feedback tools for diagnosing the levels of attainment of their students for the relevant competences. It can be seen that in all four cases, the teachers were able to

Table 7.6 Overview on the quantitative results from Switzerland in terms of diagnosis

	Assessment of level of attainment					Justification offered by the teacher about student's level of attainment									
	Addressing assessment criteria [in %][a]					Justification on what has been achieved [in %][b]					Justification on what has not yet been achieved [in %][b]				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
S1			4		96	43	57				33	67			
S2					100	93	7				13	53	20	13	
S3		10		42	48	90	10					90	5	5	
S4					100	100					91	9			

[a] 1 = none; 2 = some; 3 = half; 4 = most; 5 = all. Percentage calculated per class, see last column of Table 7.1 for number of students per class

[b] 1 = without; 2 = mostly without; 3 = balanced; 4 = mostly justified; 5 = all justified. Percentage calculated per class, see last column of Table 7.1 for number of students per class

Table 7.7 Overview on the quantitative results from Switzerland in terms of provision of guidance about next steps in learning

	Provision of guidance about concrete next steps [in %][a]				Addressing feedback in second artefact [in %][b]			
	1	2	3	4	1	2	3	4
S1		5	85	10				100
S2	7	27	47	19	No data[c]			
S3				100	5	5	60	29
S4			48	52				100

[a] 1 = non-specific guidance; 4 = very specific guidance. Percentage calculated per class, see last column of Table 7.1 for number of students per class

[b] 0 = not applicable (no improvement necessary); 1 = none; 4 = all. Percentage calculated per class, see last column of Table 7.1 for number of students per class

[c] no data because the second artefact was not available to the authors of this chapter

address the assessment criteria to a high extent. For cases S2 and S4, where rubrics were used, the teachers addressed all criteria for all students. In cases S1 and S3, the teachers worked with open comments only, but they still covered the assessment criteria in their feedback – even though the percentage is lower than where rubrics were used.

Looking at the justification of the assessment that was offered by the teachers, the results show that this does not seem to be a main component of the feedback. The variability within the cases is generally rather low which gives the impression that the degree to which the assessment is justified depends on the teachers' habits rather than on a specific assessment tool.

Table 7.7 displays the second part of the quantitative data from Switzerland, namely, the teachers' ability to provide advice for students about how to proceed in their learning. The data shows that the teachers were in all cases able to provide some guidance about next steps in learning. The variation within cases is small, which suggests that the teachers provided guidance at the same level of concreteness independent of the achievement of the students. An exception to this is case S2

Table 7.8 Benefits of written feedback as perceived by the Swiss teachers

Advantages related to clearly stated assessment criteria	Students and parents know the assessment criteria from the beginning and know what the assessment will focus on
	Easy to connect to the summative assessment in the end
	Other teachers of the same students can work on the same criteria
Advantages related to the learning of the students	The feedback can be individualised
	Comments from the teacher more reliable than from peers
	Students have the opportunity to improve their work
	Students take responsibility for their own learning and improve in their reflection abilities
	Students learn to be creative and to solve problems
Advantages related to relational and motivational aspects	Student-teacher relation is enhanced
	Students are motivated through feedback
Advantages for the teacher	Teacher gets a structured insight into what the students know

Note. Teachers were asked to reflect on their experiences regarding the implementation of formative assessment and especially feedback. From the answers, it became obvious that the teachers conflated the two terms to some extent. When they reflected on the advantages for the teacher, they mainly referred to formative assessment

where the variation is higher. This means that the teacher did not provide guidance at the same level of concreteness for all students. Looking at the student usage of feedback, all students addressed the feedback in their revisions of the original artefacts in cases S1 and S4. In case S1, this might be because the teacher provided subsequent feedbacks over the course of a month. So if a student did not work on the feedback, the teacher would comment on the same aspect again the next day. There is some variation in case S3: most students addressed the feedback to some extent but some students did not at all.

Benefits and Challenges Seen by Teachers in Using Written Feedback

Table 7.8 displays the benefits of written feedback as perceived by the Swiss teachers. The benefits can be grouped into four subcategories. Firstly, the teachers saw several advantages in the utilisation of a clear statement of assessment criteria, such as enhancing the transparency in the assessment. One of the teachers said, ‘The students and the parents know the assessment criteria from the beginning and know what the assessment will focus on’.

The second subcategory summarises advantages related to student learning: the teachers believed that written feedback is useful because it can be provided on an individual level and because it is reliable. They believed that the students do not only improve in the assessed competences but also in their self-regulation.

Thirdly, the teachers believed that the written feedback enhances the students-teacher relation but also the motivation of the students. One of the teachers reported,

Table 7.9 Challenges of written feedback as perceived by the Swiss teachers

Challenges related to teacher assessment literacy	How to come up with assessment criteria that are usable without later adaption?
	How to anticipate how the students will be working?
Challenges associated with students using the feedback	What to do if not all students are equally critical towards themselves?
	How to make sure the students take the feedback seriously?
	How to make sure the students will keep the feedback in mind over a longer time?
Practical challenges	How to minimise the teacher's workload?
	What to do if artefacts or reports that should be assessed are at home instead?

'The students were motivated through my feedback'. Finally, the teachers mentioned a benefit for themselves. By providing the written feedback, they received a structured insight into each student level of attainment.

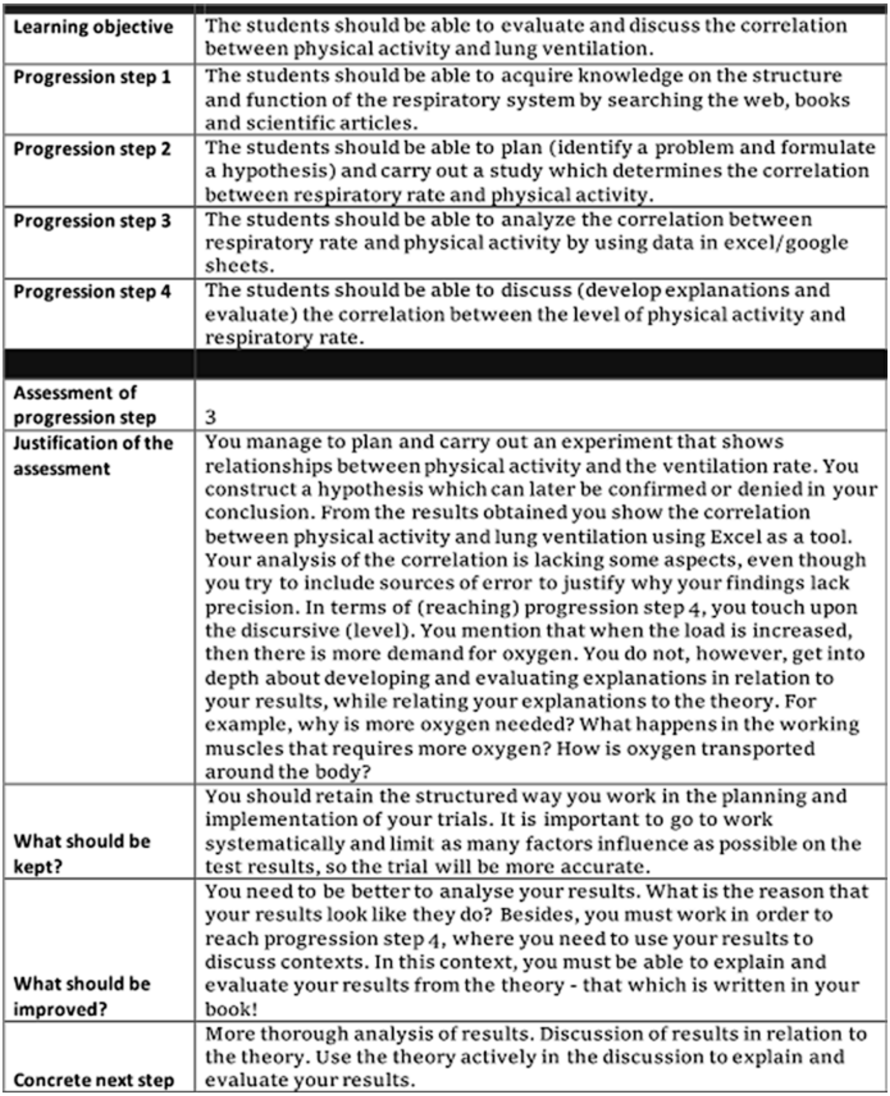
Table 7.9 displays the challenges of written feedback as perceived by the Swiss teachers, which could be grouped in three subcategories. Firstly, the teachers saw the challenges associated with the teacher providing the feedback in a standardised form. That means that assessment criteria and expected level of performance have to be anticipated beforehand and cannot be changed afterwards. One of the teachers said, 'When I notice during the unit that the assessment criteria are not so good, it is difficult to change them. If the unit is long, I can still adjust the criteria, but otherwise I have to apply the criteria as I communicated them in the beginning'.

Secondly, it was perceived as difficult to ensure the students would use the feedback meaningfully. Some students may not be able or willing to revise their work or to keep the advice provided by the teacher in mind. One teacher described, 'It is possible that some students forget the feedback when they write the next report. So it is important that the students can use the feedback soon again'. Thirdly, practical challenges were highlighted, particularly with respect to the time needed to provide the feedback, as well as the classroom management needed.

Denmark

Description of the Implementations

In Denmark, seven teachers implemented written feedback in their regular inquiry-based science units. In most implementations, the students worked in groups (two or more students), whereas in D1 some students worked alone and some students in pairs. The teachers used two different types of assessment templates. The first assessment template, which was elaborated over three or four progression steps, consisted of a description of the learning objectives (see Figure 7.2).



Learning objective	The students should be able to evaluate and discuss the correlation between physical activity and lung ventilation.
Progression step 1	The students should be able to acquire knowledge on the structure and function of the respiratory system by searching the web, books and scientific articles.
Progression step 2	The students should be able to plan (identify a problem and formulate a hypothesis) and carry out a study which determines the correlation between respiratory rate and physical activity.
Progression step 3	The students should be able to analyze the correlation between respiratory rate and physical activity by using data in excel/google sheets.
Progression step 4	The students should be able to discuss (develop explanations and evaluate) the correlation between the level of physical activity and respiratory rate.
Assessment of progression step	3
Justification of the assessment	You manage to plan and carry out an experiment that shows relationships between physical activity and the ventilation rate. You construct a hypothesis which can later be confirmed or denied in your conclusion. From the results obtained you show the correlation between physical activity and lung ventilation using Excel as a tool. Your analysis of the correlation is lacking some aspects, even though you try to include sources of error to justify why your findings lack precision. In terms of (reaching) progression step 4, you touch upon the discursive (level). You mention that when the load is increased, then there is more demand for oxygen. You do not, however, get into depth about developing and evaluating explanations in relation to your results, while relating your explanations to the theory. For example, why is more oxygen needed? What happens in the working muscles that requires more oxygen? How is oxygen transported around the body?
What should be kept?	You should retain the structured way you work in the planning and implementation of your trials. It is important to go to work systematically and limit as many factors influence as possible on the test results, so the trial will be more accurate.
What should be improved?	You need to be better to analyse your results. What is the reason that your results look like they do? Besides, you must work in order to reach progression step 4, where you need to use your results to discuss contexts. In this context, you must be able to explain and evaluate your results from the theory - that which is written in your book!
Concrete next step	More thorough analysis of results. Discussion of results in relation to the theory. Use the theory actively in the discussion to explain and evaluate your results.

Fig. 7.2 Translated example of a teacher's feedback journal (original language Danish). The feedback refers to the planning, carrying out and discussion of an investigation into the correlation between the level of physical activity and the respiratory rate

The written feedback was structured by five open text fields, namely, 'Assessment of progression step' (i.e. current progression level of the group), 'Justification of assessment' (i.e. why does the teacher class the group with this level?), 'What should be kept?', 'What should be improved?' and 'Concrete next steps'.

The second assessment template ('rubrics and open comments' template) consisted of a broken-down list of specific learning objectives where the teacher for each of these learning objectives had to mark on a three-point scale the degree to which the group has achieved this learning objective. For each learning objective, the teacher had the option of explaining his or her assessment. Finally, the template had an open field in which the teacher could point to what the group should consider in their next assignment.

The first teacher (D1, D7 and D8) teaches biology and physics/chemistry in lower secondary school and implemented this feedback method in a total of five classes (three classes for case D1, one class each for cases D7 and D8). The context for implementation D1 was a unit on indoor climate. The students were placed in groups and were asked to select a parameter relevant to indoor climate (e.g. temperature, decibel levels, CO_2 concentration). The students were asked to formulate a sketch as to how to investigate this parameter under changing conditions. The written feedback was then based on their sketch. After receiving the feedback, the students were asked to refine their sketch and carry out the investigation. For the written feedback, the teacher used the ‘open comments’ template. The context for D7 and D8 was a unit on human nutrition. The students were asked to investigate the nutritional value of various foods. For the written feedback, the teacher used the ‘rubrics and open comments’ template in case D7 and the ‘open comments’ template in case D8.

The second teacher (D4) teaches biology and physics/chemistry in lower secondary school and implemented this feedback method in one class. The context for her implementation was the same as for D1 (described above).

The third and fourth teacher (D2 and D3) teach in upper secondary school and implemented a unit about building electrical circuits in the discipline technology in one class each. The students worked with electrical circuits on breadboards. They were first asked to play around with the materials in order to make circuits with LEDs and to control their brightness. In the second part of the lesson, the students were asked to make a circuit with a bright-shining yellow LED and a weak-shining red LED. This was done as a competition where the circuit which best met the requirements won. The students were then asked to hand in a lab journal describing what they had done in the lesson. Both teachers used a variant of the ‘open comments’ template.

The fifth (D5 and D10), sixth (D6) and seventh (D9) teacher all teach biology at the upper secondary level. Their implementations concerned units about respiration (D5, one class), blood sugar regulation (D6, one class; D10, one class) and physical fitness (D9). In the implementations D5 and D6 the teachers used the ‘open comments’ template, while the teachers in the other cases used the ‘rubrics and open comments’ template. In all these implementations, the students were working with the given biology topic through establishing hypotheses, planning how to test these, carrying out an experiment and discussing the results.

In implementation D2 and D3, the main competence in focus was ‘design’ and ‘modelling’, while in all other implementations the competence in focus was ‘empirical investigation’.

Table 7.10 Overview on the quantitative results from Denmark in terms of diagnosis

	Assessment of level of attainment					Justification offered by the teacher about student's level of attainment									
	Addressing assessment criteria [in %][a]					Justification on what has been achieved					Justification on what has not yet been achieved				
	1	2	3	4	5	1	2	3	4	5	1	2	3	4	5
D1		7	28	62	3	No data					No data				
D2		50	50			No data					No data				
D3	14	43	43			No data					No data				
D4					100	No data					No data				
D5				20	80	No data					No data				
D6					100	No data					No data				
D7				77	23	No data					No data				
D8				53	47	No data					No data				
D9					100	No data					No data				
D10					100	No data					No data				

[a] 1=none; 2=some; 3=half; 4=most; 5=all. Percentage calculated per class, see last column of Table 7.1 for number of students per class

Quality of Written Feedback Provided by Teachers

Table 7.10 displays results on the Danish teachers' ability to use their respective feedback tools for diagnosing the levels of attainment of their students in the relevant competences. The results show that in all ten implementations, the teachers were able to address the assessment criteria. In eight of the ten implementations, the teachers addressed half or more of the criteria in at least 93% of all cases. The primary outliers seem to be implementations D2 and D3. In D2, the teacher addressed only some of the criteria in 50% of the cases. In D3, the teacher addressed only some of the criteria in 43% of the cases and did not address the criteria in 14% of the cases. In implementations D2 and D3, it seems that this lack of addressing the criteria was to some extent justified by the fact that the level of many of the students' first product did not mandate constructive feedback. For example, in a number of cases, the students did not hand in the assignment on time or even in a state that it was possible to provide feedback on.

During the coding of the data, two coders attempted to code independently for justification offered by the teacher about each student's level of attainment. Even after multiple iterations of coding, comparing and discussing, the coders were not able to reliably code this item (only a fair reliability with kappa=0.4 could be reached). The primary obstacle in this context was that the coders simply could not agree on *when* a particular feedback should be coded as being justified. For example, if a teacher writes to a student group 'you must provide a graph representation of your data', a case could be made that this assertion is justified in itself – because the students, all else being equal, should know that they did not represent their data, but a case could also be made that the teacher ought to explicitly justify her asser-

Table 7.11 Overview on the quantitative results from Demark in terms of provision of guidance about next steps in learning

	Provision of guidance about concrete next steps [in %][a]				Addressing feedback in second artefact [in %][b]			
	1	2	3	4	1	2	3	4
D1	10	59	31		No data			
D2			67	33	No data			
D3		72	14	14	No data			
D4			17	83	50	33	17	
D5	20	80			20	20	40	20
D6		33	67		No data			
D7	6	88	6		No data			
D8		95	5		No data			
D9		43	57			28.5	28.5	43
D10		9	82	9	No data			

[a] 1 = non-specific guidance; 4 = very specific guidance. Percentage calculated per class or student group, see last column of Table 7.1 for number of students per class

[b] 1 = none; 4 = all. Percentage calculated per student or student group, see last column of Table 7.1 for number of students per class

tion – e.g. with an assertion to the effect of 'you did not properly represent your data in a graph'. The question was often a matter of the existence (or not) of premises that in normal language use are tacit or implicit**.** Thus, in many of the discussions between the coders, their disagreement pertained to what they as coders could assume that the student group knows and what is shared (or even taken-for-granted information between the teacher and the student). These challenges do call on reflection of researchers when coding assessment data of this sort. The perspective of the coder is necessarily very different than the perspective of the student who receives the feedback, and it is naturally difficult to find a valid balance of what the coder can assume is shared information between teacher and student.

Table 7.11 displays the second part of the quantitative data from Denmark: the ability of teachers to provide advice for students about how to proceed in their learning.

The data shows that the teachers in all implementations were able to provide some guidance about next steps in learning. Only in three implementations (D1, D5 and D7) were there examples of non-specific guidance – ranging from 6% to 20% of the cases. On the other hand, only in four implementations (D2, D3, D4 and D10) were there examples of the teachers providing very specific guidance. Thus, in nine of the ten implementations, the majority of cases were coded as giving guidance to a medium specificity.

We chose to look at three implementations that differed in terms of distribution of the specificity of the guidance provided by the teacher. Implementation D4 is an example of an implementation with a generally high level of specificity in the guidance, D5 is an example of an implementation with a generally low level of specificity in the guidance, and D9 is an example of an implementation with medium range

Table 7.12 Benefits of written feedback as perceived by the Danish teachers

Advantages related to clearly stated assessment criteria	The template increased teacher awareness about giving feedback (prompted, focused and made more consistent the feedback in a way that was new)
Advantages related to the learning of the students	The fact that the templates make explicit concrete learning progressions allowed their students to better understand what the learning goals were and to better interpret where they were in the process of their learning
	The students were forced to reflect and react on concrete aspects in the assignment. The assessment becomes more a process, rather than a product, which is certainly beneficial for the student's learning
Advantages for the teacher	Using the template seemed to increase the reflectiveness of the teachers' assessment practice

specificity. Looking at the pattern of the student uptake of the feedback, there is no clear connection between the level of specificity of guidance and uptake; albeit this is a very marginal data amount to draw any conclusions from.

Benefits and Challenges Teachers See in Using Written Feedback

Table 7.12 displays the qualitative data that focuses on the benefits of written feedback as perceived by the teachers. We found three emerging categories that correspond to three of the four categories identified in the Swiss implementation: First, the Danish teachers saw a number of advantages in the clear statement of assessment criteria which enhanced transparency in the assessment. As one teacher stressed, 'It becomes much more concrete for the students; they know exactly what to keep working on the next time they do an assignment. The ensuing feedback becomes much more precise!'

Second, the Danish teachers identified benefits of using written feedback based on a template in terms of the student learning (in particular that the method induces students to use and reflect on the feedback to a higher extent than the teachers are used to). For example, one teacher noted, 'The students are to a much higher degree being involved in their own learning. They can participate in setting up their own learning objectives and "signs of learning", so that they are becoming involved in the assessment process'.

Third, the Danish teachers identified benefits relating to their role as a teacher, in particular the Danish teachers generally experienced that their level of reflection with regard to their assessment practice increased through their use of this method. As one teacher explained, 'The template works quite well; it is good as focusing the direction [of the feedback] when you are assessing […] compared to the way I used to work'. Interestingly, unlike the Swiss teachers, none of the Danish teachers made assertions about the perceived role of using written feedback to motivate students or to potentially foster and improve relationships between students and teachers.

Table 7.13 Challenges of written feedback as perceived by the Danish teachers

Challenges related to teacher assessment literacy	Formulating learning progressions is evidently a very complex and time-consuming task for teachers
	It was often difficult to select a few focal aspects to give feedback on
Challenges associated with students using the feedback	According to some teachers, some students found it difficult to decipher the many comments in the written feedback
	In particular, lower secondary school teachers stated that their students are used to oral feedback and had difficulties getting used to receiving written feedback
Practical challenges	Many teachers mention that the uses of both templates were time-consuming. Clearly the more open-ended template required more time than the template that consisted of a rubric

Table 7.13 displays the challenges of written feedback as perceived by the Danish teachers. Similar to the analysis of the Swiss implementation, we identified three emerging subcategories. First, teachers found it difficult to use a standardised template to formulate learning progressions. As one teacher stated, ‘The method is time consuming since the disciplinary learning objectives have to be translated into learning progression steps’.

Second, teachers found it difficult to get students to use the feedback. For example, one teacher stated that ‘The students had difficulties deciphering the many comments in the written feedback’. Third, the teachers identified practical issues concerning the use of this method. In particular, the Danish teachers generally stated that the use of this method is time-consuming – both in terms of designing the learning progressions and in terms of providing the feedback. As one teacher made explicit, ‘this assessment method takes a lot of time – both in terms of preparation and the implementation in practice’.

Discussion

In this chapter, the quality of written feedback that is provided based on rubrics and templates for open comments was explored. $N = 14$ teachers from three different European countries implemented written feedback in inquiry-based science units as one approach of giving feedback within formative assessment. Regarding research question 1, the feedback of the teachers was analysed in terms of different criteria which will be used to structure the following discussion. The results from the three countries will be summarised with special emphasis on similarities and differences.

To What Extent Did the Feedback Comments Take Into Account the Learning Goals Which Had Been Provided by the Teacher Beforehand?

The results from all countries show that the teachers are generally able to refer to previously specified criteria in their feedback. An important feature of good feedback (Arts et al. 2016; Nicol and Macfarlane-Dick 2006) seems to be fulfilled hereby. The results imply that if teachers set criteria for assessment, they also integrate these in their feedback which leads to a feedback that is not detached from the learning goals. This is particularly important in the context of formative assessment which aims at communicating to the students where they are in their learning compared to the learning goals and how to reach these goals. This is only possible when the learning goals are clear and referred to in the feedback.

To What Extent Did the Teacher Justify His or Her References to What the Student Had Already/Not Yet Achieved with Regard to the Targeted Competence?
In terms of justification, the results in the three countries differ. In the case of Germany, the teachers tend to justify their references to what the student has not yet achieved. It is possible that the German teachers considered these justifications more relevant than those that refer to what has already been achieved. Compared to the Germans, the Swiss teachers provided fewer justifications in their feedback. The reasons for this remain unclear. Since the importance of justified feedback is known from literature (Arts et al. 2016; Glover and Brown 2006; Orsmond and Merry 2011), there seems to be room for improvement. In Denmark, it was not possible to analyse to what extent the teachers in the study justified their references in their feedback. Even after multiple iterations of coding, comparing and discussing, the coders were not able to agree on when an assertion could be considered as justified.

To What Extent Does the Feedback Provide Specific Guidance to the Student About Concrete Next Steps?
In all of the three countries, the teachers provided specific guidance on next steps, but differences within the countries can be recognised. In Germany, two out of the five teachers provided rather abstract guidance on next steps in learning. In Switzerland, this was the case for one out of four teachers and in Denmark for four out of ten cases. The reason for this could be related to the unfamiliarity of the teachers with the materials and with the approach of written feedback in general. In the German cases, for example, it appeared that the teachers were more apt to work with the materials in the third round of implementation than in earlier rounds. For Switzerland and Denmark, this relation cannot be shown. However, the results from Germany imply that teachers need support in that respect. Another possible explanation in the case of Switzerland and Denmark could be that the artefacts of some students are of a high quality so that next steps were not necessary. As this was neither coded in Switzerland nor in Denmark, it should be clarified in further analysis.

This leads to the question whether teachers adapt their feedback in terms of the level of concreteness dependent on the student level. In Germany and in Switzerland, this was not the case. In both countries, the individual teachers provided guidance with the same level of concreteness across all their students. In Denmark, a similar picture emerges although no fine-grained analysis could be made since the teachers (if at all) only selected the level of competence attainment on a three or four level taxonomy. Considering the prerequisite that formative assessment should be provided in an individualised way (Arts et al. 2016; Nicol and Macfarlane-Dick 2006), it seems probable that not all students were able to benefit optimally from the feedback.

To What Extent Did the Student Address the Feedback on the First Artefact in His or Her Second Artefact?

Considering the use of feedback in the German cases, a variation across classes is obvious so that it is difficult to find patterns. It seems, however, that a rather abstract way of formulating next steps in learning leads to the students not using the feedback. The results from Switzerland, by contrast, imply that student use of feedback may not depend only on the concreteness of the feedback but also on the time provided to students to engage with the feedback and the continuity of working on an artefact. In Denmark, the use of the feedback was only analysed in selected cases. The results of these cases show no clear relation between specificity of guidance and uptake of the feedback by the students. The results from Denmark therefore contradict the conjecture derived from the German data.

The differences between the countries could also be related to the different ways of how the students used the feedback. The students in Germany transferred the feedback to a new context, whereas the students in Switzerland used the feedback to improve their original artefact. This means that the Swiss students had the opportunity to continuously integrate the feedback in their work. These country-specific results allow for the assumption that direct integration of feedback is easier for students compared to transferring it to new contexts. There seems to be a relation between the context and complexity of the task to which the feedback has to be applied to and the degree to which the feedback is used. If – as in the case of Germany – the feedback has to be transferred to new, more complex topics, it needs to be very concrete. Otherwise, the data from Germany suggests that the feedback will be not used or left ignored. These results are consistent with Lipnevich et al. (2014) who also found that there might be differences in the degree students make use of the feedback they receive depending on whether there is the opportunity to revise the original work or whether the feedback is to be transferred to a subsequent task.

Benefits and Challenges in Using Written Feedback

In research question 2, benefits and challenges which teachers see in using written feedback were investigated. In order to answer this question, the teachers were asked to fill out a self-reporting tool in which they reflected on the usability of written feedback in their daily teaching practice. Considering the advantages of written feedback, the teachers from all three countries mentioned aspects that relate to student learning and aspects that relate to the teachers themselves. However, when the teachers reflected on the advantages for themselves, they sometimes referred to formative assessment only and not specifically to written feedback. Nevertheless, the teachers stated that providing written feedback helps to get an insight into a student's level of achievement, which can be considered a prerequisite to provide feedback by informing the students on where they are (Hattie and Timperley 2007). This insight also provides teachers with hints on how to plan their lessons. The quotes from the teachers are well aligned with earlier findings (Andrade 2005). From the perspective of the teachers, written feedback is also useful for the students. On the one hand, they receive an individualised feedback which helps them to regulate their own learning. Again, this can be linked to the feedback theory from Hattie and Timperley (2007), where individualised feedback can tell a student where they are and what they should do next in order to reach the learning goals. On the other hand, the feedback can help to enhance student learning since they can further improve their artefacts. In this respect, written feedback appears as part of a process rather than a product. The results of the present study suggest a potential mechanism for how formative assessment supports student learning, specifically that formative assessment enhances student learning because of the individualised feedback that enables students to revise their pieces of work. The teachers from both Germany and Switzerland also mention motivational aspects. They say that the relationship between students and teachers improves through the use of feedback and that the feedback motivates the students to proceed. Besides these advantages for students and teachers, the teachers from Denmark and Switzerland also mention advantages that are related to the clearly formulated assessment criteria. They appreciate the resulting transparency and certainty for both students and teachers. The criteria allow also for easier combination of the written feedback with summative assessment and for other teachers to work on the same criteria with the same group of students. Furthermore, the Danish teachers mention that using the template helped them to reflect on their own assessment practice.

The advantages that are reported by the teachers stress the worth of written feedback provided by the teacher. However, the teachers also mention challenges that are related to the use of written feedback. These challenges are similar to the findings of Bharuthram (2015) and Luft (1999), specifically that teachers find practical difficulties in all three countries. The time needed to provide written feedback is considered a main difficulty. On the one hand, effort is needed to formulate the written feedback itself. On the other hand, the planning of units with integrated opportunities for formative assessment is also time-consuming. Furthermore, the teachers from all three countries mentioned challenges related to their own assessment literacy. The German teachers refer to the difficulties of providing reliable and effective feedback without giving away the correct solution, which is also mentioned in the studies of Tuck (2012) as well as Bruno and Santos (2010). The Danish teachers express that it is difficult to select specific aspects which should be addressed in the feedback they give. The Swiss teachers mention the challenges of using a standardised tool for written feedback in a flexible way so that the feedback is effective for all students. This is also one of the major difficulties brought up as a challenge related to student use of feedback, namely, that teachers do not know to what extent the feedback is taken seriously and if the important issues in the feedback have been recognised. These challenges are consistent with the findings in the study from Bailey and Garner (2010). Additionally, some teachers express their uncertainty on what measures to take so that students would not forget the feedback right away. These challenges support the quantitative results from Germany which showed that the students rarely used the feedback even though it was concrete.

Overall, clear benefits of written feedback can be demonstrated. However, the reported insights also show that written feedback has a number of challenges associated. The disadvantages also explain why structured formative assessment has not yet become part of normal teaching practice. Though, even in light of the challenges, it is possible to develop a number of means that could support the uptake of formative assessment in teaching practice.

Limitations of the Study
This chapter is grounded on an exploratory study. Moreover, the results are based on a small sample size and therefore cannot be generalised. Due to the small sample size, it is not possible to elaborate on differences between school levels, subjects or different ways of providing written feedback. In these respects, there is a need for further research. In subsequent intervention studies, different school levels and subjects should be controlled or systematically varied, respectively. This would imply that more teachers are needed. However, even for this exploratory study, it was difficult to find teachers who were willing to collaborate with research over a longer period of time (despite the fact that they received financial compensation). This implies that formative assessment (the focus of the whole ASSIST-ME project) was not a sufficient 'drawcard' to entice potential participants. The topic and its relevance should therefore be given more weight in teacher education and professional development.

Furthermore, the study focuses on the perspective of the teachers. Student opinions, the effects on student learning and the effects on the self-regulation of the students remained unclear. It would also be interesting to investigate how different groups of students benefitted from the written feedback. For example, it is plausible that students with different levels of language competence may have benefitted from written feedback to different degrees. Additionally, knowledge on each teacher's prior experiences in providing written feedback would be necessary to measure the effect of the tool itself.

Conclusions and Implications

In general, the results from this multi-country study suggest that the teachers involved were able to provide feedback structured by a feedback tool that related to previously defined criteria. The teachers were furthermore animated to justify their feedback and to provide guidance on next steps in learning. Therefore, relevant criteria for high quality of the feedback seem to be fulfilled. In the case of Switzerland, it was demonstrated that students used their teacher's feedback for revisions. But given the large number of students in Germany and Denmark that used little to none of their received feedback, there is still large room for improvement in this respect. Furthermore, the results suggest that it may be beneficial to provide teachers with tools, ideas and examples of good practice so that they are supported in providing students with written feedback. The results thus overall imply that written feedback can be used meaningfully for formative assessment in science learning.

Nevertheless, the results indicate that the feedback was in some cases not justified or not formulated in a concrete way. This might mean that materials and tools are not sufficient to enable teachers to provide effective feedback. In pre- and in-service training, teachers should also be introduced to the use of such tools for effective and concrete feedback including justifications on what students have or have not yet achieved. Effective and concrete feedback can, as the results imply, enhance the quality of student artefacts and the abilities of the students to transfer this feedback to new contexts. However, the transfer seems quite difficult for students even with very concrete feedback.

Despite these results, the teachers mentioned a variety of benefits. These included improved clarification of assessment criteria, the potential for enhanced student learning, motivational aspects and the potential for the teachers to adapt their own teaching.

The teachers also recognised a number of challenges related to written feedback which may inhibit the acceptance of the approach. It is suggested that these challenges should be addressed not only in pre- and in-service training but also at an educational policy level. As part of teacher training, teachers should be shown how to integrate written feedback into their teaching routine. Since written feedback is by its own nature time intensive, an important goal in this training is to enable teachers to decide at which points during their instruction the use of written feedback is most beneficial. In addition, teachers need support on how to motivate students to engage with the feedback, how to improve their work based on it and how to memorise the main aspects of the feedback. With respect to educational policy, efforts should be made to better acknowledge the importance of feedback and formative assessment, which should in turn lead to an increasing acceptance of these approaches in teaching practice. One such approach would be to give more time and scope to teachers so that written feedback can be seen as a natural part of a teacher's instruction, rather than as something 'additional' to manage within limited time available for teaching and learning.

References

Abd El Khalick, F., Boujaoude, S., Duschl, R. A., Lederman, N. G., Mamlok-Naaman, R., Hofstein, A., Niaz, M., Treagust, D., & Tuan, H. (2004). Inquiry in science education: International perspectives. *Science Education, 88*(3), 397–419.

Andrade, H. (2000). Using rubrics to promote thinking and learning. *Educational Leadership, 57*(5), 13–18.

Andrade, H. (2005). Teaching with rubrics: The good, the bad, and the ugly. *College Teaching, 53*(1), 27–31.

Andrade, H., & Du, Y. (2005). Student perspectives on rubric-referenced assessment. *Practical Assessment, Research & Evaluation, 10*(3), 1–11.

Arter, J. A., & McTighe, J. (2001). *Scoring rubrics in the classroom: Using performance criteria for assessing and improving student performance*. Thousand Oaks: Corwin Press.

Arts, J. G., Jaspers, M., & Joosten-ten Brinke, D. (2016). A case study on written comments as a form of feedback in teacher education: So much to gain. *European Journal of Teacher Education, 39*(2), 159–173.

Bailey, R., & Garner, M. (2010). Is the feedback in higher education assessment worth the paper it is written on? Teachers' reflections on their practices. *Teaching in Higher Education, 15*(2), 187–198.

Ball, D. L., & Forzani, F. M. (2011). Teaching skillful teaching. *The Effective Educator, 68*(4), 40–45.

Bennett, R. E. (2011). Formative assessment: A critical review. *Assessment in Education: Principles, Policy & Practice, 18*(1), 5–25.

Bharuthram, S. (2015). Lecturers' perceptions: The value of assessment rubrics for informing teaching practice and curriculum review and development. *Africa Education Review, 12*(3), 415–428.

Black, P., & Harrison, C. (2004). *Science inside the black box*. London: GL Assessment.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy and Practice, 5*(1), 7–73.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. London: Open University Press.

Brookhart, S. M. (2008). *How to give effective feedback to your students*. Alexandria: Association for Supervision and Curriculum Development.

Brown, G. T. L., Harris, L. R., & Harnett, J. A. (2012). Teacher beliefs about feedback within an Assessment for learning environment: Endorsement of improved learning over student well-being. *Teaching and Teacher Education, 28*(7), 968–978.

Bruno, I., & Santos, L. (2010). Written comments as a form of feedback. *Studies in Educational Evaluation, 36*, 111–120.

Burke, K. (2006). *From standards to rubrics in 6 steps*. Heatherton: Hawker Brownlow Education.

Coffey, J. E., Hammer, D., Levin, D. M., & Grant, T. (2011). The missing disciplinary substance of formative assessment. *Journal of Research in Science Teaching, 48*(10), 1109–1136.

Darling-Hammond, L., Ancess, J., & Falk, B. (1995). *Authentic assessment in action: Studies of schools and students at work*. New York: Teachers College Press.

D-EDK Deutschschweizer Erziehungsdirektoren-Konferenz. (2014). *Lehrplan 21*. http://vorlage.lehrplan.ch/downloads.php [15.12.2015].

Furtak, E. M. (2012). Linking a learning progression for natural selection to teachers' enactment of formative assessment. *Journal of Research in Science Teaching, 49*(9), 1181–1210.

Furtak, E. M., Ruiz-Primo, M. A., Shemwell, J. T., Ayala, C. C., Brandon, P. R., Shavelson, R. J., & Yin, Y. (2008). On the fidelity of implementing embedded formative assessments and its relation to student learning. *Applied Measurement in Education, 21*(4), 360–389.

Furtak, E. M., Kiemer, K., Circi, R. K., Swanson, R., de León, V., Morrison, D., & Heredia, S. C. (2016). Teachers' formative assessment abilities and their relationship to student learning: Findings from a four-year intervention study. *Instructional Science, 44*, 267–291.

Gearhart, M., Nagashima, S., Pfotenhauer, J., Clark, S., Schwab, C., Vendlinski, T., Osmundson, E., Herman, J., & Bernbaum, D. J. (2006). Developing expertise with classroom assessment in K–12 science: Learning to interpret student work. Interim findings from a 2-year study. *Educational Assessment, 11*(3–4), 237–263.

Gibbs, G., & Simpson, C. (2004). Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education, 1*, 3–31.

Glover, C., & Brown, E.. (2006). "Written Feedback for Students: Too Much, Too Detailed or Too Incomprehensible to Be Effective?" *BEE-J* 7. Retrieved from https://www.heacademy.ac.uk/sites/default/files/beej.7.1d.pdf.

Hammer, D., & van Zee, E. (2006). *Seeing the science in children's thinking: Case studies of student inquiry in physical science*. Portsmouth: Heinemann.

Harks, B., Rakoczy, K., Hattie, J., Besser, M., & Klieme, E. (2014). The effects of feedback on achievement, interest and self-evaluation: The role of feedback' s perceived usefulness. *Educational Psychology, 34*(3), 269–290. doi:10.1080/01443410.2013.785384.

Hattie, J. (2009). *Visible learning. A synthesis of over 800 meta-analyses relating to achievement*. London/New York: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112.

Jonsson, A. (2014). Rubrics as a way of providing transparency in assessment. *Assessment and Evaluation in higher education, 39*(7), 840–852.

Kazemi, E., & Franke, M. L. (2004). Teacher learning in mathematics: Using student work to promote collective inquiry. *Journal of Mathematics Teacher Education, 7*, 203–235.

Kingston, N., & Nash, B. (2011). Formative Assessment: A meta-analysis and a call for research. *Educational Measurement: Issues and Practice, 30*(4), 28–37.

Kluger, A., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254–284.

Kobarg, M. & Seidel, T. (2007). Prozessorientierte Lernbegleitung – Videoanalysen im Physikunterricht der Sekundarstufe I [Process-oriented teaching – Video analyses in high school physics instruction]. Unterrichtswissenschaft, 35(2), 148–168. Retrieved from http://www.pedocs.de/volltexte/2012/5490/

Köller, O. (2005). Formative Assessment in classrooms: A review of the empirical German literature. In OECD (Ed.), *Formative Assessment. Improving learning in secondary classrooms* (pp. 266–280). Paris: OECD.

Levin, D., Hammer, D., & Coffey, J. (2009). Novice teachers' attention to student thinking. *Journal of Teacher Education, 60*, 142–154.

Linn, M. C., Davis, E. A., & Bell, P. (Eds.). (2004). *Internet environments for science education*. Mahwah: Lawrence Erlbaum Associates Publishers.

Lipnevich, A. A., McCallen, L. N., Pace Miles, K., & Smith, J. K. (2014). Mind the gap! Students' use of exemplars and detailed rubrics as formative assessment. *Instructional Science, 42*, 539–559.

Love, N., Stiles, K. E., Mundry, S., & DiRanna, K. (2008). *The data coach's guide to improving learning for all students: Unleashing the power of collaborative inquiry*. Thousand Oaks: Corwin Press.

Luft, J. A. (1999). Rubrics: Design and use in science teacher education. *Journal of Science Teacher Education, 10*(2), 107–121.

Marzano, R. J., & Arthur, S. (1977). Teacher comments on student essays: It doesn't matter what you say. ERIC Document Reproduction Service: ED 147 864.

Mayring, P. (2004). Qualitative content analysis. In U. Flick, E. von Kardorff, & I. Steinke (Eds.), *A companion to qualitative research* (pp. 2662–2669). London: Sage.

Ministerium für Schule und Berufsbildung. (2016). Fachanforderungen Chemie. Allgemeinbildende Schulen. [Chemistry curriculum for lower secondary schools]. Retrieved from http://lehrplan.lernnetz.de/intranet1/index.php?wahl=199.

Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction – What is it and does it matter? Results from a research synthesis years 1984 to 2002. *Journal of Research in Science Teaching, 47*(4), 474–496.

Moni, R. W., & Moni, K. B. (2008). Student perceptions and use of an assessment rubric for a group concept map in physiology. *Advances in Physiology Education, 32*(1), 47–54.

Moskal, B. M. (2003). Recommendations for developing classroom performance assessments and scoring rubrics. *Practical Assessment, Research & Evaluation, 8*(14), 1–10.

Ni, Y. (1997). Performance-based assessment: Problems and design strategies. *Education Journal, 25*(2), 137–157.

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative Assessment and self-regulated learning: A Model and seven principles of good feedback practice. *Studies in Higher Education, 31*, 199–218.

Nielsen, J. A., & Dolin, J. (2016). *Evaluering mellem mestring og præstation*. [Assessment between mastery and performance]. *MONA*, *1*, 51–62.

Nunes, C. (2004). *A avaliação como regulação do processo de ensino aprendizagem da Matemática* [Assessment and regulation of the teaching in mathematics education]. Lisbon: Lisbon University.

OECD (Organisation for Economic Co-operation and Development). (2005). *Formative assessment. Improving learning in secondary classrooms*. Paris: OECD Publishing.

OECD (Organisation for Economic Co-operation and Development). (2013). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment.* OECD Reviews of Evaluation and Assessment in Education. Paris: OECD Publishing.

Orsmond, P., & Merry, S. (2011). Feedback alignment: Effective and ineffective links between tutors' and students' understanding of coursework feedback. *Assessment and Evaluation in Higher Education, 36*, 125–136.

Panadero, E., & Jonsson, A. (2013). The use of scoring rubrics for formative assessment purposes revisited: A review. *Educational Research Review, 9*, 129–144.

Parr, J., & Timperley, H. S. (2010). Feedback to writing, assessment for teaching and learning and student progress. *Assessing Writing, 15*, 68–85.

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground – A literature review of empirical research on scientific inquiry activities. *Studies in Science Education, 52*(2), 161–197.

Roth, K. J., Garnier, H. E., Chen, C., Lemmens, M., Schwille, K., & Wickler, N. I. Z. (2011). Video based lesson analysis: Effective science PD for teacher and student learning. *Journal of Research in Science Teaching, 48*(2), 117–148.

Ruiz-Primo, M. A., & Li, M. (2013). Analyzing teachers' feedback practices in response to students' work in science classrooms. *Applied Measurement in Education, 26*(3), 163–175.

Ruiz-Primo, M. A., Li, M., Ayala, C., & Shavelson, R. J. (2004). Evaluating students' science notebooks as an assessment tool. *International Journal of Science Education, 26*(12), 1477–1506.

Russ, R. S., Coffey, J. E., Hammer, D., & Hutchison, P. (2009). Making classroom assessment more accountable to scientific reasoning: A case for attending to mechanistic thinking. *Science Education, 93*, 875–891.

Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science, 18*(2), 119–144.

Santos, L., & Dias, S. (2006). *Como entendem os alunos o que lhes dizem os professores? A complexidade do feedback* [How do the students understand what the teachers say? On the complexity of feedback]. ProfMat 2006. Lisboa: APM.

Schiepe-Tiska, A., Schmidtner, S., Müller, K., Heine, J.-H., Neumann, K., & Lüdtke, O. (2016). *Naturwissenschaftlicher Unterricht in Deutschland in PISA 2015 im internationalen Vergleich* [Science teaching in Germany in PISA 2015 – results from the international comparison]. In. K. Reiss, C. Sälzer, A. Schiepe-Tiska, E. Klieme, & O. Köller (Eds.), *PISA 2015 – eine Studie zwischen Kontinuität und Innovation* (pp. 133–176). Münster: Waxmann.

Searle, D., & Dillon, D. (1980). The message of marking: Teacher written responses to student writing at intermediate grade levels. *Research in the Teaching of English, 14*, 233–242.

Smit, R., & Birri, T. (2014). Assuring the quality of standards-oriented classroom assessment with rubrics for complex competencies. *Studies in Educational Evaluation, 43*(December), 5–13.

So, W. W. M., & Lee, T. T. H. (2011). Influence of teachers'perceptions of teaching and learning on the implementation of assessment for learning in inquiry study. *Assessment in Education: Principles, Policy and Practice, 18*(4), 417–432.

Stiggins, R. J. (2002). Assessment crisis: The absence of assessment FOR learning. *Phi Delta Kappan, 83*(10), 758–765.

Stracke, E., & Kumar, V. (2010). Feedback and self-regulated learning: Insights from supervisors' and Ph.D. examiners' reports. *Reflective Practice, 11*(1), 19–32.

Tuck, J. (2012). Feedback-giving as social practice: Teachers' perspectives on feedback as institutional requirement, work and dialogue. *Teaching in Higher Education, 17*(2), 209–221.

Vögeli-Mantovani, U. (1999). *SKBF Trendbericht Nr. 3: Mehr fördern, weniger auslesen. Zur Entwicklung der schulischen Beurteilung in der Schweiz* [SKBF trend report no. 3: More support, less selection. On the development of educational assessment in Switzerland]. Aarau: Schweizerische Koordinationsstelle für Bildungsforschung.

Wiggins, G. P. (1998). *Educative assessment: Designing assessments to inform and improve student performance*. San Francisco: Jossey-Bass.

Wollenschläger, M., Möller, J., & Harms, U. (2011). *Effekte kompetenzieller Rückmeldung beim wissenschaftlichen Denken* [Effects of competential feedback on performance in scientific reasoning]. *Zeitschrift für Pädagogische Psychologie, 25*(3), 197–202.

Part III
General Reflections

Chapter 8
European Educational Systems and Assessment Practice

Robert Evans, David Cross, Michel Grangeat, Laurent Lima, Nadia Nakhili, Elie Rached, Mathias Ropohl, and Silke Rönnebeck

Introduction

As described in the introduction to this book, this project explored the fit between inquiry-based education (IBE) and the use of formative assessment. One hypothesis was that formative assessment has the potential to facilitate both teacher and student work with IBE since it can provide some assurance and direction during the challenging environment of inquiry. As part of a review done to establish the context of formative assessment in inquiry-based education, this chapter explores possible links between educational systems and teachers' practices with respect to assessment. To examine these connections, the first of the two sections uses a survey of characteristics of eight European educational policies that may have an influence on assessment practices. It produces a typology of two country groups based on degree of centralization of administration as well as teacher and school autonomy. This first section also provides four models of education in the eight countries (Cyprus, Czech Republic, Denmark, France, Finland, Germany (state of Schleswig-Holstein), Switzerland, and the United Kingdom (England)), based on comparative curricula, teacher education, and professional development and assessment practices.

R. Evans (✉)
Department of Science Education, University of Copenhagen, Copenhagen, Denmark
e-mail: evans@ind.ku.dk

D. Cross • M. Grangeat • L. Lima • N. Nakhili • E. Rached
Department of Educational Science, University of Grenoble Alpes, Grenoble, France

M. Ropohl
Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

S. Rönnebeck
Leibniz-Institute for Science and Mathematics Education (IPN), Kiel, Germany

Kiel University, Kiel, Germany

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_8

The second part of the chapter then surveys the relevant formative and summative assessment research of the same eight European countries. It builds on the first section's survey of system characteristics by examining the resulting assessment practices in each country. Given the degrees of centralization and autonomy of the countries, their assessment practices are discussed. The diversity of their existing practices shows the relative importance of summative and formative assessment in each case, the current extent of the use of formative assessment, and the challenges faced in the uptake of formative assessment methods. It concludes with the role of assessment in supporting inquiry-based education.

The goal of both parts of the chapter is to provide generalized insights into the alignment of formative and summative assessment with educational practices. The overall Pan-European affordances and challenges of assessment practices in the given contexts provide an insight into the multiple variables that account for the assessment practices in these European countries.

Links Between European Educational Systems and Assessment Practices

The possible linkage between educational systems and teachers' practices with respect to assessment was addressed by Cohen and Hill (2000) who examined the influence of assessment, curriculum, and professional development on teacher practice and student achievement. They analyzed 1994 California surveys on teachers' practices (including their formative and summative assessment practices) and students' achievement and showed that policies can affect teacher practice and student performance. As educational policies interact with national educational systems, we used their work to build the model in Fig. 8.1 that presents the relationships between educational policies, teaching practices, and learning outcomes (Lima et al. 2015).

This model proposes that the relationship between educational policies and teacher practice is mediated by:

- The curriculum (students' curricula, teachers' own curricula, and student and teacher knowledge about students' curricula)
- In-service professional development programs (which could be extended to teacher education in a broader model)
- State-wide or standardized student assessments for they allow teachers to think about the alignment among their teaching approach, the contents of the tests, and students' performances

Testing this model, Cohen and Hill (2000, p. 327) concluded that "both our practice and policy measures positively relate to student achievement. State efforts to improve instruction can affect both teaching and learning." As shown by Cross and Lepareur (2015) and Grangeat and Hudson (2015), other factors that are independent from educational policy factors, like the nature of teacher collaboration or the

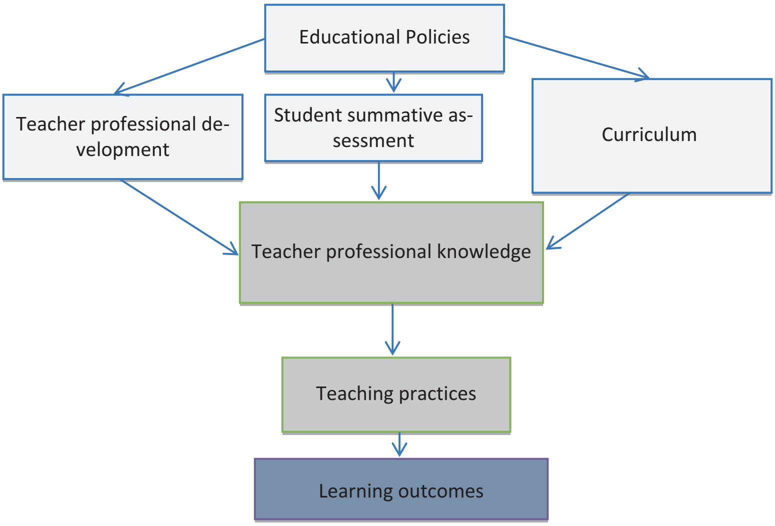

Educational Policies
Teacher professional development
Student summative assessment
Curriculum
Teacher professional knowledge
Teaching practices
Learning outcomes

Fig. 8.1 Model of the relationships between educational policies, teacher practices, and student performance (Lima et al. 2015) (The different shadings correspond to different levels of description: *in light gray*, the education system level; *in medium gray*, the teachers' level; and *in dark gray*, the students' level)

opportunities to talk about subject matter, teaching, or students learning (Wilson and Berne 1999), influence teachers' professional knowledge and practice. However, this contribution focuses on the impact of educational systems on teacher professional activity.

In the first step of the research, educational policies and system characteristics that might influence student performance were identified. The study of Hanushek and Woessmann (2014), by focusing on macro-level variables that play a role in international evaluation of student performances (PISA, TIMSS), elicits the main characteristics of educational systems that influence performance differences at an international level. The same characteristics are not necessarily relevant in explaining differences at a national level because each educational system has its own coherence and is included in a broader social system. Keeping in mind that this contribution focuses on the comparison among the national systems, it is drawn on these characteristics. Of the five main characteristics identified in the literature, three are used in this contribution for characterizing the educational systems participating in the sample (the remaining two are not relevant here):

- Accountability: The existence of an external exit exam or teacher use of regular standardized tests to monitor student performance is associated with higher student performance.
- Autonomy of schools: Students perform significantly better in schools that have autonomy in organization and personnel decisions (budget allocation, hiring teachers, choosing textbooks and instructional methods, etc.).

- Tracking, streaming, or ability grouping: In countries with early tracking (academic vs. vocational), inequality of performances linked to social background increases systematically from fourth grade to the age of 15, whereas it decreases in countries without tracking (Hanushek and Woessmann 2014, p. 167).

The two other characteristics of educational system that are of some influence on student performance concern preprimary education and the competition from private schools. They are out of range of our study that focuses on science teaching in primary and secondary schools.

The data of an online survey conducted through ASSIST-ME allows identification of some characteristics of European educational policies that might influence teachers' assessment practices as a part of their teaching practices: system organization and management, school organization and management, teacher education and professional development, science education, and forms of student assessment (see introduction to this book). This survey was conducted in eight European countries: Cyprus, Czech Republic, Denmark, France, Finland, Germany (state of Schleswig-Holstein), Switzerland, and the United Kingdom (England). A group of researchers and experts of education (three in each country chosen by the head of the national research teams involved in the research program) answered a total of 111 questions about the educational system of their national school system (including school autonomy, accountability, and tracking), teacher education and professional development, science education curriculum, and the form of student assessment at primary, lower secondary, upper secondary, and vocational secondary levels. All the questions were close ended except for three that were open ended in order to grasp the fine details of each educational system and to moderate the results from the close-ended questions. In each country, the experts had to reach a consensus before all their answers were submitted to a panel of national stakeholders. The group of stakeholders (representatives of educational institutions, head teachers, politicians, and administrators in charge of the local, regional, or national education system) was asked to react to and comment on all answers. That ensures that this information reflected both researchers and partners' views about their own educational system.

The exploration of the consortium educational systems has been conducted through multiple correspondence analysis (MCA) and cluster analysis. This type of analysis of the answers to the questionnaire allows elicitation of the main dimensions (Le Roux and Rouanet 2010) that characterize the educational systems. The analysis of the three characteristics of educational systems (accountability, autonomy, and tracking) produces a typology consisting of two groups (see Fig. 8.2). This typology is mainly based on the differences in autonomy and accountability as tracking did not seem to differ significantly among the countries involved in the analysis. First, this analysis identifies a group of centralized countries (France and Cyprus) in which the independence of schools and individual teachers is weak and the main decisions are made by the central administration. Second, this analysis identifies a group of more or less decentralized countries where the system is rather teacher centered (Finland and Denmark), rather school centered (Czech Republic)

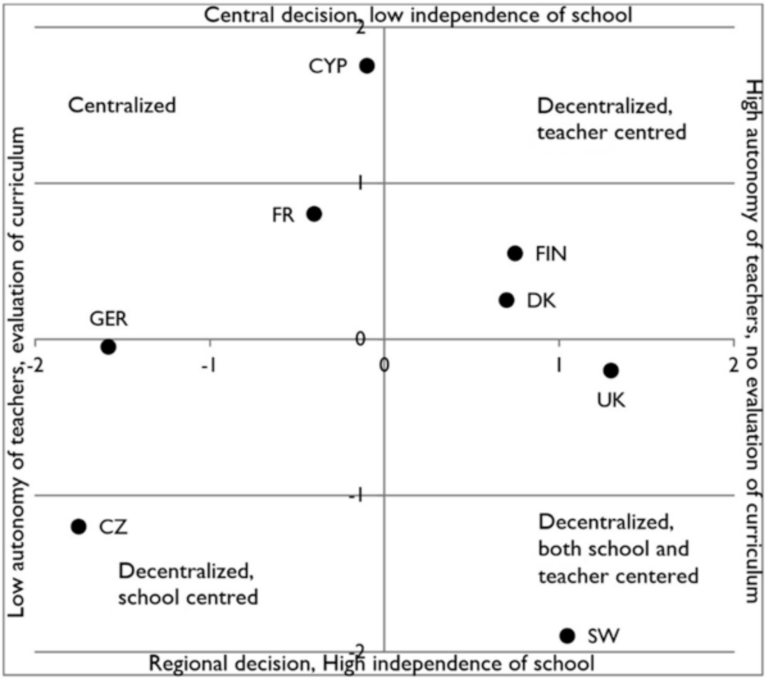


Fig. 8.2 Results of the MCA analysis on educational systems' characteristics

or represents a kind of balance between school and teacher autonomy (United Kingdom and Switzerland). For example, in centralized countries, central authorities are in charge of teacher hiring, while in decentralized countries, local authorities or schools are in charge of teacher hiring. In decentralized countries, schools and sometimes teachers have some autonomy in the implementation of the national curriculum, when it exists, but they don't have any autonomy in the centralized countries. In addition, one country is in an intermediate position between centralization and decentralization (Germany) (Lima et al. 2013, p. 6–11). Even though the position of Germany might be surprising, it is consistent with the data of the OECD (2013) that show that, in Germany, the percentage of decisions taken at the central as well as those taken at local level are at the average of OECD countries. Those taken at the regional level (länder) are a lot more common than the OECD average, and those taken at the school level are a lot less common than the OECD mean. Such contrasts between low level of school decisions and high level of regional decisions place Germany in a balanced position between fully decentralized and highly centralized countries.

Answers to the eight country questionnaire were analyzed through a quantitative method. The analysis was based on MCA with countries as subjects and questions as variables. Because the model of the relationships between educational policies and teachers' practices (including teachers' assessment practices) focuses on the curriculum, and on teacher education and professional development, the concentration of the analysis was on these three domains:

- Curriculum: Does the competence model of the students' curriculum, if it exists, specify competencies related to formative assessment? Do the curriculum's guidelines require student day-to-day assessment? Do resources for teachers exist in order to support the uptake of day-to-day assessment related to STEM?
- Teachers' education and professional development: Since when did formative assessment/summative assessment appear in initial teacher education? What is the amount of formative assessment/summative assessment in continued professional development (CPD) programs? Since when did formative assessment/summative assessment appear in CPD programs?
- Teachers' assessment practice in the classroom: How is student achievement communicated to the students? How common is it that students are involved in the assessment of their own (and others) performance? Are there dedicated meetings for helping students and parents to make sense of the assessment information and decide strategies for improving their learning?

The MCA results detected variables that distinguish the national educational systems and identified proximities between some systems on these variables. Table 8.1 illustrates the results of this analysis that shows the existence of four models in the sample. The first model depicts countries in which the introduction of formative assessment in teacher education and professional development took place more than 15 years ago, even if the practice was not important. In these countries, formative assessment practices were implicit in the official recommendations; however, they are not explicitly cited as a competency in the curriculum and not supported by formal meetings with students to discuss their assessment results. This "long-term and implicit" model comprises Denmark and Finland. The second model is also characterized by an implicit practice of formative assessment and a long-term presence of formative assessment in teachers' education but with a more recent introduction of formative assessment as an important object of teacher professional development. This "middle-term and implicit" model comprises Germany and the United Kingdom (England). The third model, "middle term and explicit," comprising France, Czech Republic, and Switzerland, is characterized by a middle-term introduction of formative assessment in teacher education and professional development (between 5 and 15 years ago), by an explicit presence of evaluation competencies in the curriculum and by the use of formal meetings with students to make sense of their assessment results. The last "short-term and explicit" model, the model of Cyprus, presents also explicit evaluation competencies in the curriculum and the use of formal meeting with students to make sense of their assessment results, but the introduction of formative/summative assessment in teachers' education and professional development is recent (less than 5 years ago). These four models are mainly defined by the older or recent introduction of formative assessment in in-service teachers' professional development and an explicit or implicit implementation of formative assessment in the classrooms.

Table 8.1 The four models of the combination among teacher education, nature of curriculum, and assessment

Models	Introduction of formative assessment in teachers' education	Introduction and amount of formative assessment in professional development CPD	Explicit presence of evaluation competencies in the curriculum	Formal meeting with students to make sense of assessment results
Implemented for long and non-explicitly (Denmark, Finland)	Long-term introduction in TE	Long term but with little amount in CPD	No	No
Implemented for some years and non-explicitly (Germany, United Kingdom)	Long-term introduction in TE	Middle term with an average to important amount in CPD	No	No
Implemented for some years and explicitly (France, Czech Republic, Switzerland)	Middle-term introduction in TE	Middle term with an average to important amount in CPD	Yes	Yes
Implemented recently and explicitly (Cyprus)	Middle-term introduction in TE	Short term but with little amount in CPD	Yes	Yes

The first and second model (long term and implicit, middle term and implicit) are associated with decentralized educational systems (Denmark, Finland, England, and Germany to some extent), while the two other models (middle term and explicit, short term and explicit) are associated with centralized (France, Cyprus) or school-centered (Czech Republic, Switzerland) educational systems (Lima et al. 2013).

These results are consistent with the model presented in Fig. 8.1 that links educational policies (in our study, the level of centralization of the system and, because of policies, curriculum specification) to teacher practices of formative assessment directly or by the means of both teacher education and teacher in-service professional development programs. However, these results come from a sample of only eight European educational systems and from teachers' practices that are indirectly described. Accordingly, they provide only some indications for a better understanding of the link that may exist between national policy and teachers' practice. The direct identification of teachers' practices through teacher interviews or classroom observations and the study of other educational systems may strengthen these results. More information about the organization of in-service professional development in each country may also specify the link between educational policies and teachers' practice.

However, our results allow us to envision possible ways to enhance the combination of formative and summative assessment in teacher practice through educational policies:

- Promoting day-to-day formative assessment in combination with summative assessment through teacher education and in-service development. As stated in the OECD report about effective teachers, teacher quality is the main factor of influence on student performance, which is potentially open to policy influence. As teacher quality is partly dependent on their education (OECD 2005, p. 26), teacher professional development seems to be a key factor for implementing formative assessment and its combination with summative assessment in teacher practice. Informal use of formative assessment and its integration in day-to-day assessment practices (Lima et al. 2015) promoted by the long-term presence of formative assessment as an object of teacher education and in-service teacher professional development seem to be favored in teacher-centered educational systems (as in Finland or Denmark).
- Providing teaching resources and guidelines for the implementation of formative assessment and its combination with summative assessment. Resources (e.g., textbooks) which help teachers in the implementation of formative assessment in their classrooms also seem to be an important means for the adoption of formative assessment practice, particularly in educational systems with a shorter-term promotion of formative assessment in schools (Lima et al. 2015).

Surveying European Use of Formative and Summative Assessment

The aim of this second part of this chapter is to complement the educational system-based perspective presented in the first part of the chapter by specifically addressing national research conducted in the ASSIST-ME partner countries with respect to the formative and summative use of assessment in science, technology, and mathematics education (STEM) (s. Chap. 3). In alignment with the objectives of the ASSIST-ME project, a specific focus thereby lies on inquiry-based teaching and learning approaches within the three domains (s. Chap. 2). The underlying assumption is that the potential success of any attempt to change the assessment practice within a country will not solely depend on the characteristics of the educational system. Of equal importance is how these conditions are implemented in practice. Results from educational research can provide helpful information not only about the challenges but also the opportunities of these implementations across different national contexts.

In order to illustrate country-specific characteristics regarding the assessment of inquiry-related competences, a survey was carried out by asking national experts from the fields of science, technology, and mathematics education. The survey consisted of ten open-ended questions that asked the national experts to summarize national research findings related to five overarching topics:

1. The role of formative and summative assessment for the teaching and learning of STEM
2. The relation between formative and summative assessment
3. Formative assessment practice
4. Challenges for the uptake of formative assessment
5. The role of assessment in supporting inquiry-based teaching and learning

As the experts were all involved in the ASSIST-ME project (see book introduction), a shared understanding of the terms formative and summative assessment could be assumed. In the following, the major findings from the reports from the eight European countries Cyprus, Czech Republic, Denmark, England, Finland, France, Germany, and Switzerland are summarized. Since the amount of available research varied considerably between countries, the intention of this section is to present spotlights of country-specific research that give a general sense of the situation of formative and summative assessment in European countries rather than to provide a comprehensive summary of the situation in each separate country. A complete description of the results can be found in Rönnebeck, Bernholt, Ropohl, Köller, and Parchmann (2013). The most interesting and striking similarity across countries is that in almost all countries, there has been little to almost no national research on formative assessment in general – or on inquiry competences in particular. For topics where no research findings existed in countries, some of the experts provided informed hypotheses regarding possible reasons for the lack of research. In Cyprus, for instance, a possible reason was seen in the fact that educational policy and teaching practice do not prioritize evidence-based research in their decisions. As a consequence, the potential of assessment data to inform policy and practice (and support learning) is often not considered. The centralization of the educational system as described in the first part of this chapter (see Fig. 8.2) was regarded as another possible cause. Teachers often lack the motivation to improve the quality of their teaching since such efforts are not rewarded by the system (e.g., teacher appointment and salaries are independent of qualifications and the quality of teaching).

The Role of Formative and Summative Assessment for STEM Teaching and Learning

With respect to the role that formative and summative assessments play in and for the teaching and learning of STEM, in almost all countries, summative assessment is considered to be predominant compared to formative assessment. In some countries like Switzerland or Germany, long traditions of summative assessment and grading exist. The same is true for Finland where students complete up to 50 tests per year, and the test results often provide the major source for grading. The character of the tests differs (e.g., from nationwide to statewide or even school-/teacher-based tests) in relation to the centralization of the educational system and, e.g., the school autonomy as described in the first part of this chapter (see Fig. 8.2). Within the last decade, however, even in systems where the authority lies with individual states or cantons, like Switzerland and Germany, a trend has been observed to establish a nationwide comparability of assessment tasks and results. The educational system in Germany, for instance, has shifted from an input to an output orientation. Nationwide, competence-oriented educational standards were implemented, and their attainment is monitored in regular intervals by national large-scale assessments. However, similar to Switzerland, the purpose of these large-scale assessments is to survey the system and not the individual student.

With respect to formative assessment, results from France show that the majority of teachers and students favor formative assessment or at least consider formative and summative assessment equally important (e.g., Issaieva et al. 2011). In Switzerland, mandatory guidelines exist in many cantons that explicitly mention formative assessment; however, no systematic surveys of formative assessment practice exist (which seems to be characteristic for the situation in many of the participating countries).

Following the seminal review by Black and Wiliam (1998) reporting on the positive effects on learning where formative assessment had been used in classrooms, several projects in England investigated the opportunities and challenges of implementing formative assessment in regular teaching practice. The results showed that such practices could be established in schools (e.g., Black et al. 2003; Wiliam et al. 2004); however, they required radical changes on the side of the teachers (e.g., Harrison 2013). A specific challenge that teachers encountered was to promote autonomy and self-regulated learning in their students (Marshall and Jane Drummond 2006). Moreover, teachers felt that the formative assessment often provided so much scaffolding that it was difficult to decide whether the learners could have been successful without this additional support.

The Relation Between Formative and Summative Assessment

In most countries, there is not much research or information about the interaction between formative and summative assessment available. Grades are regarded by some countries as a form of summative assessment that also has a potential for formative assessment purposes (see Chap. 3). A study in Germany comparing classroom-based assessments and standard-based tests found that thematically focused assessments – as needed for formative assessment – led to additional and specific information that could not be provided by summative assessments (Klieme et al. 2010). A research tradition investigating the relationship between formative and summative assessment, however, exists in England. In this tradition, formative and summative assessments are not regarded as two different types or forms of assessment. “In general terms, assessment is simply the production and interpretation of evidence of achievement. If this evidence is used to guide the next steps in progress, it is for learning [formative]; if it is used to sum up, judge, make decisions about progress so far, it is of learning [summative]” (Rönnebeck et al. 2013, p. 80). Teaching, learning, and assessment all need to be closely interlinked in the planning and implementation of any teaching program – otherwise tensions might be created or opportunities for improvement missed. One negative impact of the higher profile given to test-based results in England’s national curriculum assessment system has been shown to be not only a loss of assessment skill on the part of teachers but also a loss of confidence in their ability to make sound assessments of their students (Black et al. 2010, 2011). The given balance between school and teacher autonomy as described in the first part of this chapter (see Fig. 8.2) would seem to provide teachers with the opportunities to develop and practice formative assessment strategies. However, this affordance may be overwhelmed by this English national curriculum assessment system.

Formative Assessment Practice

Formal formative assessment seems not to exist in the investigated countries. A recent study in Denmark, however, found that when teachers assess their students, they have “an outspoken focus on learning and learning potential” and that most teachers assess “continuously and after the individual activity” (Rönnebeck et al. 2013, p. 26) – the most common forms of assessment are whole class conversations and written tests. A similar hypothesis in Finland assumes that opportunities for formative assessment exist in daily teaching practice where teachers might especially use short-term, informal formative assessment in teacher-student interactions. In Switzerland, where student and teacher attitudes toward different assessment methods have been investigated (Vögeli-Mantovani 1999), a high acceptance by teachers for oral feedback instead of grades, learning reports on progress, and student self-assessment has been observed. The acceptance among parents and students was also comparably high. The relatively high autonomy of teachers in Denmark, Finland, and Switzerland (see Fig. 8.2) may allow these teachers to individually include formative assessment strategies in their teaching.

With respect to the existence and research into the use of specific tools for formative assessments, countries differ significantly. Whereas in Finland and the Czech Republic, no such tools exist at all, in Denmark, many are available but very little research-based knowledge on how they are used exists. In Switzerland, formative assessment is systematically gaining in importance and has been supported by regulations (Vögeli-Mantovani 1999). Examples for formative assessment formats are rubrics, portfolios in mathematics, and textbooks fostering inquiry that include assessments. However, the gain in importance is not yet reflected in the daily practice in schools. Similar to Denmark, tools for formative assessment exist in Germany, but there is little research about their use. Recently, however, several studies investigated the use and effect of feedback in mathematics instruction (e.g., Rakoczy et al. 2013). The authors found no significant total feedback effects on interest and achievement development. There were, however, indirect effects on the development of interest via the perceived competence support and usefulness and on achievement development via the perceived usefulness. A mastery-approach goal orientation mediated the impact of feedback on the perceived usefulness.

In contrast, a Danish formative assessment instrument aimed at supporting students in performing inquiry processes has been used in physics. It was shown to increase the motivation, especially of girls, dramatically (Dolin 2002). Parallel to the ASSIST-ME project, a PhD thesis addressed the impact of formative assessment on students' self-regulation in the context of IBSE in France. The research aimed at analyzing the assessment practices of teachers and their effects on their students' self-regulatory process. Two approaches were compared. The first corresponded to formative assessment methods implemented by teachers in their daily practices without training. The second concerned the assessment practices implemented by these same teachers after a series of three workshops where they collaborated for designing teaching units comprising formative assessment tools and for gradually improving them. The results showed a better balance in the use of different formative assessment methods in the second situation, especially with respect to a greater empowerment of the students and a better account taken of peers as resources. Students also demonstrated more efficient self-regulation since they spent more time in elaborating problem-solving strategies and in being committed in the task (Lepareur 2016).

Challenges for the Uptake of Formative Assessment

Countries regarded different factors as impeding the uptake of formative assessment. These factors are mostly in line with results found in the international literature (Bernholt et al. 2013). A serious impediment in many countries is seen in teachers' beliefs about assessment as an instrument for generating grades and ranking students. Moreover, teachers often seem to have reservations toward formative assessment because they consider it laborious and difficult to implement (e.g., in Finland). A study in Germany points out that a dilemma between alternative assessment methods that aim at the contemplation of learning (like learning diaries) and student evaluation exists (Winter 2007). Students might not openly express their ideas, opinions, and problems if they know they will be evaluated. On the other hand, students might be demotivated if they put much effort into a portfolio, and this work does not contribute at all to their grades.

In Cyprus, research shows that although teachers seem to appreciate assessment as an integral component of teaching and a powerful means of enhancing the quality of teaching and learning, they nevertheless exhibit an inclination toward traditional assessment approaches that yield overall scores (Rönnebeck et al. 2013). Other aspects mentioned, for example, in a study from Switzerland, are a lack of time and a lack of teacher competence to differentiate between different levels of proficiency within a class (Smit 2009). In England, results from the assessment for learning (AfL) initiative provided insights into the challenges of a widespread implementation of formative assessment. While at the school level competing priorities, for example, demands for summative data and other issues of accountability, were found to be a major obstacle, the main drawback for teachers was in fully developing the dialogic classroom (Rönnebeck et al. 2013).

With respect to support that teachers need in order to implement formative assessment into their daily teaching practice, almost all countries agree on a general need for pre- and in-service teacher training. The historic variations in teacher education and professional development for the introduction of formative assessment provide some understanding of its current use (see Table 8.1). This training should address different aspects related both to learning and to assessment. Research from England indicates that teachers need to develop an in-depth pedagogical knowledge of how children learn and of their own pupils' learning needs (e.g., Watkins and Mortimore 1999). With respect to assessment, teachers need support to increase their assessment literacy. Research from Germany shows that a high diagnostic competence of the teacher positively influences his or her formative assessment practice (Klieme et al. 2010). Another important issue is the assessment of progress against personal rather than normative frameworks. Moreover, teachers need support to change their beliefs about assessment. Teacher perceptions about formative assessment are strongly influenced by how they were formed, the particular school contexts, and how they may affect practice (Sach 2012).

In this context, the importance of a strong relation between educational research and assessment practice is stressed by, for example, Denmark and the Czech Republic. An urgent need for concrete assessment tools is expressed in a study from Switzerland (Jundt and Wälti 2011). They found that ready-made mathematics units including rubrics for assessment encouraged teachers to assess complex (and therefore often neglected) competences. In Finland, a possible way to support teachers could be to involve textbook writers in the process because of the central role textbooks play in Finnish teaching and learning. From studies on school effectiveness, eventually, Cyprus concludes that mechanisms for internal evaluation need to be established and activities implemented that aim at improving teaching practice and the corresponding learning outcomes (e.g., Creemers and Kyriakides 2006).

The Role of Assessment in Supporting Inquiry-Based Teaching

No studies investigated whether assessment methods influence the uptake of inquiry in the respective countries. Switzerland, Finland, and the Czech Republic, respectively, state that inquiry is not used frequently, is uncommon, or is not a part of the regular instruction. One major reason for this is seen in the fact that inquiry is often not assessed in examinations (e.g., in Denmark and Finland) and is thus perceived as auxiliary to core teaching. However, it is assumed that with more broad support for teachers in formative and summative assessment through both preservice and continuing teaching education (see Table 8.1), inquiry could gain significance. In Denmark, there has been some research on how summative and formative assessment could be used to promote learning in inquiry education (the "assessment dialogue" Christensen 2004). In Germany, the implementation of educational standards (which include inquiry competences) required the development of competence models – and thus assessment items – for inquiry for monitoring purposes. In Denmark, a new examination for lower secondary includes attention to inquiry processes.

Overall

The diverse use and relationship of formative and summative practices in the eight European countries provide an overview of irregular connections in the Fig. 8.1 model of the relationships between educational policies and teacher practices. The lack of a European-wide standard for using formative and summative assessment along with inquiry-based STEM teaching and learning provides both challenges and opportunities. The diversity between and within countries gives us a number of "natural experiments" where various uses of assessment tools and teacher training can be compared and contrasted for insights into research-based changes. Our research is useful in identifying some of these "experiments" which might be worthwhile exploring with further research.

The next chapter moves from this general overview of Pan-European perspectives of assessment to a deeper focus on teacher perspectives collected from teacher questionnaires as well as interactions with teachers in eight countries.

References

Bernholt, S., Rönnebeck, S., Ropohl, M., Köller, O., & Parchmann, I. (2013). *Report on current state of the art in formative and summative assessment in IBE in STM - Part I.* (ASSIST-ME Report Series No. 1). Copenhagen: University of Copenhagen.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice, 5*(1), 7–74.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. Buckingham: Open University Press.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*(2), 215–232.

Black, P., Harrison, C., Hodgen, J., Marshall, B., & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*(4), 451–469.

Christensen, T. S. (2004): *Integreret Evaluering – En undersøgelse af den fagligt evaluerende lærer-elevsamtale som evalueringsredskab i Gymnasial Undervisning*. PhD Dissertation, University of Southern Denmark. Retrieved from http://static.sdu.dk/mediafiles/Files/Om_

SDU/Fakulteterne/Humaniora/Phd/afhandlinger/2005/Afhandlinger%2042_spanget%20pdf.pdf

Cohen, D. K., & Hill, H. C. (2000). Instructional policy and classroom performance: The mathematics reform in California. *Teachers College Record, 102*(2), 294–343.

Creemers, B. P., & Kyriakides, L. (2006). Critical analysis of the current approaches to modelling educational effectiveness: The importance of establishing a dynamic model. *School Effectiveness and School Improvement, 17*(3), 347–366.

Cross, D., & Lepareur, C. (2015). PCK at stake in teacher–student interaction in relation to students' difficulties. In *Understanding science teachers' professional knowledge growth* (pp. 47–61). Rotterdam: Sense Publishers.

Dolin, J. (2002). *Fysikfaget i forandring – læring og undervisning i fysik i gymnasiet med fokus på dialogiske processer, autenticitet og kompetenceudvikling*. PhD Dissertation, Roskilde University. Retrieved from http://rudar.ruc.dk/handle/1800/1645

Grangeat, M., & Hudson, B. (2015). 12. A new model for understanding the growth of science teacher professional knowledge. *Understanding Science Teachers' Professional Knowledge Growth*, 205.

Hanushek, E. A., & Woessmann, L. (2014). Institutional structures of the education system and student achievement: A review of Cross-country economic research. In *Educational Policy Evaluation through International Comparative Assessments* (pp. 145–175). New York: Waxmann Verlag.

Harrison, C. (2013). Collaborative action research as a tool for generating formative feedback on teachers' classroom assessment practice: The KREST project. *Teachers and Teaching, 19*(2), 202–213.

Issaieva, É., Pini, G., & Crahay, M. (2011). Positionnements des enseignants et des élèves du primaire face à l'évaluation: une convergence existe-t-elle? *Revue française de pédagogie, 3*, 5–26.

Jundt, W., & Wälti, B. (2011). Mathematische Beurteilungsumgebungen; 7. *Schuljahr. Zug: Klett und Schulverlag*.

Klieme, E., Bürgermeister, A., Harks, B., Blum, W., Leiß, D., & Rakoczy, K. (2010). *Leistungsbeurteilung und Kompetenzmodellierung im Mathematikunterricht. Projekt Co2CA* (pp. 64–74).

Lepareur, C. (2016). L'évaluation dans les enseignements scientifiques fondés sur l'investigation: Effets de différentes modalités d'évaluation formative sur l'autorégulation des apprentissages. Thèse de doctorat. Université Grenoble Alpes. France

Le Roux, B., & Rouanet, H. (2010). *Multiple correspondence analysis* (Vol. 163). Los Angeles: Sage.

Lima, L., Cross, D., Nakhili, N., Grangeat, M., & Bressoux, P. (2013). Educational system factors influencing student assessment methods in science, technology and mathematics education (Delivrable 3.4). Report for the FP7 UE project Assess Inquiry in Science, Technology and Mathematics Education -ASSIST-ME. http://assistme.ku.dk/deliverables/wp3/Deliverable_3.4_-_Part_1.pdf

Lima, L., Cross, D., Grangeat, M., & Nakhili, N. (2015, January). Évaluation Formative dans l'enseignement des sciences et mathématiques dans 8 pays européens: résultats de la première étape du projet ASSIST-ME. In *Conditions enseignantes, conditions pour enseigner: réalités, enjeux, défis*.

Marshall, B., & Jane Drummond, M. (2006). How teachers engage with assessment for learning: Lessons from the classroom. *Research Papers in Education, 21*(02), 133–149.

OECD. (2013). *Education policy outlook*. Germany: OECD. http://www.oecd.org/edu/EDUCATION%20POLICY%20OUTLOOK%20GERMANY_EN.pdf.

Organisation for Economic Co-operation and Development, Santiago, P., & Source OECD. (2005). Teachers matter: Attracting, developing and retaining effective teachers. Organisation for Economic Co-operation and Development.

Rakoczy, K., Harks, B., Klieme, E., Blum, W., & Hochweber, J. (2013). Written feedback in mathematics: Mediated by students' perception, moderated by goal orientation. *Learning and Instruction, 27*, 63–73.

Rönnebeck, S., Bernholt, S., Ropohl, M., Köller, O., & Parchmann, I. (2013). *National reports of partner countries reviewing research on formative and summative assessment in their countries*, *No. D 2.3*. Kiel, Germany: IPN.

Sach, E. (2012). Teachers and testing: An investigation into teachers' perceptions of formative assessment. *Educational Studies, 38*(3), 261–276.

Smit, R. (2009). *Die formative Beurteilung und ihr Nutzen für die Entwicklung von Lernkompetenz: eine empirische Studie in der Sekundarstufe I*. Schneider-Verlag Hohengehren.

Vögeli-Mantovani, U. (1999). Mehr fördern, weniger auslesen. *Zur Entwicklung der schulischen Beurteilung in der Schweiz. Trendbericht SKBF*, (3).

Watkins, C., & Mortimore, P. (1999). Pedagogy: What do we know. *Understanding pedagogy and its impact on learning*, 1–19.

Wiliam, D., Lee, C., Harrison, C., & Black, P. (2004). Teachers developing assessment for learning: Impact on student achievement. *Assessment in Education: Principles, Policy & Practice, 11*(1), 49–65.

Wilson, S. M., & Berne, J. (1999). Chapter 6: Teacher learning and the Acquisition of Professional Knowledge: An examination of research on contemporary Professional development. *Review of Research in Education, 24*(1), 173–209.

Winter, F. (2007). Fragen der Leistungsbewertung beim Lerntagebuch und Portfolio. *Lernprozesse dokumentieren, reflektieren und beurteilen. Lerntagebuch und Portfolio in Bildungsforschung und Bildungspraxis* (pp. 109–129).

Chapter 9
Teacher Perspectives About Using Formative Assessment

Robert Evans, Rose Clesham, Jens Dolin, Alena Hošpesová, Sofie Birch Jensen, Jan Alexis Nielsen, Iva Stuchlíková, Sofie Tidemand, and Iva Žlábková

Teacher Questionnaire

Pan-European views of the project teacher's development were sampled with a questionnaire to track overall teacher change in seven countries. As described in this book's introductory chapter, teachers were participants in the ASSIST-ME project aimed at investigating the uses of formative assessment strategies along with inquiry-based teaching. Both the pre-study and final questionnaires were distributed to all participating teachers. The purpose was to gain insight into the development in teachers' perceptions of the relevance of inquiry-based education (IBE), formative assessment, competence-oriented teaching and their confident use of these aspects of assessment compared to that of control groups composed of similar collegial teachers. All of these aspects were examined in three dimensions: the teachers were asked how often they used the different methods in their teaching, how important they thought they were and how competent they felt in employing the aspects in their teaching.

The questionnaire assessed confidence by asking about respondents' self-efficacy when using formative assessment in their classroom practices. There were five questions regarding the use and perception of IBE: how teachers work with engaging their students through interesting or unusual questions and how they get them to identify investigable questions, plan investigations, have collaborative discussions and work with real-life problems.

R. Evans (✉) • J. Dolin • S.B. Jensen • J.A. Nielsen • S. Tidemand
Department of Science Education, University of Copenhagen, Copenhagen, Denmark
e-mail: evans@ind.ku.dk

R. Clesham
Pearson, London, UK

A. Hošpesová • I. Stuchlíková • I. Žlábková
Jihočeská univerzita, University of South Bohemia, České Budějovice, Czech Republic

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_9

Questions about competences focused on six competences: investigation in science, problem-solving in mathematics, design in engineering, argumentation, modelling and innovation across all subjects.

The questionnaire was distributed to all local working group (LWG) teachers across partners as well as to a control group of teachers for each LWG during the first round of implementation. This resulted in 110 LWG sets of answers and 57 control group responses by the end of the first implementation round (see Table 9.1). Only at the end of the third implementation round 2 years later, the identical questionnaire was distributed once more to all respondents who answered the questionnaire in the first round. However, partners where the first-round response rate was too low to yield meaningful results were omitted from the final questionnaire round. We did not attempt to classify the nonrespondents, thus reducing some of the generalizability of this study. This means that by the end of the 2 years of implementations, the questionnaire was distributed to 101 LWG teachers and 52 control group members from seven partners (two from France).

Except for the 12 items relevant to self-efficacy, the questionnaire responses have not been analysed quantitatively due to the incomplete pre- and post-data sets. The self-efficacy items are an exception since it is useful to examine quantitative changes in single questions across countries.

Changes in items regarding IBE teaching suggest that developing inquiry-based teaching units was the focal point for many teachers throughout the implementations. This is supported by the work on the LWG meetings in Denmark where substantial amounts of time, especially in the first implementation round, were spent developing inquiry teaching units. The questions about assessment also hint at some change in perceptions about formative aspects of feedback and questions related to specific ways of assessing students. However, clear trends across cultures were not evident.

Table 9.1 A summary of the response rates from the pre- and post-questionnaires from both participant teachers (LWG) and those not participating (control). Cyprus and the United Kingdom were dropped from the final survey due to low response rates in the first round

	Pre-round response rates				Post-round response rates			
Country	LWG	Rate	Control	Rate	LWG	Rate	Control	Rate
Denmark	19	100%	7	58%	16	94%	4	57%
Germany	12	86%	6	100%	4	40%	2	33%
Cyprus	7	41%	3	9%	0	NA	0	NA
Switzerland	20	95%	15	83%	16	89%	13	87%
France (A)	17	68%	NA	NA	11	65%	NA	NA
United Kingdom	2	12%	2	33%	0	NA	0	NA
Finland	8	89%	2	67%	8	89%	1	50%
France (B)	7	64%	5	83%	6	67%	3	60%
Czech Republic	18	86%	17	81%	14	82%	12	75%
Totals	110		57		75		35	

First Perspective: Changes in Teacher Self-Efficacies While Using Formative Assessment

Teacher Self-Efficacy Beliefs

Introducing teachers to strategies for formative assessment and facilitating their use inevitably includes consideration of the importance of changes in teacher's personal beliefs about their capacities to adapt these approaches successfully to their classrooms. Without concomitant belief changes, teachers are less likely to successfully use less familiar methods and to continue to improve their skill at mastering them. Our efforts to introduce and facilitate the formative assessment methods—teachers' written feedback to students, classroom peer-to-peer assessment, 'on-the-fly' teacher feedback to students and structured assessment dialogue in the classroom—were grounded in strategies for enhancing teacher beliefs about their abilities to use them (see this book's introductory chapter for details about these formative assessment methods). This section of *teacher perspectives* looks at changes in the teacher capacity belief of self-efficacy while using strategies for enhancing these beliefs given the circumstances in which formative assessment methods were trialled. The findings of the teacher questionnaire which addressed capacity self-efficacy beliefs among study teachers and those not a part of the study, before and after using the formative assessment methods, are shared.

The Role of Self-Efficacy

'Self-efficacy' is the capacity belief based on Albert Bandura's work that posits that such beliefs '… contribute significantly to human motivation and attainments' (Bandura 1992). Beliefs in one's own ability to manage and implement a given challenge, such as using formative assessment with a class not accustomed to it, are instrumental in meeting the challenge (Bandura 1992). We each hold these expectations about our future ability to perform tasks based on previous life experiences. As teachers grow professionally by strategies to facilitate learning, they feel more confident about replicating those teaching methods that are successful and less confident about trials either which do not succeed or for which they receive no evidence of their success.

For example, if teachers attempt to use peer-to-peer assessment for the first time they, based on previous experiences with unfamiliar methodologies, typically will have some doubts about their chances for success. At the same time, these doubts may be balanced by positive expectations from previous successful experiences with new strategies. Considering these doubts, positive expectations and their current teaching environment, teachers will have individual levels of self-efficacy about how successful they expect to be. Contributing to this level of self-efficacy is a teacher's general confidence as a teacher at attempting new methods of instruction. However, general self-confidence is not the same as self-efficacy beliefs, since efficacy beliefs are targeted at specific future behaviours, whereas self-confidence is non-specific. We can simultaneously have a high confidence in our teaching ability yet low self-efficacy when confronted with a specific teaching demand such as using an unfamiliar kind of formative assessment. A higher general confidence can positively influence our self-efficacies, but its effect may be diminished as attempts at a given task provide task-specific feedback about our abilities (Bandura 1997).

Therefore, as teachers implement, in this case, peer-to-peer formative assessment, the feedback they get about how the trial goes will either raise or lower their self-efficacies for peer-to-peer formative assessment. If their self-efficacies for using peer-to-peer assessment were rather low to begin with and if they cannot see evidence of successful use, they may be averse to trying the strategy again. However, if to begin with, their self-efficacies were robust and they get some credible positive feedback, their self-efficacy for using peer-to-peer assessment is slightly raised, and the chances of them tying it again are enhanced.

Essential for any change in self-efficacy is authentic feedback about the degree of success for a teaching action. Sources of such feedback include self-reflection, student activation and various indicators of student success as well as perspectives provided by colleagues and/or other observers.

The perceived validity of these sources of feedback success is important since if teachers do not have confidence in how well their use of feedback methods is working, then they are not likely to use them in the future.

Opportunities for Enhancing Self-Efficacy

Albert Bandura (1995, 1997) identified four experiential factors that determine self-efficacy expectations. He categorizes them as 'enactive mastery experience', 'vicarious experience', 'verbal persuasion' and 'affective states' (Bandura 1997). Mastery experiences are past efforts at the same or similar teaching tasks from which teachers judge for themselves how well they were able to achieve a 'novel' teaching method. Their self-reflections about the extent to which they succeed in implementing something different strongly influence their future personal expectations for using this teaching method again. In the case of facilitating peer-to-peer assessment in their classroom, how teachers judge student experiences and other indicators of performance influences how they feel about using peer-to-peer feedback.

The influence of mastery experiences on future behaviour is high. Consequently, misinterpretations of success are especially important to avoid when trying 'new' teaching methods. For example, if a teacher's self-reflections tend to be very critical, then they may avoid follow-up trials based on these harsh judgements about their first successes. Conversely, teachers who are more objective and perhaps have a generally higher self-confidence may note where they succeeded and what to change the next time. Teachers who get feedback on the use of unfamiliar teaching methods, such as by objectively examining student outcomes, are more likely to have valid changes in self-efficacies.

Teacher self-efficacies are also influenced vicariously through seeing how their peers handle a trial of, for example, peer-to-peer student assessment. When they work in a teaching group to implement such a method and then discuss with their colleagues the degree of success, they adjust their own self-reflections to those of others with whom they compare themselves. These comparisons influence individual self-efficacies, particularly among teachers who have had little experience of their own and look to peers with more experience for indicators of success (Bandura 1997).

Similarly, 'verbal persuasion' that teachers receive from those who they respect such as other teachers, administrators or university faculty has an effect on their individual perceptions of self-efficacy. These sources of verbal coaching, when valid and not just kindly supportive, influence self-efficacies. The veracity of the feedback from significant others can also serve to align self-reflections with the reality of a teacher's success and hence influence self-efficacy. Even when this feedback is negative and therefore likely to depress self-efficacy, if it is combined with suggestions for ameliorating the difficulties, the negative effects may be limited (Bandura 1997).

The affective perceptions of teachers attempting unfamiliar teaching methods, perhaps like peer-to-peer assessment, can have a negative influence on their self-efficacy in that their performance may be hindered by negative affective messages that reduce the otherwise positive feedback of their efforts. The disequilibrium of attempting untried teaching methods when compared to usual procedures can increase negative affective effects for some teachers and hence depress their self-efficacies. This is particularly true for inexperienced teachers and for teachers who depend only on their own self-reflections. Vicarious feedback from observing peers and valid encouragement from respected colleagues can help reduce attributions to inability.

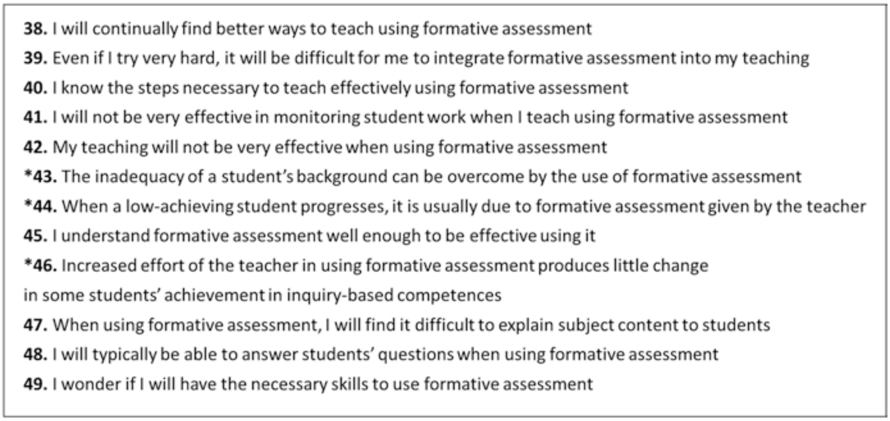
The Environment in Which Formative Assessment Is Used

With a goal of testing the usefulness of formative assessment methods, we included with our implementations opportunities for all of Bandura's methods for self-efficacy change (Bandura 1995, 1997) to be used. For each of the trials in eight country sites, local working groups (LWGs) of experienced teachers were formed. Over the course of 2 years, before, during and after feedback implementations, the teachers met with one another and project leaders to plan activations and discuss the results. During implementations, LWG teachers tried the assessment methods multiple times and reflected both individually and as local groups on the results of their trials. After concluding their trials, the LWGs along with the project leaders in each country met to discuss the outcomes. Before and after concluding their project work, all of the teachers in the LWGs, as well as teacher colleagues in each country who did not participate in the trials, answered questions about their experiences on a standard teacher questionnaire that was translated into each country's language.

These teacher trials with formative assessment methods were designed to provide opportunities for 'mastery' of the less familiar methods since they were tried multiple times with intervals between for reflection and feedback. Since the project engaged experienced teachers, their self-reflections after repeated lesson trials were likely to have influenced their self-efficacies for each of the methods they used. In addition, since they met with peers in their LWGs before, during and after trials, the opportunities for vicarious influences from the group were frequent. Concomitantly, there were opportunities for influential members of the LWGs as well as project leaders to affect teacher self-efficacies through social persuasion at meetings where the processes and results of the trials were discussed. It was hoped that the engagement of experienced teachers who volunteered for the project, along with frequent opportunities for LWG reflection and feedback from project leaders, helped control any possible negative affective consequence effects of trying new classroom teaching methods.

How Self-Efficacy Was Assessed with the Teacher Questionnaire

The teacher questionnaire was administered to all participating teachers as well as to a sample of similar teachers from the same countries who were not involved with the study, both before and after trials with the formative assessment methods. It contained 12 items whose aim was to assess the self-efficacy of teachers unfamiliar with various formative assessment methods. These items (see Fig. 9.1) were derived from a commonly used international instrument for science teaching self-efficacy (Enochs and Riggs 1990; Bleicher 2004). The two constructs of self-efficacy and outcome expectations, both of which are theoretically part of capacity beliefs (Bandura 1997), were both represented among the questionnaire items. Nine questions required teacher efficacious projections of their future capability at performing a given teaching action; three questions (*s in Fig. 9.1) queried the likely outcome of teaching efforts given the circumstances in which the teaching occurred. Both perspectives are indicative of the likelihood that a teacher who has experienced using these formative assessment methods is to use them again since it is not only their judgement of their own ability to use a method (self-efficacy) but also whether, given the circumstances of their classroom and students, they would actually be able to implement it (outcome expectations).



38. I will continually find better ways to teach using formative assessment

39. Even if I try very hard, it will be difficult for me to integrate formative assessment into my teaching

40. I know the steps necessary to teach effectively using formative assessment

41. I will not be very effective in monitoring student work when I teach using formative assessment

42. My teaching will not be very effective when using formative assessment

***43.** The inadequacy of a student's background can be overcome by the use of formative assessment

***44.** When a low-achieving student progresses, it is usually due to formative assessment given by the teacher

45. I understand formative assessment well enough to be effective using it

***46.** Increased effort of the teacher in using formative assessment produces little change in some students' achievement in inquiry-based competences

47. When using formative assessment, I will find it difficult to explain subject content to students

48. I will typically be able to answer students' questions when using formative assessment

49. I wonder if I will have the necessary skills to use formative assessment

Fig. 9.1 Teacher questionnaire items aimed at assessing self-efficacy (*outcome expectations)

Teachers in each of six countries responded to five-point Likert scales for each of the 12 items. The polarities of some items were reversed to reduce response sets. Eighty-three teachers out of a possible 95 (64 participating teachers and 31 control teachers) in six countries responded to both the pre- (administered in 2014) and post-teacher (2016) questionnaire administered in their own languages by local researchers. The full response rate of 87% of teachers completing both pre- and post-questionnaires was high. This high response rate can be attributed to the fact that the teachers were all known to researchers who had worked with them for several years and the surveys were not anonymous. There may have been some bias since the desire of these participating teachers to 'please' the researchers may have affected scores. On the other hand, the 2 years between pre- and post-administrations would have made it unlikely that teachers recalled their pre-responses when completing the post-questions. The 33% respondents who were control teachers were subject to as much bias since they had no participation or relationship to the researchers and probably did not recall their pre-responses when completing the post-survey.

Since the 12 items in the questionnaire do not represent a standardized instrument for collectively measuring self-efficacy, the individual item results are more useful in assessing change than aggregated scores. Because self-efficacy is an individual's capacity belief, summative data for all six reporting countries is more useful than individual or country data in judging the potential of these experienced teachers to raise their self-efficacies while using the formative assessment methods. Individual and country changes in self-efficacy have the potential to inform individual and country success with these methods. Consequently, we chose an overall cross-country perspective to gain general feedback on the trials of formative assessment.

Changes in Self-Efficacy Beliefs While Using Formative Assessment Methods

A hypothesized outcome of the trials of assessment methods was for positive changes in teacher's self-efficacies to occur when using the given formative assessment methods. The results of the 12 questionnaire items about self-efficacy provide relevant indicators. Table 9.2 contains the changes in mean scores on the 12 pre- and post-items for the project teachers (LWGs) and other teachers (control groups). Overall (means), there were no changes in self-efficacy for the project teachers (+0.06 out of 5 points), while the collegial teachers (control groups) who had no exposure to the formative assessment methods of the project reduced their self-reported efficacies (−0.49 out of 5 points) over the course of the study. Since these 12 questions did not comprise a comprehensive and validated instrument, it is more useful to look at changes in individual items since each assesses a different aspect of capacity belief. Seven items 38, 39, 40, 41, 42, 45 and 49 (see Table 9.2) directly address teacher projections about their future use of formative assessment.

The projections in these seven questions are not confounded by the outcome expectations of items 43, 44 and 46, since these three questions ask teachers about characteristics of their teaching situation which may influence their self-efficacy, but they do not assess the self-efficacies directly associated with the use of the four formative assessment strategies of the project. Nor are they tangential to using formative assessment, as are items 47 and 48, which address teacher content knowledge along with formative assessment.

For the seven items (38, 39, 40, 41, 42, 45 and 49) directly addressing self-efficacy, questions 39, 40, 41, 42 and 49 all showed significant differences in the pre- to post-changes for the two groups of teachers. In each of these five questions which allowed teachers to indicate their future confidence in using formative assessment methods, the LWG teachers were more confident and the control teachers less so. Even for question 49 where teachers made a general assessment of whether they '… will have the necessary skills to use formative assessment', the unchanged pre- to post-responses of the LWG teachers (−0.02) compared favourably with the drop in self-efficacy (−0.8) for non-participating teachers. The non-significant differences in pre- to post-changes for the two groups for questions 38 and 45 may indicate an overall positive outlook for using formative assessment as compared to responses to the five questions (39, 40, 41, 42 and 49) which were more specific and less based on a general future projection.

Table 9.2. Changes in the mean pre- to post-scores for 12 items (Fig. 9.1), for local working groups and for controls for six countries (outcome expectation questions shaded). Five-point scale LWG n = 54; control n = 29

	#38	#39	#40	#41	#42	#43	#44	#45	#46	#47	#48	#49
LWGs	0.04	0.27	0.50	0.58	0.60	-0.21	0.25	0.46	0.35	0.63	0.01	-0.02
Control Groups	-0.1	-0.19	-0.31	0.296	0.019	0.09	-0.31	0.3	-0.8	0.438	-0.65	-0.8

Self-Efficacy Beliefs as Indicators of Teacher Change

Higher teacher self-efficacies have been associated with the use of inquiry methods and student-centred teaching approaches such as provided by formative assessment (Czerniak 1990). Therefore, some evidence of our expectation that project teachers would increase their self-efficacies for using formative assessment methods along with inquiry would be a positive indicator of the potential of introducing these formative assessment methods into classrooms. If teachers' experiences with innovative formative assessment had significantly decreased their self-efficacies, then we would need to re-examine our procedures for their introduction into classrooms. The influence of teacher self-efficacy beliefs on teacher's roles, planning, lessons and student achievement is strong (Tobin et al. 1994). Therefore, positive changes in self-efficacies for experienced teachers trying unfamiliar methods such as innovative formative assessment provide alignment between these methods and those associated with other student-centred approaches in that they all are associated with higher self-efficacies. The observation that experienced teachers in this study had increases in self-efficacy when using innovative formative assessment methods provided encouragement for further efforts to introduce them to classrooms.

Second Perspective: Changes in Teachers Subjective Theories of Assessment—A Czech Perspective

This section on teacher's perspectives uses teacher interviews and case studies to look in depth at the challenges which implementation of formative assessment into everyday teaching faces. One frequently mentioned obstacle is a lack of appropriate support for teachers (e.g. Bernholt et al. 2013; Brown 2004). To arrange the support properly, especially for teachers who introduce new forms of assessment, it is important to understand the teacher's perspectives and their expectations related to assessment and its functions and values in teaching.

The research on practice of formative assessment is rather broad in scope (c.f. Bernholt et al. 2013); it covers various forms of assessment and various educational settings. For example, Hogson and Pyle (2010) summarized research on assessment for learning in primary science. Their report reviewed several different contexts of research, including the development of self-assessment skills, the use of different techniques for eliciting peer assessment, the links between feedback from peers and from teachers and the relationship of a formative classroom climate to peer- and self-assessment activities.

However, the current literature does not cover the issue of changes in teachers' conceptual understanding of formative assessment. With this incomplete understanding of teacher perspectives about formative assessment in inquiry teaching, especially at the primary school level, researchers in Czech Republic focused on further understanding of the teachers' points of view.

When implementing or upscaling new practice, teacher understanding of their own practice is challenged (Whitehead 1989). It has been argued that formal theories of educational processes are very often useless for eliciting the change in the practice of teaching unless they are constructed from authentic teacher experience. This view is reflected in Korthagen et al. (2001) concept of personal practical knowledge. It is knowledge that builds on gestalts of experience but can be reflected and is accessible to conscious description and communication, which helps in further refinement of it in interpersonal settings. Another term, which has become popular in the German-speaking countries from around the 1980s of the twentieth century, is subjective theory. The term ‘subjective theories’ (ST) has been used to describe the fact that humans construct their own theories while constantly reflecting on the reality they perceive (e.g. Groeben et al. 1988). Subjective theories which often arise spontaneously, under pressure and without conscious control, have an argumentative structure which is implicit, comprise more liberal assessment criteria, etc. (Janík 2005, p. 478). Subjective theories of teaching help to understand teaching as a series of deliberate, reflected actions. Most teachers develop subjective theories that allow them to justify their actions during class (c.f. Edmondson 1998). It can be assumed that teachers develop particular parts of their subjective theories when introducing new practice. The confrontation of their new teaching experience with former theoretical knowledge that has influenced their initial expectations leads to a gradual refinement of their subjective theories and thus further influences implementation of the new techniques.

Objectives of the Study of ‘Teacher Perspectives’ in Czech Republic

In Czech Republic, formative assessment is seen as more or less embedded in common Czech teaching culture, but typically only in the forms where the feedback is provided by the teacher (see also Chap. 4: On-the-fly assessment). Peer assessment is perceived mostly as a supplementary option (e.g. Košťálová and Straková 2008) probably also due to the fact that peer interactions as a form of learning communication are not frequently used (Šeďová et al. 2012).

However, as peer assessment is a promising but also rather challenging method of assessment in inquiry-based lessons in elementary mathematics (Hodgen and Marshall 2005), Czech researchers investigated teachers’ subjective theories of formative assessment and their development over time during a trial implementation of formative peer assessment in inquiry-based elementary mathematics.

How Perspectives of Teachers Were Assessed

A working group of six elementary teachers participated in this study. Together with researchers, teachers developed inquiry-based teaching units of primary mathematics, where formative assessment instances were implemented and reviewed. The teachers in pairs of two taught the teaching units in second-, fourth- and fifth-grade classes. The units consisted of a sequence of 4–6 inquiry-based tasks, implemented mostly in 2-h blocks.

Teachers reflected upon their subjective theory of formative assessment before and after the experimental teaching units. The reflection was elicited and organized by the Struktur Lege Technik (SLT, Scheele 1992). SLT is a structured interview that enables externalization of teachers' subjective theories of formative assessment in inquiry-based learning and consequently allows their description and clarification.

The first step whose purpose was to determine the content of the relevant cognitions was done by conducting a semi-standardized interview. As formative assessment is a relatively new concept for teachers, the researchers also offered several metaphors to facilitate broad conceptualizations. The second step, which involved reconstructing the structure of the subjective theory, was facilitated by providing paper cards bearing the main concepts and ideas expressed by the teacher in the previous interviews. The teachers were asked to organize the cards and describe the connections between them. The final step was an overall description of teacher understanding of formative assessment as depicted by explicated subjective theory. This process of explication of subjective theory of formative assessment was done twice, before and after the experimental teaching period that lasted for 8–10 weeks. Besides reconstruction of subjective theory of formative assessment, the researchers collected teachers' commentaries after each enacted lesson.

How the Interviews Were Analysed

Data were analysed in a qualitative manner. The subjective theories were first analysed by thematic coding. The initial subjective theory and the one created after the period of experimental teaching were compared. An inductive approach (Boyatzis 1998) where thematic codes were organized within a template (Crabtree and Miller 1999) was used to search for changes in teachers' subjective theories. The changes were analysed first as individual case studies, and then the whole group of teachers was studied in search for similarities.

What Was Learned from the Czech Teachers' Perspectives

The reflective interviews with teachers after the completion of particular lessons showed that implementation of peer assessment during inquiry-based tasks in mathematics was rather difficult and challenging for both teachers and students. The main problems that were identified by the teachers were time and resource demands (worksheets, assessment tools, teacher assistant time, etc.) and novelty of the assessment method. The experimental teaching thus provided an important new experience, which could stimulate the development of teachers' subjective theories of assessment.

Thematic coding was led by expectations driven by the inquiry and formative assessment literature (Black and Wiliam 2009; Eastwell 2009). The main thematic codes were benefits of formative assessment, obstacles of implementation, relationships between formative assessment and inquiry-based learning, relationships of formative and summative assessment within inquiry, advantages and disadvantages of formative assessment for students, learning tasks relevance for formative assessment and teachers' roles in formative assessment.

The development of the six teachers' subjective theories was summarized as case descriptions. Following are two examples, showing the difference in depth of conceptual understanding of formative assessment and subsequent views on prospective actions which should be taken for further implementation of formative assessment in a teacher's practice. These two teachers worked as a pair in the first round of experimental teaching; they taught in parallel classes, in the same school, and they discussed their teaching units together. Though they worked in similar conditions and have similar lengths of teaching practice and similar limited experience with inquiry-based teaching and no previous experience with formative peer assessment, their subjective theories of formative assessment developed differently.

Case Studies

Teacher A

This fifth-grade teacher had no previous experience with formative assessment, less than 5 years of practice and taught the topic 'big numbers'.

Before she started to use formative methods of teaching, Teacher A expected that the use of formative assessment in inquiry-based math lessons would provide effective continuous feedback for students—contrary to summative marks ('…when the students see the marks, they never see the mistakes. The mark means end of the work.'). She felt uncertain whether she would master inquiry and formative assessment together. As a main means of support, she considered worksheets/forms/rubrics for students. The main obstacle for the implementation of formative assessment which she saw was in parents' views on learning ('…only working with the student book is seen as sound learning, anything else is seen as entertainment or relaxation').

During the teaching trials, she realized that students frequently were not able to provide effective feedback for their peers, and she did not know how to help them. She also reflected that she had difficulties in summing up the lessons in a final whole class discussion, which could institutionalize the new piece of knowledge (Brousseau & Novotná 2012). After the experimental teaching, she reconstructed her subjective theory of formative assessment and revealed that she perceived inquiry-based learning as more important for her attitude towards teaching than formative assessment. The salient problem of formative assessment in inquiry-based lessons for her was that students were not able to assess the work of their peers. She did not find the prepared worksheets to be a valuable support for her formative assessment trials. Instead, she would like to observe a more experienced colleague's teaching. Further, she saw the importance of development of the students' assessment skills. She believed that students value more peer assessment than the assessment provided by the teacher. The parents' opinion was still seen as an obstacle ('… I do not know how I could defend the time we have spent on it and explain to parents that we did not get enough practice of the tasks in the student book').

Overall, Teacher A expected in correspondence with theory that formative peer assessment would provide continuous feedback to her students, but she found it very difficult to facilitate peer assessment in a way that will lead to an effective and formative feedback. She expressed a need for support. She did not perceive formative assessment as a natural ingredient of inquiry-based learning, but rather as something that is additive to inquiry and that she could therefore concentrate on later when she masters teaching inquiry and the anticipated reluctance of parents. As a whole, she especially considered the benefit that formative assessment can bring to her teaching, paid less attention to the perspective of student learning and dealt in greater extent with classroom external factors (support for teacher, attitude of parents).

Teacher F

This fifth-grade teacher also had no previous experience with formative assessment, practised teaching for 15 years and, as Teacher A, taught the topic 'big numbers'.

Teacher F saw formative assessment as a tool which should help students not only in learning to assess (self and the peers) but also in mastering the curriculum. The goals of formative assessment in inquiry education which she foresaw before the trials were that the students would learn about peer assessment and how it should meaningfully be provided for the recipient and what components of peer assessment are necessary for the recipient to get enough hints to see how to proceed. She thought that the teachers' role is the most important in the implementation of formative assessment, and thus it is important that the teacher has enough knowledge about formative assessment and is willing to use it. The teacher has to communicate to students what is necessary for good formative assessment, to make clear to students the principles of solution or criteria for assessment. Some teacher training therefore should precede the implementation of formative assessment in IBE, and also assistence in classes would be helpful, as the implementation is time demanding and not all teachers have a proper readiness. She did not see any problem in persuading the parents that this way of work would be productive; neither did the large number of students in the classroom seem to be a problem for her. On the other hand, she believed that formative assessment is more fruitful than summative assessment and understood that formative assessment cannot replace summative assessment.

During the trials, she appreciated cooperation with the researchers, possibilities to consult about the tasks and the process of teaching and peer assessment. She reflected on her role, which she found even more important and difficult than she foresaw, and on the difficulties students had with providing peer assessment. She appreciated that the trials helped her students find out how to work during inquiry and to recognize that assessing others' work is not an easy task and that they need to 'know' (criteria of assessment and solution of the task) before being able to assess. She thought that some children prefer the feedback from peers to her feedback, especially because they can express their thoughts in similar language. Written peer assessment still seems too difficult; the students were able to be more precise and detailed when using oral assessment. It would be good to start with simpler tasks and with some task for training, e.g. working on a series of similar tasks and only at the end ask the students for peer assessment. She also mentioned that the students would prefer to have an opportunity to see and discuss more solutions before assessing. She experienced that time was a large issue as she realized that students may need more time to think over assessment; she maybe gave too much feedback herself and that some of it the students probably would communicate on their own when having enough time.

Overall, Teacher F dealt much more with the process of formative assessment, its relation to learning and summative assessment. She was concerned with her own role in teaching students to formatively assess their peers. She acknowledged the role of students' language and problems that the students may have in writing formative assessment with enough precision. She also considered the time issue, not only from the organization of peer-assessment perspective but also from the cognitive one (time to think). As a whole, she was concerned about classroom internal factors that affect the implementation of formative peer feedback.

Czech Teachers' Concerns

To analyse the results of the entire group of teachers, Czech researchers searched for thematic code with higher occurrence. Attention was paid especially to those codes which mostly diminished after the teachers accomplished the teaching units with formative assessment in their classes, to codes that newly emerged and to codes that were more frequently mentioned (including during the post-lesson interviews).

The concerns that emerged in interviews before the implementation of formative assessment but were much less mentioned in the interviews after the experimental teaching were:

- The relation of formative assessment to summative assessment (mostly feelings that formative assessment is needed and necessary, but summative assessment is seen as preferred, inevitable and expected by administration and parents). It seemed that teachers got this point during their experimental teaching.
- The role of parents and the relationship of parents (feelings that parents expect marks and see peer assessment rather as an entertaining or distractive activity than an essential part of learning). This stayed important for Teacher A; other teachers turned their attention to other issues after the experimental teaching.

The newly emerged concerns or issues which were more intense after gaining the experience of formative assessment implementation were:

- Time demands (the combination of inquiry task and peer assessment seemed to be highly demanding with respect to time and allocation of students' attention; teachers recognized the importance of appropriate complexity of the task)
- Importance of peer assessment for students' learning (both in terms of deepened understanding of assessors and peer-language explanation for assesses)
- The difficulties that students had and possibilities to overcome them

How Czech Teachers' Perspectives on Formative Assessment Developed Over Time

The Struktur Lege Technik helped to delineate the key issues that the teachers considered about formative assessment before and after the trials. It revealed their worries about their own capability to facilitate student peer assessment and expectations of prospective positive outcomes and showed how they perceive formative assessment in the structure of their teaching goals. Teachers gained experience with the method of peer assessment, and they evaluated its benefits and limits. They would recommend developing simpler tasks of shorter duration and simpler and more structured tools for providing peer feedback. The research also revealed that it is very important to develop more deeply an understanding of teaching goals, and especially the role of formative assessment within them. The results indicate that teachers need time to develop a deeper understanding of the role of formative peer assessment in inquiry-based instructions and that there are big differences in the ways in which they conceptualize formative assessment. The subjective theory of formative assessment exploration is a good tool both for investigations of teachers' perspectives on the implementation of formative assessment and for teachers' preparation for facilitating student practices.

Third Perspective: Teacher's Experiences About Using an Internet-Based Application to Facilitate Formative Assessment

Another perspective of teacher's experiences with formative assessment in the context of IBE was made possible through an inquiry platform which was developed to support four assessment methods (peer-to-peer, teachers' written, on-the-fly and structured assessment dialogue). The platform was designed to scaffold a version of inquiry-based education (IBE) with support for formative assessment. It was more appropriately purposed to support formative assessment through grading/feedback and peer assessment rather than on-the-fly and structured dialogue. Key features were aligned with Marshall et al. (2009) 4E x 2 model of inquiry-based instruction (see Figs. 9.2 and 9.3):

- Idea creation and sharing between peers and between student, teacher, and experts
- Formation of inquiry approaches—either structured or open ended
- Mechanisms (e.g. messages, chat boxes, mind maps) to capture ideas and/or metacognitive processes of individual students or groups
- Mechanisms (e.g. audio, video, photographs, spreadsheets, graphs) to allow students/groups to collect a wide range of evidence (data) types
- Mechanisms to allow the organization, manipulation and analysis of different forms of evidence (data)
- Mechanisms to reflect on the reliability and validity of evidence (data)
- An environment to capture the developmental stages of student metacognition
- Mechanisms for students to track their confidence ratings on their inquiry skills over time
- Messaging, chat, discussion, help and feedback forums that could be used in peer or teacher/learner interactions or cached and explored for research purposes
- The facility to engage in discussion and participation of any inquiry or formative approach in- or outside of a formal classroom situation
- The potential for formative and or summative assessment
- A light web-based design, avoiding downloading and capacity issues for laptops or mobile devices

How the Platform Could Support the 4E × 2 Model of Inquiry-Based Instruction

Rather than offering many of the closed inquiry affordances, such as Operation Aries (Koenig et al. 2010) or SimScientists, (Pellegrino and Quellmalz 2010), the platform was developed to be as adaptable as possible depending on the experience, expertise and interests of the end users (teachers and students). It allowed teachers

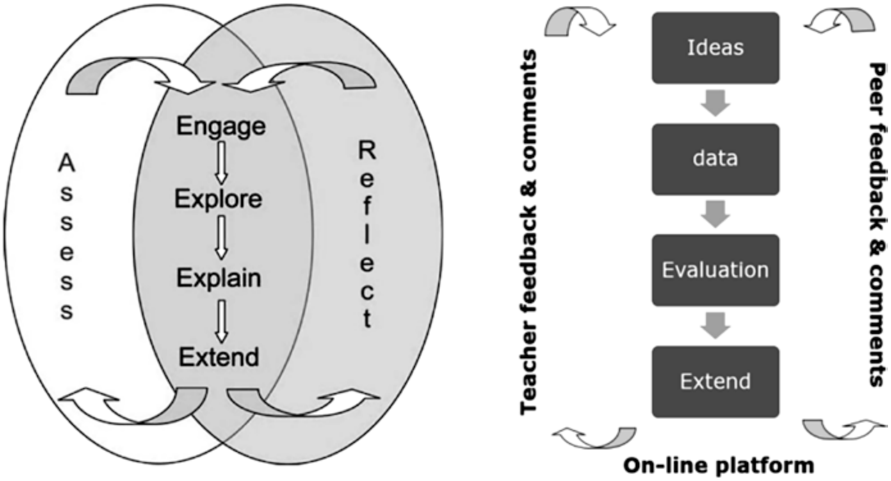
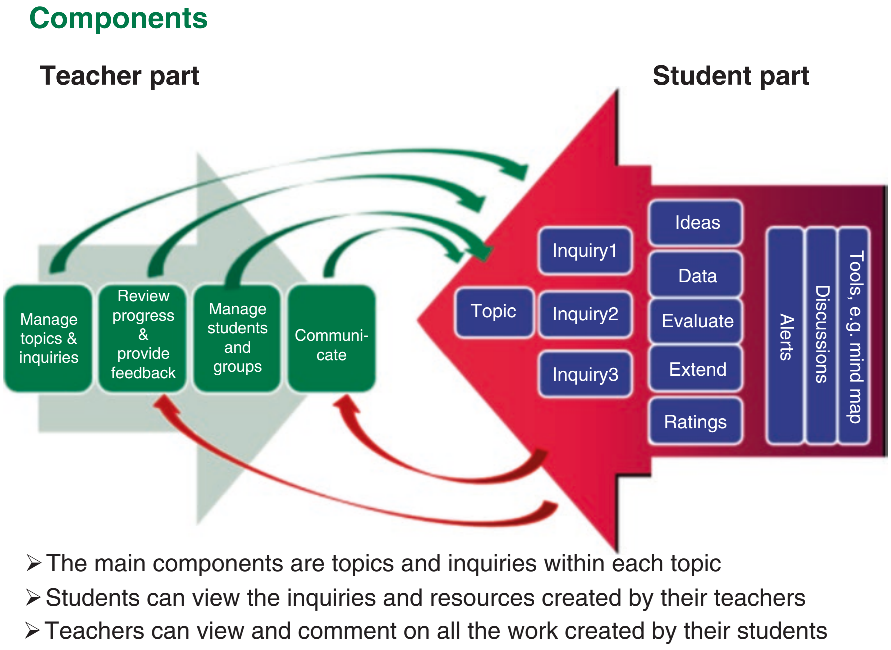


Fig. 9.2 A comparison of elements of IBE (Marshall et al. 2009) with the online platform


Components

Teacher part

Student part

- The main components are topics and inquiries within each topic
- Students can view the inquiries and resources created by their teachers
- Teachers can view and comment on all the work created by their students

Fig. 9.3 Screenshot of the features of the online platform

to select very open-ended screens (and inquiry topics) for particular classes, more structured screens (and topics) for others and the facility to customize the language and approach of any given area of inquiry. The platform could be used to build up particular inquiry skills or metacognitive approaches and then build up to more holistic end-to-end inquiries. It is important to note that although the platform could allow students' freedom and creativity for inquiry, the teacher remained in control of the learning environment and the areas of inquiry they wanted students to explore. Students could work on individual projects or be assigned to groups. They could also be working on a few separate inquiries at the same time—everything being collected in a named inquiry area.

From the teachers' points of view, Fig. 9.4 summarizes the pedagogical and operational features of the platform.

From the students' points of view, Fig. 9.5 summarizes the features that are built into the platform for students' use and illustrates formative benefits of the platform.

It was possible for the platform to be used only for the development of ideas and the collection and collation of various forms of evidence (data) in order to establish an initial area for inquiry. Therefore, inquiry work did not have to work in a structured linear path; the teacher could choose to direct the path. It was also offered in an adapted primary and secondary school design and language. These adaptations also included appropriately simple or complex user mechanisms.

Everything that the student or groups did on the platform could be evidenced and shared with other students, teachers and researchers. All available evidence could be downloaded, explored or presented to others. The intention was for the platform to be as flexible as possible to support the creative ways in which inquiry-based and formative assessment methods could be utilized by teachers and students. This also included the possibility that the platform could be used alongside any non-technological approach and therefore could be considered an enabling or blended learning tool to support varied ways of implementing the assessment methods.

Affordances and Challenges of Using the Platform

The affordances that the platform provided were designed to be as adaptable and flexible as possible in terms of providing an array of conversational, planning, data, video capturing and analytical mechanisms and tools to explore IBE and provide for formative assessment. The online environment provided students and teachers with an integrated environment to facilitate the creation of ideas, plan, execute and evaluate investigations but also collaborate and share their work through peer and teacher assessment.

The development of the platform was not primarily focused on the collection of inquiry-based evidence for the purposes of summative assessment. Rather, it was designed to support the progression of formative e-assessment in broader assess-

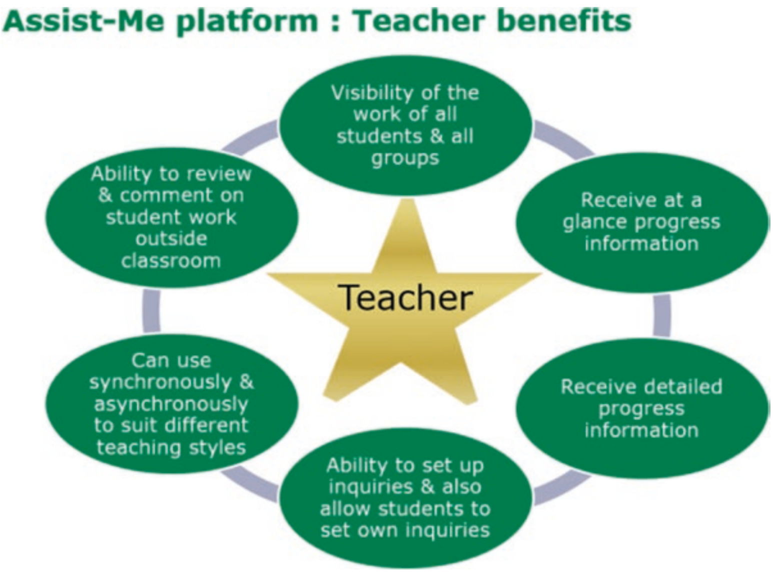

Assist-Me platform : Teacher benefits
Visibility of the work of all students & all groups
Receive at a glance progress information
Receive detailed progress information
Ability to set up inquiries & also allow students to set own inquiries
Can use synchronously & asynchronously to suit different teaching styles
Ability to review & comment on student work outside classroom
Teacher

Fig. 9.4 Teacher features of the platform

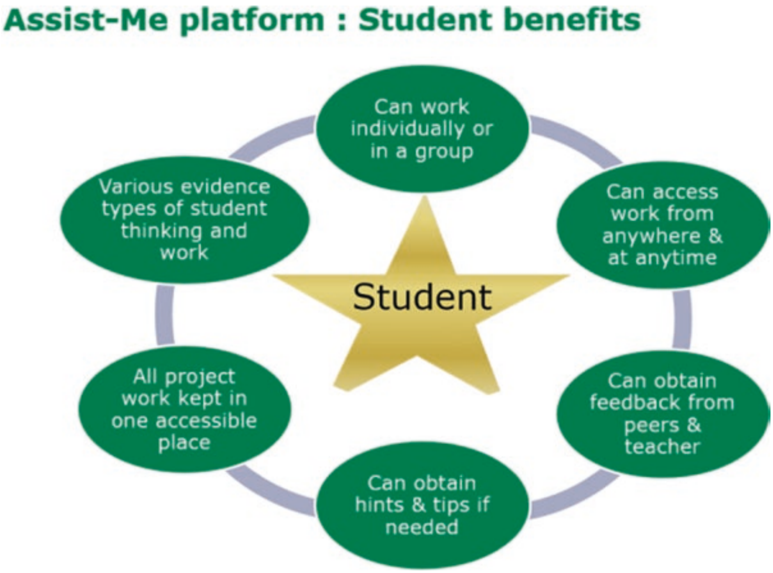

Assist-Me platform : Student benefits
Can work individually or in a group
Can access work from anywhere & at anytime
Can obtain feedback from peers & teacher
Can obtain hints & tips if needed
All project work kept in one accessible place
Various evidence types of student thinking and work
Student

Fig. 9.5 Student features of the platform

ment contexts and in particular personalizing the use of such technology to drive opportunities for assessment for learning (Ripley 2006). Integrating technology of this type into teaching and learning and making use of the mechanisms available in the platform enabled, for example, the creation of online e-portfolios that can support both formative and self-regulated learning activities (Crisp 2007; Coombs 2010). In addition, the inclusion of dialogue and conversations provides rich evidence of reflective thinking and can capture qualitative evidence as part of self-assessment learning tasks, an agenda promoted by Futurelab (2007).

The platform was trialled over a number of phases, with small amendments made between phases based on requests and feedback by the participating countries. Teacher groups were provided with face-to-face and online training sessions in order to familiarize themselves with the platform and the ways in which it could be used alongside traditional methods to support inquiry and formative assessment methods. Some partner groups were happy to explore and experiment using the platform, some groups lacked either technological or assessment knowledge to effectively engage with the platform and there were some groups who were philosophically opposed to using technology for formative purposes. One clear outcome from the research was that as much as teachers requested sophisticated tools and mechanisms built into the platform from the start, it became clear as the research phases progressed that they much preferred simpler interfaces to gain confidence and platform usage by teachers and students.

In general terms, the platform had mixed levels of interest and usage. These results reflected the difficulties and challenges that teachers had in implementing formative approaches in IBE. Many of the teachers reported low levels of efficacy in terms of implementing traditional forms of formative assessment. This was clearly compounded when an online environment was also made available to them and their students. Many of the features of the online platform, particularly the inclusion of messaging, chat and uploading of differing file types (photo, video, aural), have similarities to those found in social media and ‘apps’. While most students are now familiar and confident working with these forms of technology, it might be assumed that some of their teachers are less comfortable with them. The challenge of familiarizing themselves with sophisticated new technologies, setting up and managing groups while also facilitating formative assessment strategies, proved to be a step too far for most of the countries and teacher groups. There was evidence of trialling the platform, however little of concerted classroom use.

The findings for the use of traditional forms of IBE indicated that there is a concerted need for support and professional development to implement formative assessment into classroom practice. The same issue applies when countries work towards integrating technology into the curriculum and pedagogies. Provision of equipment, software and guidance, however, easy to use, will itself not be enough for most teachers to make use of the affordances that technology can offer. Whether technology takes the form of an integrated platform such as the one developed or just makes use of available mobile technology found in devices such as mobile phones, tablets and their applications, the gap between the resources that could be used and what are currently being used needs to be closed in order to enhance formative assessment opportunities and encourage deep learning.

Summary Perspective

Together, these teacher perspectives reveal participant teacher experiences that result in some increase in beliefs in their abilities to use formative assessment in their teaching as well as clear concerns about the challenges to be expected. Although the lack of clear generalizable trends across cultures may be due to the multiple relevant variables within education in diverse settings, the growth in useful knowledge and confidence in using formative assessment methods among teachers points to realistic perspectives for further work at national levels. The promise of an Internet-based platform to facilitate inquiry teaching and learning as well as formative assessment remains an attractive potential that was not fully tested in this project's classrooms. Possibilities for the introduction and even institutionalization of formative assessment methods into science inquiry classrooms are informed by these teacher perspectives.

References

Bandura, A. (1992). Exercise of personal agency through the self-efficacy mechanism. In R. Schwarzer (Ed.), *Self-efficacy: Thought control of action* (pp. 3–38). Washington, DC: Hemisphere.

Bandura, A. (1995). *Exercise of personal and collective efficacy in changing societies*. New York: Cambridge University Press.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. Englewood Cliffs: Macmillan.

Bernholt, S., Rönnebeck, S., Ropohl, M., Köller, O., & Parchmann, I. (2013). *National reports of partner countries reviewing research on formative and summative assessment in their countries*. Kiel: Leibniz-Institute for Science and Mathematics Education (IPN).

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education)*, *21*(1), 5.

Bleicher, R. E. (2004). Revisiting the STEBI-B: Measuring self-efficacy in preservice elementary teachers. *School Science and Mathematics, 104*(8), 383–391.

Boyatzis, R. E. (1998). *Transforming qualitative information: Thematic analysis and code development*. Thousand Oaks: sage.

Brousseau, G., & Novotná, J. (2012). *Úvod do teorie didaktických situací v matematice*. Pedagogická fakulta: Univerzita Karlova.

Brown, G. T. (2004). Teachers' conceptions of assessment: Implications for policy and professional development. *Assessment in Education: Principles, Policy & Practice, 11*(3), 301–318.

Coombs, S. (2010, May). Critical thinking, portfolio assessment and e-scaffolding of continuing professional development for knowledge elicitation. In *Global Learn* (Vol. 2010, no. 1, pp. 4010–4014).

Crabtree, B., & Miller, W. (1999). A template approach to text analysis: Developing and using codebooks. In B. F. Crabtree & W. L. Miller (Eds.), *Doing qualitative research* (pp. 163–177). Newbury Park: Sage Publications.

Crisp, G. (2007). *The e-assessment handbook*. Continuum.

Czerniak, C. M. (1990). *A study of self-efficacy, anxiety, and science knowledge in preservice elementary teachers*. Atlanta: National Association for Research in Science Teaching.

Eastwell, P. (2009). Letters: Inquiry learning: Elements of confusion and frustration. *The American Biology Teacher, 71*(5), 263–266.

Edmondson, W. J. (1998). Subjective parameters describing teaching roles: Towards a theory of tertiary foreign language instruction. *Fremdsprachen Lehren und Lernen, 27*, 81–105.

Enochs, L. G., & Riggs, I. M. (1990). Further development of an elementary science teaching efficacy belief instrument: A preservice elementary scale. *School Science and Mathematics, 90*(8), 694–706.

Futurelab (2007). *E-assessment- an update on research, policy and practice* (Report 10). Available at http://archive.futurelab.org.uk/resources/documents/lit_reviews/Assessment_Review_update.pdf

Groeben, N., Wahl, D., Schlee, J., & Scheele, B. (1988). *Das Forschungsprogramm subjektive Theorien: eine Einführung in die Psychologie des reflexiven Subjekts* (p. 364). Francke.

Hodgen, J., & Marshall, B. (2005). Assessment for learning in English and mathematics: A comparison. *Curriculum Journal, 16*(2), 153–176.

Hodgson, C., & Pyle, K. (2010). A literature review of Assessment for Learning in science. Slough: Nfer.

Janík, T. (2005). Zkoumání subjektivních teorií pomocí techniky strukturování konceptů (SLT). *Pedagogická revue, 57*(5), 477–496.

Koenig, A. D., Lee, J. J., Iseli, M., & Wainess, R. (2010). A conceptual framework for assessing performance in games and simulations. CRESST Report 771. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. Los Angeles: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST). Lederman, L. C., & Fumitoshi, K. (1995).

Korthagen, F. A., Kessels, J., Koster, B., Lagerwerf, B., & Wubbels, T. (2001). *Linking practice and theory: The pedagogy of realistic teacher education*. New York: Routledge.

Košťálová, H., & Straková, J. (2008). Evaluation: trust, dialogue, growth. SKAV, os.

Marshall, J. C., Horton, B., & Smart, J. (2009). 4E× 2 instructional model: Uniting three learning constructs to improve praxis in science and mathematics classrooms. *Journal of Science Teacher Education, 20*(6), 501–516.

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2), 119–134.

Ripley, M. (2006). *The four changing faces of e-assessment 2006–2016.* Available at: http://www.xplora.org/ww/en/pub/insight/thematic_dossiers/articles/e_assessment/eassessment2.htm

Scheele, B. (Ed.). (1992). Struktur-Lege-Verfahren als Dialog-Konsens-Methodik: Ein Zwischenfazit zur Forschungsentwicklung bei der rekonstruktiven Erhebung Subjektiver Theorien. Aschendorff.

Šeďová, K., Švaříček, R., & Šalamounová, Z. (2012). Komunikace ve školní třídě. Orbis scholae, 147.

Tobin, K., Tippins, D. J., & Gallard, A. J. (1994). Research on instructional strategies for teaching science. *Handbook of research on science teaching and learning, 45*, 93.

Whitehead, J. (1989). Creating a living educational theory from questions of the kind, 'How do I improve my practice?'. *Cambridge Journal of Education, 19*(1), 41–52.

Chapter 10
Policy Aspects: How to Change Practice and in What Direction

Jens Dolin, Jesper Bruun, Costas P. Constantinou, Justin Dillon, Doris Jorde, and Peter Labudde

When it was set up in 2012, the ASSIST-ME project was responding to a perceived need within education in Europe to reform evaluation and assessment practices. Some years earlier, in 2009, the OECD had launched the programme 'Review on Evaluation and Assessment Frameworks for Improving School Outcomes' (www.oecd.org/edu/evaluationpolicy) 'designed to respond to the strong interest in evaluation and assessment issues evident at national and international levels' (ibid.). As stated in the final report, one of the key issues dealt with was: 'How can assessment and evaluation policies work together more effectively to improve student outcomes in primary and secondary schools?' (OECD 2013). The reports expressed a growing understanding among the OECD countries that evaluation and assessment systems are key components in improving school systems. Together with a number of other reports aimed at education policy-makers, the need for a greater focus on formative assessment and its integration with summative assessment became more and more obvious (Looney 2011; OECD/CERI 2005). A report 'emphasise[d] the importance of seeing evaluation and assessment not as ends in themselves, but instead as important tools for achieving improved student outcomes' (OECD 2011, p.1). This position suggests an emerging awareness among policy-makers that evaluation for accountability reasons has, to a large degree, hindered formative processes.

J. Dolin (✉) • J. Bruun
Department of Science Education, University of Copenhagen, Copenhagen, Denmark
e-mail: dolin@ind.ku.dk

C.P. Constantinou
Department of Educational Sciences, University of Cyprus, Latsia, Cyprus

J. Dillon
University of Bristol, Bristol, UK

D. Jorde
ILS, University of Oslo, Oslo, Norway

P. Labudde
Fachhochschule Nordwestschweiz, Pädagogische Hochschule, Basel, Switzerland

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science
Education Research 4, DOI 10.1007/978-3-319-63248-3_10

The importance of assessment *for* learning has, at the same time, become widely known in education circles, thanks, to some degree, to John Hattie's meta-analysis *Visible Learning* (Hattie 2008), which showed that formative assessment is one of the most effective ways to enhance student learning.

Especially within science education, both researchers and teachers have for a long time been complaining over the many tests often testing only rote learning and simple knowledge (Harlen 2007), and reports have pointed at solutions:

> Tests are dominated by questions that require recall – a relatively undemanding cognitive task and, in addition, often have limited validity and reliability. … Transforming this situation requires the development of assessment items that are more challenging; cover a wider range of skills and competencies; and make use of a greater variety of approaches – in particular, diagnostic and formative assessment. (Osborne and Dillon 2008, p. 8)

This growing imbalance between a dominating test system assessing relatively simple skills and the need for introducing formative assessments able to capture more advanced competences has been a focus for EU projects such as SAILS (http://sails-project.eu/). The ASSIST-ME project responded to a call from the European Union's research and innovation funding programme for 2007–2013, the so-called FP7 programme. The call recognised that merely developing new assessment items was not sufficient; it also had to be a political priority to reform the educational system to give room for formative processes in a system with strong emphasis on summative assessments. The FP7 call contained an explicit demand to enter into the political world, and some of the expected outcomes were identified:

> The research should be 'use-inspired' and lead to identification of the factors (including cultural) that undermine the effective uptake of formative assessment appropriately combined with summative assessment in different contexts […] The actions should include policy recommendations and appropriate dissemination activities […] the project will provide policy makers with data and guidelines for an informed decision making.

The ASSIST-ME project has taken up these challenges. The project partners defined dual aims: (1) to provide a research base on the effective uptake of formative and summative assessment for inquiry-based, competence-oriented Science, Technology and Mathematics education in primary and secondary education in different educational contexts in Europe and (2) to use this research base to give policy-makers and other stakeholders guidelines for ensuring that assessment enhances learning in STM education (Dolin 2013, p. 5).

This chapter describes how the second aim, the policy-oriented aspects, was dealt with in ASSIST-ME. It will put the project experiences into a wider European perspective by referring to other initiatives. A key focus is on how educational policies may encourage or restrict the contribution that assessment might make to students' learning. In many countries, the formative-summative dichotomy has, to a large degree, been created by the national use of externally based summative assessments. We know that such tests often distort the learning agenda while being inadequate in validity and even often in reliability (see Chap. 3 in this volume). Various examples will be given from different national contexts of the barriers to introducing more formative assessment, and also examples of national policies given more emphasis to the formative aspects, and maybe combining them with summative assessments.

In accordance with the dual goals of the ASSIST-ME project, it was organised as two parallel strands. In one strand ASSIST-ME teachers and researchers worked together in teacher action research processes designing and implementing formative assessment processes. This work was organised in local working groups (LWGs) which met regularly to prepare and discuss the implementation. Researchers analysed data to answer the research questions. This work is presented in Chapters. 4, 5, 6, 7 and 8 in this volume. The other strand dealt with the policy aspects of the project. In each participating country, National Stakeholder Panels (NSPs) were established involving representatives from different stakeholder groups influencing educational decisions in the country. In this chapter, we will describe how the project used social network analysis (Scott and Carrington 2011) for identifying and selecting NSP members as well as the work and outcomes of the national NSPs. We will also show how the NSPs have given feedback to, and informed, the ASSIST-ME project. In a wider perspective, we will analyse how research results have and can influence STM education, both the educational practices and the political climate and decisions framing education. At this point, we will go beyond ASSIST-ME and draw upon other project experiences across Europe. Finally, the policy recommendations for the transformation process based on the ASSIST-ME experiences will be put forward.

Linking Teachers, Researchers and Policy-Makers

Despite the increased focus in the last decade on using research as an evidence base for educational decisions, it is not the norm for educational research projects to explicitly address policy issues. Most researchers see it as a virtue to separate research, seen as objective and independent, from policy, which is interwoven with interests and weakly underpinned opinions. Of course, the reality is far more complicated. Most research is financed by foundations, companies, interest groups or a public programme – with specific aims and specific expectations in terms of the product – and researchers are therefore becoming more aware of the impact of their findings and of the necessity of engaging in dialogue with policy.

Research tells us, however, that it is not easy for research to affect policy due to different logics and different discourses (Fensham 2009). In his article ‘Speaking Truth to Power with Powerful Results: Impacting Public Awareness and Public Policy’, Mack Shelley II (2009) underlines the need for eclecticism in research and its interface with expertise and policy. The point is that to communicate research findings to decision-makers, you need to break down the barriers between the research world and the policy world through better communication and an understandable and usable message. This can only be achieved if the two sides meet to exchange ideas and understandings and accept each other’s respective capacities and influence.

Education researchers primarily report their results to the research community which is most often isolated from the political processes, not because researchers have little knowledge about their national policy matters but to ensure their academic independence. This dissociation between research and policy goes back to the building of the independent universities and their insistence on academic freedom.

One of the consequences of this approach has been the development of different logics and discourses within these separated areas of activity – with the result that educational research sometimes has little influence on policy. To change this situation, it is necessary to understand how policy is made and how it is implemented. It is also important to know the discourse of policy and to communicate results in a way politicians can understand and use. If you want influence on educational policy, you need to engage directly with the relevant policy-makers in a language they can understand – it is of no use to stand outside talking to each other in your own jargon.

In broad terms, both educational research and education policy development aim to contribute towards improved teacher effectiveness as well as enhanced student learning, creativity and emergent autonomy. Educational research is constrained by methodology and the reliability of evidence. Policy-making is constrained, additionally, by convention, existing practices, finite resources, administrative structures, professional capacities and stakeholder interests.

Improving student learning in STM is an immensely challenging goal. As they grow, children develop their own perspectives, styles and interests. The extent to which they purposefully engage with science learning has a crucial influence on their progress. Meeting the aim of attaining a minimum level of science literacy for all requires concerted effort on behalf of teachers and families over extensive periods of time. Many things can and often do go wrong. Systemic constraints add an extra level of complexity to this inherently challenging task.

The Science Education Expert Group brought together by the European Commission sought to critically analyse prior efforts to promote reform in STEM education and to utilise diverse perspectives in using existing evidence to formulate recommendations and priorities for renewed efforts. This dual emphasis on diversity of participation and evidence-informed policy development is valuable in providing a bridge between science education research and an informed approach to formulating future priorities.

European policy-making has a natural ally in international research activities and communities such as European Science Education Research Association that promote these activities. Collaboration across different educational systems adds value and thoughtfulness to reform efforts, especially if that collaboration takes the available research into account. In addition, it has the potential to promote learning from best practices. Even so, the divergent priorities of research and policy reform sustain widely differing cultures in the communities that promote them sometimes with deep mistrust between them. In this sense, it is not surprising that the gap between research and practice appears to be widening.

The subsidiarity principle places responsibility for European education at the level of Member States. This tends to severely limit the power of European policy-making to influence classroom practices. At local level, the divergence between science education research and policy tends to be even greater. At this level, administrative constraints, conventions and stakeholder interests tend to be more pronounced with a correspondingly stronger influence in the majority of educational systems.

Where does all this leave us?

The need for evidence-informed policy development is stronger than ever. To respond to this, there needs to be a concerted effort to strengthen capacities and promote structures that enhance bilateral communication between policy-makers and educational researchers. Formulating measurable objectives and committing to them over extensive periods of different politicians is equally important for this shift to happen.

The same situation of discursive misunderstanding and deeply rooted mistrust also applies to the relationship between teachers and policy-makers and to a lesser extent to the relationship between researchers and teachers. One model for these relationships is a triangle with research, policy and practice at the corners (see Fig. 10.1).

In the model, researchers and teachers work together to define and carry out research that both parties find useful. The work is mediated by policy input. Researchers and policy-makers exchange research findings and policy demands for new research, with teachers being involved in the communication process. Teachers and policy-makers build a mutual respect for each other's profession and recognise each other's distinct roles, while researchers can deliver knowledge to the process. For each of the agents, it is a question of changing their respective areas of practice with due respect to the legitimate interest and expertise of the other two agents. Researchers need to pursue research questions relevant to teachers and policy-makers. Teachers need to change teaching practices in the light of research and policy demands based on societal decisions. Policy-makers need to change educational policy in accordance with research and with teacher requirements for fulfilling their job.

The ASSIST-ME project has mainly worked on the research-teacher and the research-policy relations. The first relation was practised in the LWGs, and the links from teachers to policy were managed by the researchers, although teachers participated in some NSP meetings. But the fact that teachers' voices were reported to

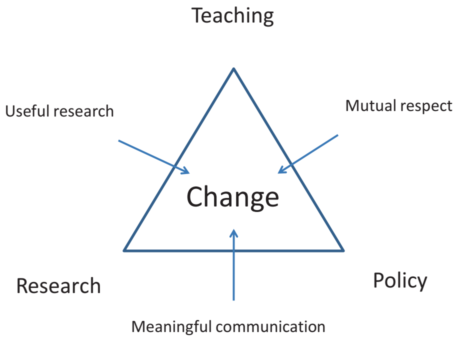


Fig. 10.1 A model of the relationships between research, policy and practice

policy-makers by researchers did put the onus on the researchers to get it right. The research-policy relation was practised in the NSPs, and in this chapter, we will mainly deal with the work of the NSPs and refer to some of the policy-related discussions in the LWGs.

The ASSIST-ME Approach to Policy Issues

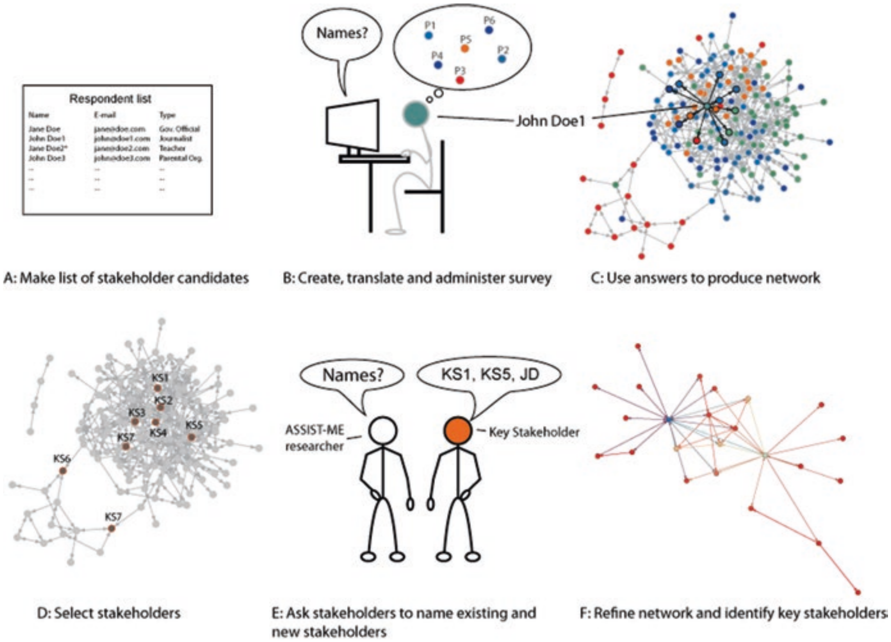
The NSPs played a pivotal role in including policy-makers in the project and in giving them influence and co-ownership of the research findings. The starting point was an acknowledgement that policy-makers were looking for solutions to problems they realise were important. We also knew from our experience that there were no quick fixes; one piece of evidence wouldn't do the trick – we needed sustained lobbying, which required consistent involvement from the selected representatives from the different stakeholders. We believed that most stakeholders want to influence policy not for their own self-interest but for what they saw as higher motives, often coloured by their organisational background. The challenge was how do you systematically identify the people who represent key stakeholders, how do you make them interested in getting involved and how do you sustain their interest throughout the project's lifetime?

Selecting Key Stakeholders Through Network Analysis

The identification of key people influencing STM education was managed through a social network analysis process. The method is described in detail in ASSIST-ME Deliverable D5.1, downloadable from www.assistme.ku.dk, and Fig. 10.2 illustrates the steps in the method. The first list of Danish stakeholder candidates was drawn up by selected national researchers. They were told that the stakeholder network should map out the people in different organisations – not the organisations themselves – who were influential in making educational changes.

Based on this first list, we identified the following groups of key stakeholders in the project who it was crucial to involve in any project aimed at influencing STM education:

- *Government and municipalities (central as well as local levels)*: Governments drive policy development in the area of innovative science strategies at a general level, such as the role of assessment, the introduction of new assessment methods, the overall curriculum goals, the framing conditions for teacher education and in-service training, etc. Depending on the educational system and culture, municipalities and local authorities can be crucial in facilitating cultural change in STM education.


Respondent list
Name E-mail Type
Jane Doe jane@doe.com Gov. Official
John Doe1 john@doe1.com Journalist
Jane Doe2* jane@doe2.com Teacher
John Doe3 john@doe3.com Parental Org.
Names?
P1
P6
P5
P2
P4
P3
John Doe1
KS1
KS2
KS3
KS4
KS5
KS7
KS6
KS7
Names?
KS1, KS5, JD
ASSIST-ME researcher
Key Stakeholder

A: Make list of stakeholder candidates

B: Create, translate and administer survey

C: Use answers to produce network

D: Select stakeholders

E: Ask stakeholders to name existing and new stakeholders

F: Refine network and identify key stakeholders

Fig. 10.2 A graphical illustration of the steps in identifying and selecting key stakeholders

- *Media*: Journalists in the leading print media and broadcasting corporations have substantial influence on policy and decision-makers, especially with respect to agenda setting. Close communication with media representatives is crucial in order to facilitate the public discussions of how assessment strategies influence the outcomes of teaching. This approach will support policy-makers in working with cultural change in STM education.
- *Business and industry*: With the key role STM education plays in the economic development of all EU countries, organisations representing business interests have a strong incentive to be involved in the project. In some countries, such as Denmark, private foundations, often with strong connections to private companies, also have a major influence in deciding which areas and development trends they give funding for development and research.
- *Teachers*: Teachers are primary agents for implementing real changes in the classroom. It is vitally important that they are directly or indirectly involved in any education research projects. This involvement may be via those teachers well known for their contributions to public discourse, or, more importantly, those who are involved in teacher organisations. In Denmark, teachers are organised in two teacher unions, one for compulsory school teachers (primary and lower secondary level) and one for upper secondary school levels. Both have significant influence on educational policy. As well as being members of unions, most teachers are also members of a subject teacher association, which are involved in curriculum changes.

- *School leaders and school owners*: School leaders are usually responsible for organising and supporting implementation of most educational changes. Their involvement is essential in any project leading to policy recommendations. Depending on national conditions, school owners are equally important. In Denmark, compulsory schools are mostly 'owned' (meaning financed and steered on major issues) by the municipalities, while some compulsory schools are private, and nearly all upper secondary schools (gymnasiums) are independent units under national legislation.
- *Teacher trainers and professional development providers*: Teacher trainers and professional development providers should be conversant and fully aware of the project objectives and findings. They will then be able to adopt and adapt those aspects that will encourage more widespread uptake of the project ideas and findings.
- *Research communities of STM education*: The results of the project have been disseminated in conference presentations and by publishing in journal papers. Feedback from the research community is crucial in validating the project's findings in a wider context.

These groupings were communicated to the ASSIST-ME partners, and they guided the partners in their selection of NSP members so as to have as broad a representation as possible.

In each country, an iterative process was carried through. The ideal process is illustrated in Fig. 10.3, and due to unforeseen obstacles in the partner countries, the process varied a bit from country to country. The ideal process has these phases:

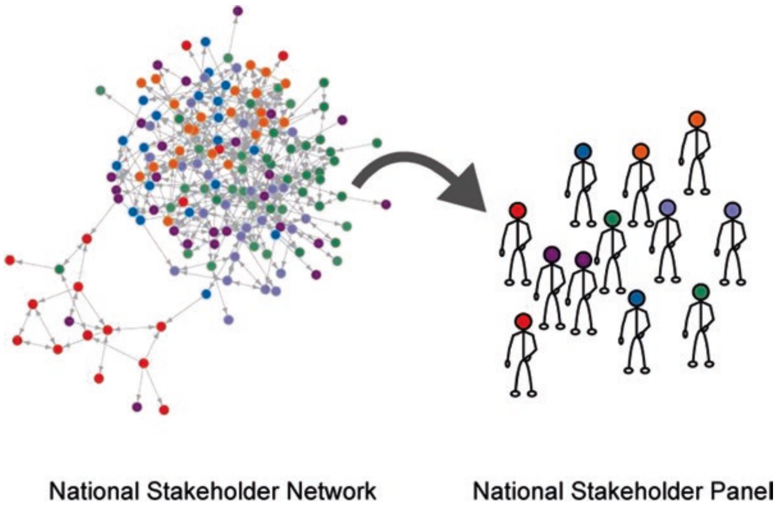


Fig. 10.3 The purpose of the social network analysis method is to create stakeholder networks that can be used to pick out relevant persons for the National Stakeholder Panels

A. The researchers made up a list of stakeholders covering the above groupings, including the name, email address and to which organisation the stakeholder belonged. In a country like Denmark, we expected the list to be comprised of approximately 500 persons.
B. Each person on the list was sent a SurveyXact questionnaire (in the national language) asking: 'Which persons do you consider to have impact on changing assessment practices at the school, local and national level?' For each person they named, they gave an email address (if possible) and organisation affiliation. They were also asked to indicate whether the person had impact on changing assessment practices at the school level, local level or national level.
C. Based on the survey data, the Danish group produced a network of stakeholder candidates and gave it a first analysis in collaboration with the national researchers.
D. Stakeholders who held central positions in the network of stakeholder candidates were listed, i.e. persons with many incoming connections but also with a large diversity in types of stakeholder candidates to which they were connected. The selection was based on a dialogue between the ASSIST-ME partners and the Danish team. If the ASSIST-ME partner believed that an important stakeholder candidate was missing from the network, the stakeholder could be included in the new list of stakeholders. The number of stakeholders was expected to be between 50 and 100 persons.
E. Each stakeholder was presented the list of stakeholders (i.e. the list mentioned in 'D') and was asked to choose the persons they believed were important. They could select between the names on the list, and they could also supply names they believed were missing.
F. Based on the stakeholder connections found in step E, the Danish team refined the network of stakeholders to those individuals who were identified as having impact on changes in assessment practices. This final network was analysed in detail by the Danish team in collaboration with each ASSIST-ME partner to select key stakeholders relevant for inviting to the National Stakeholder Panels.

The next phase was to recruit the members from the refined list of stakeholders (determined in F). This was very much a question of finding the right balance between choosing the most influential individuals who, at the same time, would find the membership of a NSP important enough to give priority to attend the meetings. Some potential members were contacted personally. The final list of selected stakeholders was mailed to all on the list, so they could see the optimal composition of the NSP.

The resulting composition of the NSPs varied from country to country, depending on country size, the status of the researchers and the national policy culture (Bruun et al. 2015). All the researcher teams had high-profile researchers. In small countries like Denmark, it was easier to get access to relevant key persons, whereas in larger countries or in countries with a decentralised administration, it seemed more difficult to find policy-makers willing to represent a sector. In large, centralised countries, it was possible to pinpoint key persons and to have them accept membership in the NSP.

Once this NSP network was established, it was essential to deliver concrete and relevant input to them in order to keep the enthusiasm and interest alive. If not, members only attended meetings irregularly or they substituted their own attendance a less influential person from the organisation.

How Did the NSPs Work? Agendas, Successes and Problems

At the beginning of the project, the Danish partner published National Stakeholder Panel Guidelines for the NSP meetings beginning in month 12 of ASSIST-ME to provide goals and suggestions for the first meeting. The guidelines were one of the planned outputs ('deliverables') of the project. The NSPs were expected to have face-to-face work meetings three times during the project:

- Early December 2013 (Month 12)
- March 2015 (Month 27)
- January 2016 (Month 36)

The Danish group sent out draft agendas before every NSP meeting. The agenda reflected the current project needs for members' input and guidance and asked for their reflections on the problems currently dealt with that could contribute to answering our research questions. Each partner could supplement the common agenda with local issues, and the partner was responsible for taking minutes of the meeting and sending them (in English) to the Danish group, which was responsible for this part of the project, and uploading them on the common communication platform. The outcome of the NSP meetings will be presented in the next section in a thematic form.

First NSP Meeting (Month 12)

The first meeting settled the role of the NSP. It outlined the project research question and the problems that the researchers wanted to address. Members were asked to consider how the project could be managed for co-ownership. The following questions were asked:

How could you secure a meaningful communication between researchers, teachers and stakeholders, including policy-makers?

How could relevant stakeholders be invited to take co-ownership of the research results and how could a partnership between researchers, policy-makers and teachers be established in order to secure relevant actions following implementation guidelines?

Second NSP Meeting (Month 27)

The second meeting addressed the assessment issues at the focus of ASSIST-ME: It posed the following questions:

1. Do you see any reason to change the assessment/examination culture in your country?
2. If not – why not? If yes – why, and in which way?
3. What will be the best strategy for changing the assessments in a direction that takes the ASSIST-ME results into consideration?
4. In which ways can you – as a NSP – help the changing process?
5. If we should apply for a successor, a follow-up for ASSIST-ME, which research questions should we then pursue?

Third NSP Meeting (Month 36)

The third and last meeting was held after the final implementation round. In this meeting, the common questions, arising from the previous meetings, were discussed in the light of the preliminary findings. The questions were discussed at an ASSIST-ME Management Board meeting, and the following list was agreed upon:

1. What position/role describes you best?
2. From your perspective, describe how students' learning is assessed in your country. Please describe both formative assessment for learning (e.g. teachers' feedback to students in the daily teaching) and summative assessment of learning (e.g. exams). Please indicate if these practices differ across educational levels from grade 1 to 12 (baccalaureate).
3. Is learning about formative and summative assessment an important aspect of teacher education and teacher professional development TPD?
4. Is it desirable to try to combine formative and summative assessment?
5. Are there any nationwide (or regional-wide) high-stakes assessments in your country?

   If yes: At which level(s)?
6. Do you see any reason to change the assessment/examination culture in your country?
7. What changes, if any, do you find necessary in the examinations at different levels to make it reflect the competence goals (both subject specific and generic) in the curriculum?
8. Do you have any influence on the change of the assessment system in your country?
9. If so, will you use your influence in any change process – and in what direction? How can you best change the assessments/examinations in the desired direction?

NSP Discussions

The following is a summary of some of the discussions that took place in the national NSPs. Common agreements are discussed first, and then specific country points of view are provided.

Learning About Formative and Summative Assessment in Teacher Education and Teacher Professional Development (TPD)

All panels agreed that assessment is an important aspect of teacher education and TPD. There was substantial agreement that the weight assigned to formative assessment in teacher education in TPD programmes is contingent on conceptualising the assessment of inquiry and process competences. The panels identified a range of needs that informed the project:

- Instruments, tools, guidelines and examples of good assessment practice. However, it is not sufficient to provide teachers with diagnostic instruments – they also have to understand the underlying principles of these instruments.
- Teachers need to be convinced that they can handle competence-oriented formative assessment.
- Teachers need clear competence descriptions that can be used as a basis for formative assessment.
- Regarding in-service teachers, there is a need for teaching innovation projects that integrate teaching institutions (e.g. schools) and education research groups.

It was also stressed that it is not sufficient to provide teachers with materials and discuss these in short (e.g. one day) TPD activities. The implementation has to be accompanied in practice by long-term TPD. This is in accordance with the vast amount of TPD research that indicates that one-shot workshops do not work (Goldenberg and Gallimore 1991; Lieberman and Pointer Mace 2008; Scott 2010).

The Desirability of Combining Formative and Summative Assessment

The NSPs reported a number of ways in which summative and formative assessment are combined in practice, for example, students' work on projects is assessed formatively during lessons and is assessed summatively at the end of the course (the summative assessment is then provided as oral or written feedback). Alternatively, several summative tests can be used during a teaching unit/course and serve formative functions. However, it was generally agreed that consistency is important and there needs to be an alignment between teaching and assessment, whether it is for formative or summative purposes, there needs to be the same visible criteria. In terms of alignment of formative and summative assessment, it was mentioned in one of the panels 'that there is lack of systematic implementation of the two types of assessment. Hence, combining the two types becomes an ever more difficult task'.

Throughout the discussion, it emerged that at some points, and for some purposes, assessment can (and should) only be formative. It is important to consider that formative and summative assessment serve different functions. In one of the panels, it was argued that the only way to combine formative and summative assessment is by evaluating student portfolios to monitor students' learning progress.

Do You See Any Reason to Change the Assessment/Examination Culture in Your Country?

There was a strong feeling among all groups that changes in assessment culture should be adapted to the relevant contexts – not just in terms of national or cultural context but also educational level, subject, etc. All countries highlighted areas needing some kind of attention or change, but it was also stressed that assessment is a sensitive topic and, as such, quick fixes should be avoided.

In this section, we will look at points that emerged from the NSPs which are not common across countries and which reflect particular cultural or systemic issues.

There was a discussion among some panels about how the focus on formative assessment should not detract from the role of summative assessment, especially since parents are used to, and expect summative assessment:

- 'We are not in the stage where the formative assessment could be part of everyday teaching. There is still prevailing demand for grading – the teachers need them to make final certificate and parents are used to work with them too'. (Czech minutes)
- 'Also, regarding marking, there's a long history in France concerning this tradition of marking that is not easy to change. One participant mentioned that the society is quite competitive and we should educate students about it as well'. (French minutes)
- 'The panel speaks clearly in favour of a strict separation of formative and summative assessment to avoid a confusion of learning and achievement situations. In addition, it stresses that summative assessment cannot (and should not) be completely abolished'. (German minutes)
- 'Also, many parents still advocate summative testing (grades)'. (Finnish minutes)
- 'Stronger involvement of parents: the parents want formative assessment on the one hand, but at the other hand also want to know the "worth of any artefact" (summative assessment)'. (Swiss minutes)

There was a general consensus across the partner nations that something needs to be changed in the assessment culture to enhance the status of formative assessment, albeit there were different opinions about what should be changed and how this could be done:

- 'The reason for the change: The school assessment doesn't support quality of students' learning with respect to understanding of content. The assessment should help student to "learn with understanding" and achieve better understanding of the content'. (Czech minutes)
- 'Even if teachers often will promote formative assessment, the overall discourse in the school system is on summative assessment. The focus on figures in the school has exploded and students have high attention on their marks. More formative assessment could be a useful reflection tool for students'. (Danish minutes)
- 'In general the NSP indicated that, on one hand, there's a need to engage teachers in an attitude that foster more assessments for learning. On the other hand, they indicated that the (official) educational system position is quite heterogenic regarding assessment'. (French minutes)
- 'The panel feels that there had been an assessment culture at schools once but it has gotten lost to a huge extent. If it would be possible to reimplement it, this would have positive influence on school development, teaching and learning. In this context, the school leaders are crucial'. (German minutes)
- 'Large-scale assessments with innovative assessment formats could initiate more innovative teaching in the classrooms (positive teaching to the tests). So far, the existing regional-wide assessments are rather traditional. So one could also fear that more such tests kill innovative and creative teaching'. (Swiss minutes)
- '… the majority of existing diagnostic tests are devoted to measuring students' content knowledge, without placing any emphasis to on their attitudes and or skills. In addition, it would be more productive to base students' assessment on a wide variety of tools and methods, such as portfolio, individual and cooperative work'. (Cypriot minutes)
- 'More emphasis should be put on feedback and focus more on learning (what is learnt instead of what is not learnt). Generally, people should be more aware about the diversity of assessment methods and practices'. (Finnish minutes)

What Changes, If Any, Do You Find Necessary in the Examinations at Different Levels to Make Them Reflect the Competence Goals (Both Subject Specific and Generic) in the Curriculum?

The role and the validity of the examinations were central issues for most NSPs, and across countries, there was an awareness that changes were needed:

- ‘The panel admits that changes in the final examinations do have a steering influence. Changes in final examinations thus always have to be preceded by changes in instruction. The examination tasks should then be changed carefully, for example, by introducing tasks that cover the concepts introduced by the educational standards. Science is part of the final examinations almost exclusively in the ‘Abitur’. Here, these new tasks could be related to experimental methods, modelling or scientific ways of thinking’. (German minutes)
- ‘Assessment could be more diverse than it is nowadays, is the impression. It should be based on diverse evidence of learning. The new core curriculum emphasizes more competences than previous one, so assessment must chance as well’. (Finnish minutes)
- ‘There is a need to engage students in assessment in a way that focuses on assessment for learning. Students work in a different way with peers than with the teacher. There is a need to change the assessment culture so that the way in which students work with the teacher resembles the way they work with peers. Teachers need to change their teaching practices to integrate formative assessment’…. ‘there is [however] a tradition for summative assessment and grading, which is not easy to change. The French NSP suggests collaborations with teaching institutions and research collaborative groups in establishing teaching innovation projects’. (French minutes)
- ‘There is a need to change the assessment culture because the used assessment strengthens only the external motivation […] students and teachers don’t think enough about the learning goals, characteristics of quality performance and products [and since] the school assessment doesn’t support quality of students’ learning with respect to understanding of content … teachers are mainly focused on the fact whether the students have learnt the topic or not […] teachers don’t discuss the mistakes with students very often. However, there is still a demand for grades, the teacher needs them to make final certificates, and parents are used to work with them too’. (Czech minutes)

What Will Be the Best Strategy for Changing the Assessments in a Direction That Takes the ASSIST-ME Results into Consideration? And in Which Ways Can You, as a NSP, Help the Changing Process?

- ‘Examples of tasks with the described competencies as well as guidance for assessment and examples of students’ work. These materials will help future teachers and in-practice teachers with formative assessment. In addition, there is a need to expand the understanding of school assessment (…) as it should not be seen only as a tool, but also as an (educational) goal (…) we could use some results from this project to promote this change’. (Czech minutes)

- 'You need to combine initiatives on a local level with centrally decided changes in the examination system. Start experiments with no-marking classes (within the given regulations), increased emphasis on formative assessment etc, and evaluate the results. At the same time design new examination forms and concrete strategies for assessment, research on the implementation and present the results for the Ministry of Education'. (Danish minutes)
- 'There is a call for integrating formative assessment into pre-service teacher training. Another strategy is to generate projects by the local institutions that integrate formative assessment practices in their projects'. (…) All the NSP members support actively the French ASSIST-ME Conference in Grenoble in October 2016. They will take part in some panel discussions during the conference. Each NSP member can help the changing process in its own level:

Research associations (ARDIST, ARDM) can help for dissemination and the sharing of research results (theoretical and experimental) within the research network that include a lot of teacher educators. They can also share results with teachers' associations in order to reach teachers.

The DEGESCO (General Direction of School) can present information on the national site EDUSCOL (An official site for school educators and teachers that aim at informing and supporting teachers). DEGESCO can also pass information onto National Education school inspectors.

National Education school inspectors can support our action in the National Plan of Formation for in-service teachers' professional development. They said that we need to invest in pre-service teacher's education. Collaborative research groups (in-service teachers, teacher educators and researchers) can develop specific training actions about formative assessment in in-service teachers' training'. (French minutes)

Summing Up the NSP Discussions and Work

All NSPs had very engaged discussions, reflecting the importance that all stakeholder representatives attributed to assessment. Despite quite different foci and nuances, as illustrated by the excerpts above, it was possible to extract some common agreements and recommendations which will be outlined in the following 'general recommendations' section.

On reflection on the importance of having a forum for dialogue across political interests and power relations in order to establish mutual trust and understanding, it was clear that the NSP constituted a free space for exchange of ideas and points of view. It was also clear that many controversial questions were openly discussed. The discussions often revolved around the accountability purposes of assessment versus the learning purposes and the potential contradictions between the two purposes.

Laveault (2015) refers to two accountability orientations proposed by Blackmore (1988):

- Policy targeted at improving the management of the school system: Economic-Bureaucratic Accountability (EBA)
- Policy targeted at improving students' learning: Ethical-Professional Accountability (EPA).

An EBA orientation will increase students' performance and achievement levels through enhanced efficiency in the use of human and material resources. 'Teachers are directly held responsible for students' achievement results and therefore, should use AfL [Assessment for Learning] to improve them. Hence, in such a context, "The results are what matters, and the processes are validated only by performance" (Jaafar and Anderson 2007, p. 211)' (Laveault (2015), p. 23f). This approach has as the only purpose a clear accountability use. An EPA orientation will be based on a shared responsibility, and 'Emphasis is put on teachers working together as a professional learning community and on students' improved learning skills and sustained achievement levels (Jaafar and Anderson 2007)' (Laveault 2015, p. 23f). This approach attempts to combine a learning purpose with an accountability policy. A more expanded discussion of the problems involved in such an approach can be found in Chap. 3.

The tension between these two orientations was clearly articulated in the NSPs, and it became clear to all stakeholder representatives that a solution was necessary and that it could only be developed through strengthening the cooperation between teachers, researchers and key stakeholders, especially from official bodies like the Ministries of Education.

General Recommendations

The ASSIST-ME Lyon meeting in February 2016 had as a main theme the summing up the findings and formulating the project's general recommendations, which was reported to the EU as Deliverable D7.3. All partners brought with them the minutes from the NSPs and also minutes from the last meeting in their LWGs where teachers had discussed questions very much in line with the questions for the third NSP meeting.

Thematic groups were formed with members from all partner institutions. The groups wrote down statements everyone could agree upon based on the different national statements and conclusions. These statements were then edited together into five broad recommendations for policy-makers and other key stakeholders on how formative assessment of inquiry-based teaching and learning might be done more effectively across a range of countries. This process builds on the trial implementation of the four ASSIST-ME assessment methods (i.e. marking (grading and written comments), self and peer feedback, on the fly interaction and structured assessment dialogue) and recommends how these approaches can be strengthened and how existing assessment systems might be modified to enable formative assessment to function effectively in STM classrooms. Many of the recommendations are neither new nor ground breaking to a researcher in the field of STM education, but they have been carefully discussed among stakeholders from the partner countries involved in educational issues, and for many policy-makers they constituted rather strong statements.

Recommendation 1: A Competence-Oriented, Inquiry-Based Pedagogy Is Important

An inquiry-based teaching and learning approach helps young people develop critical thinking and scientific reasoning that are important in creating citizens who can make sense of the world they live in and make informed decisions. Inquiry-based teaching and learning has proved its efficacy at both primary and secondary levels in increasing interest and attainments levels in STM subjects, while at the same time stimulating teacher motivation. The ASSIST-ME project confirms this understanding and goes further, in defining and operationalising key competencies within STM subjects that help students utilise and develop scientific knowledge and processes.

The project points at ways to implement such a competence approach in different educational cultures and recommends adjusting educational policies to make this possible.

Recommendation 2: Focus on Formative Assessment to Support Competence-Based Inquiry Learning

Formative assessment provides both the time-frames and opportunities to look at how students develop competencies. ASSIST-ME has collected solid evidence of the huge learning potential of formative assessment methods via student goal orientation, making the learning journey visible and explicit. It has also supported teachers to identify the best next steps in student learning. However, the project has also revealed that formative assessment is not an integrated part of current STM teaching and that, for many teachers and students, it is difficult to implement in a structured form.

It is therefore necessary to promote a teaching approach integrating formative assessment into the classroom culture and to frame the educational condition, resources and the curriculum to make it happen.

Recommendation 3: Reduce the Emphasis on Summative Assessment to Give Room for Formative Assessment

ASSIST-ME found that the summative assessment load needs to be reduced to allow teachers time to focus more on formative aspects, on assessment for learning, and to highlight and emphasise those aspects of learning that we value within the STM community. Curricular material, textbooks and resources need to include specific and detailed reference to their formative potential and to their use in the classroom so that both teachers and students are focused on how assessment can support learning.

It is recommended to develop national assessment policies that recognise the different purposes and potential involved in the interactions between formative and summative assessment and that makes it possible to realise the full potential of formative assessment processes.

Recommendation 4: Develop New Forms of Examination Able to Capture STM Competencies

The ASSIST-ME implementations have made it clear that a big gap exists in many countries, between the examinations at the end of a course and the learning processes during the course. While the teaching-learning processes are aiming at developing the learners' STM competences, the examinations often fail to assess these properly. To bridge this gap, summative assessments should be more in alignment with the formative processes in everyday teaching and should be designed to assess the STM competences in a valid and reliable way. There is evidence from the project that classroom practice is heavily influenced by the well-recognised backwash from summative examinations. The development of examination forms that assess a broader range of STM competences will have a positive impact on the teaching of these competencies.

It is necessary to develop new types of examination that are able to capture the central STM competencies and also be aligned to the formative approaches in the classroom.

Recommendation 5: Teachers Need Support in Implementing and Enacting Classroom Assessment of STM Competencies

ASSIST-ME has developed formative assessment tools able to support teachers in defining and articulating appropriate feedback comments for students, thereby strengthening the assessment literacy of both teachers and students. Assessment tools alone are insufficient, though, teachers need to adapt the tools to the educational contexts that exist in the local environment. This requires support from peers and educators in translating tools for specific contexts. The ASSIST-ME model may provide an effective format for these programmes: three meetings per year, involving a team of teachers and researchers or teacher educators in designing and testing new teaching units in an iterative process dedicated to the ongoing improvement of the inquiry activities and their assessment tools.

ASSIST-ME has identified a strong need for professional development programmes (pre-service, induction and in-service) that support teacher understanding of formative assessment and inquiry-based teaching and learning and facilitate the implementation and enactment of formative assessment processes in STM classrooms at both primary and secondary level.

The Impact from ASSIST-ME on Education and Educational Policy

It is difficult to measure the long-term impact of an educational research project such as ASSIST-ME. Education is a complex interplay between the teachers in the classroom and school leaders; both framed and conditioned by educational policy. Research can influence how some teachers teach their classes and how some school leaders support and promote new teaching approaches – within the given frame. It can also affect the policy system to change the conditions for teaching, thus from a top-down position steer or support teachers' work. The ASSIST-ME project intended to do both.

Through the work in the LWGs teachers have been supported in changing their assessment practice. They have been introduced to various assessment methods and their implementation in the classroom as described in Chaps. 4, 5, 6 and 7. This work has been disseminated through national channels such as national conferences, teacher journals, teacher networks, in-service teacher training, etc.

The national policy level has been influenced in various ways from country to country.

In Denmark, a new Act for upper secondary education was passed through the parliament in 2016 with the intention of implementing new assessment forms and formats inspired by ASSIST-ME. An in-service programme for teachers and school leaders focusing on increasing student feedback is running in the spring 2017 with participation of ASSIST-ME researchers. The Danish Ministry of Education is launching a school experiment programme allowing schools to minimise summative assessments and to put more emphasis on formative assessments. A new Danish strategy for science education K-12 is formulated during spring 2017 – and ASSIST-ME knowledge is used and influencing the strategic goals and the initiatives adopted to realise the goals.

Each ASSIST-ME partner has their own story about influence, depending on the relations they have been able to establish to policy-makers. But besides these direct effects on policy, the project has had impact on the research field associated with formative assessment through documentation of features that seem to enhance or impede the enactment of the assessment methods.

A Wider European Perspective

To put ASSIST-ME in a wider European perspective, this section will review some of the different projects that sought to influence educational policy in Europe. The first is the 'mother' of all the inquiry-based science education projects within EUs Seventh Framework Programme in Science in Society. The programme rolled out in 2016, and ASSIST-ME was one of the last funded projects in this initiative. The following three country-specific cases are discussed in the next section: The Norwegian assessment for learning project illustrates a strong state controlled TPD programme, the New Standards and Curricula in Switzerland is an example of a curriculum project in alignment with many of the ideas in ASSIST-ME, and, finally, the English case incorporates reflections on the impact – or lack of impact – of a well-known science education policy initiative.

The cases together with the ASSIST-ME experiences provide a background for the perspectives of the chapter.

The Rocard Report

The publication Science Education Now: A renewed pedagogy for the future of Europe (European Commission 2007) – the so-called Rocard report – is an example of how science education policy was formed within the EU's Seventh Framework Programme in Science in Society. In response to a report from the OECD in 2006 (Evolution of Student Interest in Science and Technology Studies – Policy Report; Global Science Forum), a high level expert group on science education was formed to make recommendations based on a review of projects that seemed to be having a positive impact on recruitment to the sciences, and the report identified the necessary preconditions for increased implementation throughout Europe. The estimated level of support within the Science and Society (SIS) programme at that time was estimated to be 60 million euros over a 6-year period. The leader of this committee was Michel Rocard who was a member of the European Parliament and former Prime Minister of France. The other five members were leading scientists in Europe with only one science educator, Doris Jorde, on the committee. Policy was made by looking at existing documentation on what was working in science education, particularly with inquiry-based science teaching (IBST), as well as the identification of successful projects (wherein Sinus plus and Pollen were recognised as good examples for scaling-up). Michel Rocard gave the committee the authority required to prioritise science education within the EU.

What followed was the release of project funding to IBST projects involving science educators and informal science learning environments throughout Europe. The number of projects, number of participants from every country in Europe and number of ideas for promoting IBST are daunting in retrospect. An overview of the projects can be seen at www.scientix.eu.

A Norwegian Case

The Norwegian National Curriculum (2006) is written in the form of competency goals for all subjects in grades 1–13 (1–7, 8–10, 11–13). Competency goals in the integrated science subject (naturfag) occur in grades 2, 4, 7, 10 and 11. Biology, Chemistry, Physics, and Earth Science have competency goals for grades 12 and 13. National exams are administered after grade 10 (pupils may be selected for an oral examination) and at the end of grade 13 (students may be chosen for either oral or written examination). All children have the 'right' to formative evaluation in all subjects according to the national rules and regulations governing schools.

Based on the information above, one could say that Norway has very little testing compared to many other countries. Norway participates in TIMSS and PISA, but these tests are designed to identify trends at the national level. The Department of Education places trust in local school evaluation, supporting these efforts through national programmes for TPD in evaluation and the development of web pages (http://www.udir.no/laring-og-trivsel/vurdering/). These help explain and provide tools for evaluation (both formative and summative).

In 2010, the Department of Education launched a national TPD programme in 'Vurdering for læring' (assessment for learning) to improve school practice and thereby improve student learning outcomes. As of 2016, seven groups of teachers (representing over 300 municipalities) have participated in the programme in which schools build up communities of practice. Four principles guide the TPD programme.

Students learn best when they:

1. Understand what they are to learn and what is expected of them
2. Receive information on the quality of their work
3. Receive information on how to improve their work
4. Are involved in their own learning processes (through evaluation of own work and evaluation of progress)

The focus of the professional development programme is to move thinking away from concentrating on the learning activities students are doing to a focus on what students are learning.

As entire schools participate in these programmes, it is important to ask if the information is improving the way science teachers work in their own classrooms with assessment. Are the ideas presented above applicable for the practices employed in teaching science?

Do we know how to assess student progression in science lessons? Do we have a good 'tool box' in science for helping teacher to make student thinking visible?

Perhaps what is more important for us to think about when working with general ideas of assessment in schools is whether we, as science educators, have taken this type of pedagogical language into our own way of thinking about assessment in science classrooms. In science TPD and pre-service teacher education programmes, science educators often use their own language of assessment ideas and language, not necessarily corresponding with the literature presented in the 'assessment for learning' national TPD programmes.

Combining the four points above with how we work with science teachers (pre-service and in-service), the following questions are pertinent:

Conducting laboratory exercises is common in science lessons. When science teachers decide that students will conduct a lab: Are students clear about why they are doing the lab? Do teachers let students know what will be expected of them? Are clear goals for the lab articulated to students before they begin, rather than simply providing recipes for a procedure? Are students allowed to engage in the process of designing the lab?

Do teachers let students know how the lab will be assessed? Is it only the final lab report that is important or is the entire process of doing the lab also assessed? If the whole process is important, how will it be documented? Are students given the opportunity to discuss their results with others? Are students given the opportunity to use modern technology when collecting data in the lab (smart phones for example? and in their laboratory reports?

Are students given feedback on their lab reports? Are students given the opportunity to wonder about outcomes from the lab, including alternative interpretations? Are students asked to place the lab into an historical context or perhaps wonder about the importance of the outcomes in a modern science perspective?

The assessment for learning project in the Norwegian curriculum places a focus on what students are learning, not just on what students are doing. In science lessons, it is easy for science teachers to see activity and assume that learning is happening. However, it is not until we look for evidence of learning, through formative assessment tools, that we improve student outcomes.

New Standards and Curricula in Switzerland: A Focus on Inquiry-Based Learning and Assessment

In the last 15 years, Switzerland has begun to implement new standards and curricula which will lead to substantial changes in the Swiss educational system. In science, inquiry-based learning and formative and summative assessment play an important role in both standards and curricula.

National Standards Triggered by PISA results, the Swiss Conference of Cantonal Ministers of Education (EDK 2011; Labudde 2007; Labudde et al. 2012) initiated the project HarmoS (*Harmonisierung obligatorische Schule Schweiz*: Harmonisation of the Compulsory School Switzerland). Switzerland has a federal political system, that is, each of the 26 cantons has its own educational system. The mobility of the population demands for more harmonisation in education. HarmoS intends to establish comprehensive competency levels and standards in specific core areas for compulsory schools in Switzerland including science. The standards are defined for the end of grades 2, 6 and 9, that is, they are based on a progression model. The national standards in science include six skills: (1) asking questions and investigating; (2) exploiting information sources; (3) organising, structuring and modelling; (4) assessing and judging; (5) developing and realising; and (6) communicating and exchanging views. As an overall skill 'working self-reliantly and reflecting on one's own work' is added. Each of the skills – and their subskills – has been described meticulously and with a degree of sophistication. Many of them are related to inquiry-based learning. The following examples belong to the skill 'asking questions and investigating': (a) at the end of grade 6, that is, 12-year-old students, and (b) at the end of grade 9, i.e. 15-year-old students:

(a) Grade 6: 'Students can perceive simple situations and phenomenon with different senses; they can observe and describe them. In regard to the situations and phenomenon, they are able to ask questions and formulate problems. Guided by the teacher, students can perform investigations and experiments; they can carry out estimations and measurements, collect and interpret data'.
(b) Grade 9: 'Students can perceive situations and phenomenon with different senses; they can observe and describe them. In regard to the situations and phenomenon, they are able to ask different questions and to formulate problems and simple hypotheses, and to determine variables in order to check them. They can plan and perform investigations and experiments. Doing so, they are able to carry out specific estimations and experiments, to collect and interpret data, and to answer their questions and to give their view on the hypotheses'.

Curriculum 21 The national standards and competences, as defined by HarmoS, were the frame for the development of the new curricula, one for each of the three linguistic regions of Switzerland, that is, German, French and Italian. For the German speaking part of Switzerland, that is, 21 cantons out of 26 cantons, the curriculum was named 'Curriculum 21' (*Lehrplan 21*). The subjects 'Nature-People-Society' (grades K-6, i.e. age 4–12) and 'Nature and Technology' (grades 7–9, i.e. age 12–15) focus on both inquiry-based learning and formative assessment.

- The curriculum defines hundreds of so-called competences; dozens of them are related to inquiry-based learning. For example: 'Students can investigate, reflect, and present information about nature and technology on their own. […] Students can plan, implement and interpret investigations in regard of the interaction of plants and soils' (D-EDK 2014a, NT 1.3NT 9.3c).

- The so-called basics of Curriculum 21 describe formative and summative assessment: ‘Formative assessment: During the lessons, the students receive encouraging and constructive feedback, which they can use developing competences and which supports their learning process. This feedback is adapted to the individual learner, it integrates aspects of self-assessment. [...] Summative assessment focuses on the actual performance level of students. Primarily, it is based on the objectives of the curriculum [...]’ (D-EDK 2014b, p. 9–10).

The quotes are paradigmatic examples showing how inquiry-based learning and formative assessment as two different instructional strategies are integrated in Curriculum 21. Both have already played an important role in previous curricula, but they are enforced by the new curriculum which will be implemented gradually in the next few years. Each canton has its own schedule for the implementation; the first two cantons already started in August 2015 and the last ones will start in 2020.

Teacher Training Programmes and Continuous Professional Development (CPD) Teacher training colleges and institutions of CPD play an important role in the cantonal educational systems; CPD is well developed and well recognised by teachers and other stakeholders. For more than 20 years, these institutions offer modules and courses that promote Inquiry Based Learning (IBL). However formative assessment has seldom been an explicit part of programmes or CPD. This situation will change, though; with the implementation of Curriculum 21, there will be large programmes of CPD in all cantons. Teachers should learn about the objectives and content of the new curriculum. In science, a main emphasis will be on both IBL and formative assessment. The same is true for teacher training programmes across Switzerland.

Politics and Administration In general, both politicians and administrators support the ideas of IBL and formative assessment. In a small country like Switzerland, with a population of only 8 million people, there are many relationships between cantonal ministries of education, teacher training colleges, teacher unions and associations and teachers. Therefore, a new development or a new concept that is well accepted by different kinds of stakeholders, like the concept of IBL, will also be accepted by other stakeholders. In addition to formative assessment, the Swiss Conference of Cantonal Ministers is implementing national monitoring that includes nationwide tests in the main subjects, that is, mathematics, first and second language, and science. The monitoring will be based on a representative sample; it should yield results with regard to the implementation of standards and curricula: Which competence levels do the students achieve by the end of grades 2, 6 and 9? This monitoring is a form of low-stakes assessment. Furthermore and in contrast to the national monitoring, some cantons are planning and implementing high-stakes assessments at the end of grade 6 and 9. The results of these assessments will be part of the final certificate that the students get at the end of primary and secondary school.

The Role of ASSIST-ME The project ASSIST-ME only started in 2013, that is, it could not have had an influence on the established national standards and only a marginal influence on Curriculum 2014. However, it has had and will continue to have effects on the implementation of the new curriculum, on teacher programmes and CPD and on formative and summative assessment. The Centre for Science and Technology Education (CSTE), which is the Swiss partner in ASSIST-ME, and its members are responsible – in most cases together with other institutions – for the following activities and projects:

1. Educating pre-service teachers at our university as well as being engaged in CPD, for example, in the big programme SWiSE (2017, Swiss Science Education, customers: Teachers, ministries of education, universities, and foundations; since 2011).
2. Planning and developing checks in science for the end of grade 8 and 9; IBL and experiments play an important role in these checks (customers: ministries of education of the four cantons of Northwestern Switzerland; annually since 2014).
3. Organisation of a 1-day conference on formative and summative assessment in science for cantonal ministries (about 30 specialists of cantonal ministries; November 15, 2016).
4. Part of the advisory board for the national monitoring (customer: Swiss Conference of Cantonal Ministers, since 2016).
5. Writing science school books for grades 7–9 with an emphasis on IBL and with hints for formative and summative assessment in the teacher edition (customer: a Swiss publishing company; 2014–2020).
6. Helping to implement Curriculum 21, in particular the ideas of IBL and formative assessment.

These activities and projects show that ASSIST-ME has had and will continue to have an impact in Switzerland. Thus, there are a lot of synergistic effects between the European project and several Swiss projects.

An English Case

Over a decade ago the Nuffield Foundation sponsored two seminars which resulted in a report entitled ‘Science Education in Europe: Critical Reflections’ (Osborne and Dillon 2008). Although the majority of the attendees were science education researchers, there were a scientist and a policy-maker from the EU at each seminar. The overall approach was inspired by the series of Nuffield Foundation seminars that led to the seminal UK science education report, Beyond 2000 (Millar and Osborne 1998), which has been cited (according to Google Scholar) almost 1500 times.

‘Science Education in Europe’ examined three aspects of science learning: curriculum, pedagogy and assessment. The report was highly critical of many aspects of the then current policy and practice. In terms of assessment, the report's authors concluded that:

> For too long, assessment has received minimal attention. Tests are dominated by questions that require recall – a relatively undemanding cognitive task and, in addition, often having limited validity and reliability. Yet, in many countries, the results of a range of tests, both national and international, are regarded as valid and reliable measures of the effectiveness of school science education. Teachers naturally, therefore, teach to the test, restricting and fragmenting the content and using a limited pedagogy. (Osborne and Dillon 2008, p. 9)

Those comments still hold true today in many, if not all, of the ASSIST-ME partner countries. The report went on to identify a possible course of action:

> Transforming this situation requires the development of assessment items that are more challenging; cover a wider range of skills and competencies; and make use of a greater variety of approaches – in particular, diagnostic and formative assessment. (p.9)

A number of individuals and institutions have done just that although implementing new assessment strategies across educational systems has proved problematic.

The report made a number of recommendations including:

EU governments should invest significantly in research and development in assessment in science education. The aim should be to develop items and methods that assess the skills, knowledge and competencies expected of a scientifically literate citizen. (p.9)

The EU did invest in research and development in assessment in science education but how far have we come and where are we going? If anything, the dominance of large-scale international testing such as PISA has grown over the last decade. While some have argued that such comparisons have been used to lever up standards in a number of countries, others argue that such tests distort policy and thus classroom practice.

Reflecting on the impact of 'Science Education in Europe' one could argue that while the European Commission was sympathetic to its messages, systemic change has foundered on the way that politicians interpret the results of the OECD PISA assessments together with a natural inertia in education systems to any radical change in student assessment.

Nowhere in the world is student assessment as much of a political football as in England. With its long history of practical work in school science and its concern for science for all, one might expect that England might have the answers to many of the questions that drove the ASSIST-ME project. The pioneering work of Paul Black and Dylan Wiliam is well known to anyone interested in formative assessment, and it was a starting point for many of the ASSIST-ME developments. But how far has England come?

In February 2017, the Wellcome Trust in collaboration with the Gatsby Charitable Foundation announced a new scheme which focused on 'Assessing Practical Science Skills in Schools and Colleges'. The scheme 'supports researchers who want to explore the best ways to assess students' practical science skills'. Funding would be made available to support researchers who wanted to address 'the challenge of assessing students' practical science skills in a way that is valid, reliable and feasible. Rather oddly Wellcome added 'We might also consider other ideas that don't meet these criteria, but develop new ways to assess students' practical science skills'. There does seem to be an element of wheel reinvention here. Surely in 2017 enough is known about how to assess students' practical skills in science? Should not the priority be transforming science education as the Nuffield report recommended through 'making use of a greater variety of approaches – in particular, diagnostic and formative assessment' (Osborne and Dillon 2008, p. 9).

Discussion of science education policy in England tends to be dominated by the learned societies and by the Association for Science Education, the leading professional organisation for those involved with the teaching of science. Despite the many working parties, committees and conferences that have been devoted to assessment in science in England, the level of thinking does not seem that advanced. Furthermore, the impact that these discussions, pontifications and reports have on policy-makers seems limited. Some measure of where the debate is now in England is indicated by this quote from SCORE – the Science Community Representing Education.

> Assessment largely determines what students are taught, and has an enormous influence on the style and emphasis of teaching and learning. Therefore, it is essential that awarding organisations, Ofqual and the Department for Education work together across organisations, and with others, to ensure that an effective, evidence-based mechanism for assessing practical work is developed alongside content. (SCORE n.d.)

It might be 'essential' that the Department for Education works with other organisations but the mechanism for making this happen is not clear. All too often in science education in England, the rhetoric and the reality are far apart.

Conclusions and Perspectives

So, what can we conclude? The science education research community seems to have accepted that changing classroom assessment practices towards more emphasis on formative processes is a key dimension of raising attainment in school science. But the dominating political power still favours the summative elements.

The ASSIST-ME project has demonstrated that bringing researchers and teachers and policy-makers in close dialogue about concrete issues regarding assessment gives an enhanced awareness and understanding of the role of formative and summative use of assessment. This has influenced the attitude of the individual policy-maker and given him or her a more nuanced view on problems related to assessment and an informed openness for debating solutions.

It is also evident that ASSIST-ME has had influence on the educational policy in the participating countries. The extent of influence differs, of course, from country to country, and many effects will only be visible after a longer period of time. But all partners report of impact on the educational and political scene in their country. Many have arranged national conferences including policy-makers, and all partners give various examples on how their researchers have been invited to participate in curriculum development, teacher professional development workshops, expert groups on assessment, etc. Through such activities and through the NSPs, the results from ASSIST-ME have been spread and have affected the public discourse.

It has been crucial that dissemination of information and the debates in the NSPs were based on research results originated from collaboration between researchers and teachers. The research foundation gave the background for the debate a high degree of legitimacy, both among teachers and policy-makers, and thus some seriousness and credibility to the outcome. It has also been fruitful that NSP members were involved indirectly in the research process. It gave them insight into the complexity of educational research and a certain humbleness towards quick solutions.

This has not necessarily led to consensus, which is a rare thing in policy, and it has not eliminated all resistance to change. Much resistance to change has legitimate reasons seen from the point of view of the actor. Teachers have limited capacity for change if not giving the necessary supports. Professionals need time and training for professional development. Policy-makers are often restricted in their actions by the point of views they normally represent, talking on behalf of their policy or interest organisation. But the ASSIST-ME approach has pointed at some ways forward for affecting educational policy.

What needs attention in the future is whether the lack of profound changes in the assessment systems can be tracked down to inertia in the educational environment or to resistance to change among policy-makers. It is necessary to pose the question: Is the lack of change in the directions indicated by research such as ASSIST-ME due to lack of knowledge among policy-makers – or is it rooted in values among the same policy-makers opposing the research findings?

Maybe it is about time to realise the political character of educational policy issues, especially issues related to assessment. When things don't change, it could be because strong economic and political powers don't want them to change. Some policy actors might simply find a change in the fundamental ways the educational system is currently functioning a threat to their basic values. In this respect, change in the assessment system is a question of value clarification and value change. This approach is, perhaps, key to understanding where we go next.

References

Blackmore, J. (1988). *Assessment and accountability*. Victoria: Deakin University.

Bruun, J., Dolin, J., & Evans, R. (2015). *At the policy-research interface: Usefulness of social network analysis in identifying and selecting key stakeholders*. NARST2015.

D-EDK, Deutschschweizer Konferenz der kantonalen Erziehungsdirektoren. (2014a). *Lehrplan 21: Natur und Technik, 3. Zyklus*. Luzern: D-EDK.

D-EDK, Deutschschweizer Konferenz der kantonalen Erziehungsdirektoren. (2014b). *Lehrplan 21: Grundlagen*. Luzern: D-EDK.

Dolin, J. (Ed.). (2013). *ASSIST-ME proposal*. http://assistme.ku.dk/resources/. Accessed 18 Feb 2017.

EDK, Schweizerische Konferenz der kantonalen Erziehungsdirektoren. (2011). *Grundkompetenzen für die Naturwissenschaften.* Bern: Schweizerische Konferenz der kantonalen Erziehungsdirektoren. Retrieved March 6, 2017, from http://edudoc.ch/record/96787/files/grundkomp_nawi_d.pdf

European Commission. (2007). *Science education now: A renewed pedagogy for the future of Europe*. Luxembourg: Office for Official Publications of the European Communities.

Fensham, P. (2009). *The link between policy and practice in science education: The role of research*. Wiley InterScience.

Goldenberg, C., & Gallimore, R. (1991). Changing teaching takes more than a one-shot workshop. *Educational Leadership, 49*(3), 69–72.

Harlen, W. (2007). Holding up a mirror to classroom practice. *Primary Science Review, 100*, 29–31.

Hattie, J. (2008). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. New York: Routledge.

Jaafar, S. B., & Anderson, S. (2007). Policy trends and tensions in accountability for educational management and services in Canada. *The Alberta Journal of Educational Research, 53*(2), 207–227.

Labudde, P. (2007). How to develop, implement and assess standards in science education? 12 Challenges from a Swiss perspective. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Making it comparable: Standards in science education* (pp. 277–301). Münster: Waxmann.

Labudde, P., Nidegger, C., Adamina, M., & Gingins, F. (2012). The development, validation, and implementation of standards in science education: Chances and difficulties in the Swiss project HarmoS. In: S. Bernholt, K. Neumann, & P. Nentwig (Hrsg.). *Making it tangible: Learning outcomes in science education* (S. 235–259). Münster/New York/München/Berlin: Waxmann.

Laveault, D. (2015). Assessment policy enactment in education systems: A few reasons to be optimistic. In D. Laveault & L. Allal (Eds.), *Assessment for learning: Meeting the challenge of implementation*. Springer.

Lieberman, A., & Pointer Mace, D. H. (2008). Teacher learning: The key to educational reform. *Journal of Teacher Education, 59*(3), 226–234.

Looney, J. W. (2011). *Integrating formative and summative assessment: Progress toward a seamless system?* Paris: OECD Publishing.

Millar, R., & Osborne, J. (Eds.). (1998). *Beyond 2000: Science education for the future: A report with ten recommendations*. King's College London, School of Education.

OECD. (2011). *Evaluation and assessment frameworks for improving school outcomes. Common policy challenges*. Paris: OECD Publishing. http://www.oecd.org/edu/school/oecdreviewonevaluationandassessmentframeworksforimprovingschooloutcomespapersandstudies.htm. Accessed 18 Feb 2017.

OECD. (2013). *Synergies for better learning: An international perspective on evaluation and assessment.* OECD reviews of evaluation and assessment in education. Paris: OECD. http://www.oecd.org/edu/school/synergies-for-better-learning.htm. Accessed 18 Feb 2017.

OECD/CERI. (2005). *Assessment for learning – Formative assessment*. Paris: OECD Publishing.

OECD. (2006). *Evolution of student interest in science and technology studies*. Policy Report. OECD, Global Science Forum.

Osborne, J., & Dillon, J. (2008). *Science education in Europe: Critical reflections*. London: The Nuffield Foundation.

SCORE. (n.d.). *SCORE principles: The assessment of practical work*. Available at: http://www.score-education.org/media/14286/score%20principles%20for%20the%20assessment%20of%20practical%20work%202014.pdf. Accessed 18 Feb 2017.

Scott, S. (2010). The theory and practice divide in relation to teacher professional development. In J. O. Lindberg & A. D. Olofsson (Eds.), *Online learning communities and teacher professional development: Methods for improved education delivery* (pp. 20–40). Hershey: IGI Global.

Scott, J., & Carrington, P. J. (2011). *The SAGE handbook of social network analysis*. Los Angeles: SAGE.

Shelley II, M. C. (2009). Speaking Truth to Power with Powerful Results: Impacting Public Awareness and Public Policy. In: Shelley II, M.C. et al. (eds.), *Quality Research in Literacy and Science Education*. Springer.

SWiSE. (2017). *Swiss science education.* Retrieved March 6, 2017, from www.swise.ch

Chapter 11
Transforming Assessment Research: Recommendations for Future Research

Jan Alexis Nielsen, Jens Dolin, and Sofie Tidemand

Introduction

As a research project, ASSIST-ME produced a large number of results both within and across the eight participating partner countries using a variety of research methods. Based on the preceding chapters, this chapter will organise, prioritise and summarise the principal outcomes. It seems reasonable to assume that many of the findings presented in the preceding chapters can inform further research with the fields of classroom assessment or science education (or both). For example, it is quite clear that the concept of inquiry teaching, while being central in the field of science education for two decades, still is difficult to define in clear and uniform terms (Rönnebeck et al. 2016). In concert, the chapters at the very least provide a state-of-the-art terminology about inquiry-related learning outcomes and how they are assessed that can act as a strong scaffold for future research on inquiry Science, Technology and Mathematics (STM) teaching.

We will in this chapter identify and outline current gaps in research into assessment practice and tie the results of the ASSIST-ME project onto this outline. In this way, the chapter will present concrete research vistas that are still needed in international assessment research. The chapter concludes with a key theme that appears across many of the chapters in this volume, namely, issues concerning the *operationalisation* of complex learning goals into teaching and assessment activities.

J.A. Nielsen (✉) • J. Dolin • S. Tidemand
Department of Science Education, University of Copenhagen, Copenhagen, Denmark
e-mail: janielsen@ind.ku.dk

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3_11

Cross-Cutting Trajectories

We start by extrapolating three trajectories across the chapters that seem to be particularly promising for future research. These pertain to (i) using competences as a theoretical foundation in assessment, (ii) placing summative and formative assessment on a continuum and (iii) identifying the need for teachers to be supported when introducing new assessment formats. These trajectories will subsequently be discussed in the ensuing sections of this chapter.

First, Ropohl et al.'s (Chap. 1) exposition of *competences* as a way of parsing learning objectives in science education and Rönnebeck et al.'s (Chap. 2) delineation of how competences related to STM inquiry could be *assessed*, provide a very direct vision of how the field of science education could understand and make operational concepts such as *Bildung*, scientific literacy and inquiry teaching (for a discussion of the connection between scientific literacy and *Bildung*, see Sjöström and Eilks 2017). Ropohl et al.'s (Chap. 1) analysis of the concept of competence indicates that the concept is multifaceted and often used ambiguously, but in so doing, the chapter provides much needed clarity and indicates ways forward by approaching competences from what they call a 'holistic' perspective. Further, Ropohl et al. provide a general vantage point for an understanding of competences within science, technology and mathematics education. In particular, Ropohl et al.'s push to conceptualise complex learning objectives in competence terms forms a backdrop for Rönnebeck et al.'s (Chap. 2) analysis of assessment of learning in inquiry teaching. Rönnebeck et al.'s analysis builds on a detailed systematic literature review reported in Bernholt et al. (2013), and the chapter provides a much needed translation of inquiry-related learning outcomes into competence terms, not just for research to reach a shared understanding of inquiry but also indirectly to support teaching and assessment practice concerning inquiry teaching.

Second, while researchers within the field of classroom assessment have long distinguished between formative and summative assessment, Dolin, Black, Harlen and Thiberghien (Chap. 3) provide a framework for understanding the dynamics of the interplay between formative and summative assessment. In particular, the chapter contains concrete ideas for ways of linking formative and summative forms of assessment and for how formative and summative forms of assessment can be seen as belonging to a spectrum rather than as binary forms of assessment. Now, these ideas have bearing on the empirical studies of both *structured assessment dialogues* (Dolin et al., Chap. 3) – a regimented assessment procedure that allows for multifarious layers of assessment activities and perspectives – and the scaffolded approach to teacher *written feedback* explored by Holmeier et al. (Chap. 7).

Third, it seems that the teachers who were involved in the various empirical studies presented in this volume in general needed support to use the various assessment formats as well as to plan and implement inquiry STM teaching (see Chaps. 4, 5, 6, 7 and 9). In particular, at least three chapters conclude that concerted professional development efforts are needed to support teachers to provide formative feedback in inquiry teaching (see Chaps. 4, 7 and 9). The need for support notwithstanding, it seems equally clear that the process of repeatedly trying out the different methods for assessing inquiry-related competences did provide a strong basis for the teachers to establish an understanding of inquiry (or rather inquiry-related competences) that in turn stabilised their formative assessment practice.

Recommendations for Future Research Foci and Methods

We next present what we find to be key lacunae in classroom assessment research that future research should address. As such, we reflect on key research foci that we feel should be pursued in the future. Further, we identify an underlying theme that reoccurs in various ways across many of the research findings in the chapters of this volume. We present this theme, argue why this theme is important and provide a terminological and analytical framework that may open a new research vista for classroom assessment research.

The chapters in this volume all address issues that are boundary objects (Star and Griesemer 1989). Indeed, the issues that are explored can be approached from the perspective of classroom assessment research which we can take to be a strand of general education research *and* from the perspective of science education research. When looking across the issues in the preceding chapters, it is difficult to define exactly when it is advisable to draw on background research from the general educational field and when to draw on research about assessment in subject-specific contexts – such as the STM subjects. Clearly, drawing on classroom assessment research in general would have afforded a much more comprehensive background into typical issues surrounding teachers' assessment practices regardless of which subject the teachers are teaching. But it is still an open question whether assessment practices differ across disciplines and if so how they differ and what such differences signify (Ruiz-Primo and Li 2012). Also, focussing on the assessment of complex inquiry competences may to some extent preclude the application of much of the existing research into subject-specific assessment practices that will often focus on assessment of concrete subject content. For example, for science education, there exists some research on the development and validation of learning progressions (Wilson 2009), but such work typically focusses on concrete subject-specific conceptual content rather than generic competences such as those that are the aims of inquiry teaching. Generic competences are here taken to be competences that can be at play in or the aim of multiple different disciplines, e.g. communication competences, innovation competences, problem-solving competences, collaboration competences and argumentation competences (see, e.g. Belova et al. in press).

Future research endeavours must take this into account: research into assessment in relation to inquiry teaching still requires a fair amount of extrapolation – either by assimilating findings from general education assessment research or by transposing findings and principles about assessment of more specific disciplinary skills.

Cowie (2012) recently argued that '[t]he current focus on more directly aligning the system of assessment (national to classroom), curriculum, and pedagogy comes with the prospect of this suggesting the need for research that tracks development at all levels of the system and across all stakeholders' (p. 484). The primary idea behind Cowie's call for comprehensive research of this type is the fact that the *assessment culture* of an educational system manifests itself in various ways across stakeholder types – for example, teachers' narratives about the assessment culture in which they are actors will probably be very different from the narratives of parents (compare, e.g. Brookhart 2012) or government officials overseeing centralised assessment (see Moss 2008, for an argument to a similar effect). To be sure, the ASSIST-ME project that formed the basis of the preceding chapters took an important step in this direction by adopting an involvement of key stakeholder groups: researchers, teachers, leaders and policymakers (see Dolin et al., Chap. 10). In the project, this stereoscopic involvement ultimately resulted from a *change agenda* – with the aim of actively capitalising on the research results in order to impact teaching and assessment practice (see Dolin 2012). Arguably, more stakeholder types – most notably the learners – could and should play a larger role in research that comprehensively tracks the development of assessment culture in increasingly aligned educational systems.

An important issue in assessment research that had a somewhat implicit role in this volume is that of *validity*. The coding of the data in the project did entail some aspects of assessment validity – in particular in the case of peer feedback (Chap. 6) and to some extent in the case of written feedback (where the coders had to focus on the level of justification of the teachers' feedback to his/her students; Chap. 7) – and of course validity is a key aspect considered in the theoretical outcome of the project focussing on how summative and formative assessment can be linked constructively (Chap. 3). Future large-scale cross-national research projects similar to the ASSIST-ME project may focus more directly on validity aspects. Indeed, the role of validity in educational assessment research is difficult to circumvent (for a historical overview and an exposition of the importance of validity in educational assessment, see Newton and Shaw 2014). At this point it is relevant to emphasise that the ASSIST-ME project worked with complex learning objectives (inquiry competences in STM) that were not necessarily familiar to the participating teachers and that this may indicate validity concerns (even though reliability concerns may trump validity concerns in matters about formative assessment; see Chap. 3).

Operationalisation of Learning Goals: A New Research Vista?

In this section, we want to point to an underlying thematic process that we find permeates the research findings in most of the preceding chapters. We provide a first attempt to define that process and in so doing point to new research vistas for research on teachers' assessment practices. Let us start by making some observations from three chapters that describe empirical studies.

Evidence from the use of *structured assessment dialogues* (see Dolin et al., Chap. 6) indicates that the translation of relatively complex learning goals into more concrete and operational constructs that can function on the level of assessment criteria proved to be important for the quality of feedback. Further, teachers' practice of engaging in on-the-fly formative interaction benefits from a close exposition of the assessment criteria or the construction of rubrics (see Harrison et al., Chap. 4). Similarly, the work put into translating more complex competence goals into criteria for the feedback templates used for written feedback seems to have been beneficial (see Holmeier et al., Chap. 7). To be sure, such templates that delineate the potential progression trajectories of students' competence development can aid the teacher in providing valuable feedback. But beyond this, the very fact that competence development is given a typified description seemed to help some of the participating teachers in making assessment transparent to their students.

What connects these findings is that the participating teachers went through a process of translating learning goals and that, in the context of these studies, this process to some extent was necessary for establishing high-quality assessment practices. Translating learning goals belongs to a process of *operationalising* learning goals. As such, the process has the aim of *making operational* learning goals, and the rationale behind the process is that the operationalised learning goals may provide *better* guidance for the teacher than the initial learning goals – by guidance we mean guidance on how to structure his or her teaching and on how to assess students' level of attainment of the learning goals. We have depicted a graphical model of the process of operationalising learning objectives in Fig. 11.1.

Take, for example, the following predefined learning goal for biology in a Danish upper secondary school: '[students should be able to] assess far-reaching biological issues and their significance on a local and global level' (Danish Ministry of Education 2013). On its face value, this particular learning goal arguably provides little guidance on how to structure teaching and assessment activities. In order for that learning goal to be instructive for a teacher (e.g. a teacher that is confronted with this learning goal for the first time), there must be a process through which the reader can *negotiate the meaning* of the goal (this negotiation of meaning could be

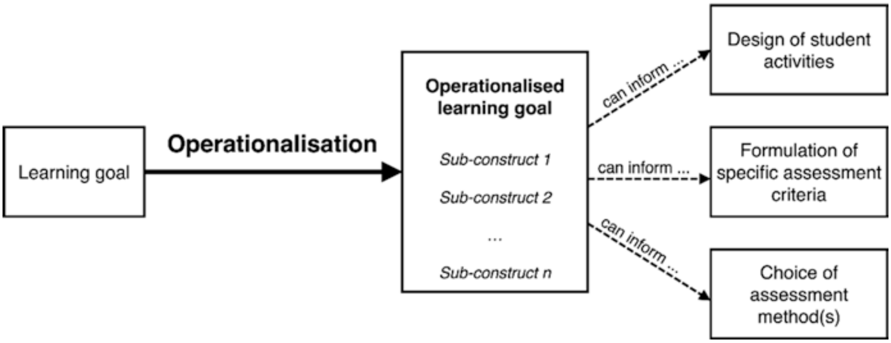

Learning goal
Operationalisation
Operationalised learning goal
Sub-construct 1
Sub-construct 2
...
Sub-construct n
can inform ...
can inform ...
can inform ...
Design of student activities
Formulation of specific assessment criteria
Choice of assessment method(s)

Fig. 11.1 A graphic illustration of the process of operationalising a learning goal. Notice that the product – the operationalised learning goal – does not necessarily involve sub-constructs; in many cases the product is an interpretation of the initial learning goal that enables the teacher to make decisions vis-á-vis teaching and assessment activities

scaffolded by attaching some sort of commentary to the curriculum). This kind of negotiation of meaning is exactly what the concept of operationalisation signifies. As we will describe below, the operationalisation of a learning goal can be more or less specific – ranging from a general interpretation to a detailed parcelling out of the initial learning goal into sub-constructs that can indicate assessment criteria on various taxonomical levels (see, e.g. Biggs and Collis 2014; Krathwohl 2002). So, at the very least, the operationalised learning goal, as a product of the process of operationalisation, is the teachers' interpretation of the initial learning goal.

The process that we are describing here should resonate quite well with both practitioners and researchers. Indeed, there is nothing new in that process. But it seems to us that it is important to specify the process in more detail than has been done in the existing literature. In fact, the process has often remained implicit in theoretical expositions. Clearly, the process of operationalising learning goals must be an important part of what Kattmann, Duit, Gropengiesser and Komorek (1996) called *Educational Reconstruction* – the process through disciplinary content is reconstructed into a curriculum or into teaching activities. Similarly, the process of operationalising learning goals would be a part of what Chevallard (1991) called the *internal didactic transposition*, i.e. the process with which a teacher transposes the aims of a curriculum into actual teaching (Winsløw 2011). But in both conceptualisations of the process from the disciplinary knowledge over the curriculum to the classroom activities, what we have called the process of operationalisation is at best only implied. While there seems to be no established body of work related to the particular process that we want to refer to as operationalising learning goals, the terminology we have chosen should be familiar to the field. For example, in the German curriculum, competence goals are fleshed out using 'operators' (German: *operatoren*) that are action verbs describing student activities that should be expected when the student is developing a particular competence.[1] Notice that our usage of operationalisation as a term indicates an active part of the teacher. While a curriculum can perform a part of the task of making learning goals operational, the teacher will need to perform at least a minimal operationalisation him- or herself (such as described by the notion of the internal didactic transposition).

We contend that any teaching practice will involve at least minimal processes of operationalisations of the sort we are describing here. But it seems reasonable to assume that the more complex the initial learning goals are, or the more unfamiliar a learning goal is to the teacher, the more there is a need for support for operationalisations. Arguably, the extent to which teachers go through interpretative processes that can be categorised as operationalisation of learning goals will vary between different educational cultures (see, e.g. Desurmont et al. 2008). A reasonable hypothesis could be that teachers in systems that belong to what can loosely be called the north and continental western European tradition have a strong tradition for going through such processes, for these are systems where the teacher traditionally has a relatively autonomous role of designing his or her teaching using a curriculum as a guide. But the explicit familiarity with processes of operationalisation

[1] See, e.g. https://lehrerfortbildung-bw.de/u_gewi/gwg/gym/bp2004/fb1/modul1/geo/operator/

can also differ between educational levels within one country. For example, findings from the ASSIST-ME project indicate that Danish lower secondary school teachers were much more familiar with processes that resemble operationalisation of learning goals than teachers from upper secondary school – probably due to differences between the curricula for lower and upper secondary school in Denmark (Nielsen and Dolin 2016).

It seems to us that there is a real need for thematising how to operationalise learning goals that teachers perceive as new and/or unclear (such as 'innovation competence'; see Nielsen 2015) or for other reasons perceive as unclear (such as is often the case with learning goals that relate to technology issues in science teaching; see Bungum 2006). Indeed, the previous chapters in this volume indicate that teachers for whom inquiry teaching introduced a new set of learning goals need substantial support in identifying viable strategies to plan their teaching and operate during their teaching vis-à-vis identifying and acting on opportunities to provide formative assessment (see, e.g. Harrison et al., Chap. 4).

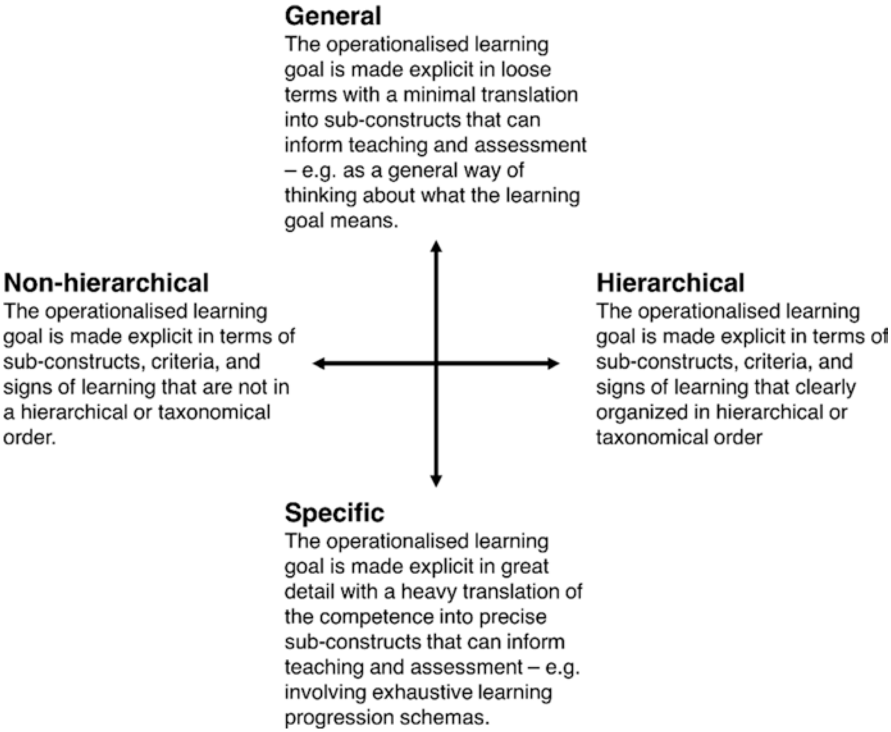
We hypothesise that the process of operationalisation is of paramount importance – whether or not a given learning goal is complex, unclear or novel. As has been argued by Dysthe et al. (2008), if assessment 'criteria are explicitly formulated as reifications of continuous negotiations and participation, they become part of a meaningful learning process[; …] [e]xplicit criteria cannot be understood in isolation from the negotiation process' (p. 127). Indeed, the findings in the preceding chapters corroborate this statement. Crudely put, a criterion in itself is not yet operational for teaching and assessment.

The term 'operationalisation' is frequently used in the field of validity research. For example, when investigating the validity of a construct, the key question is whether that construct was appropriately operationalised for functional measurement (see, e.g. Drost 2011). But we do *not* want invoking specific psychometric connotations with our usage of 'operationalisation' here. Our way of using 'operationalisation' also relates to questions regarding validity of assessment (*am I, as a teacher, really assessing the construct that I intend to assess?*), but we aim to signify a process which is closer to actual teaching practice and which pertains not just to assessment. Further, the term 'operationalisation' has been used in curriculum research. For example, Wiek et al. (2015) use 'operationalisation' to signify a process of making explicit a given general competence through a set of 'specific learning objectives for different educational levels' so as to inform curriculum design (p. 242). Again, we want to use 'operationalisation' here as signifying a process that is closer to actual teaching practice, rather than something that occurs during the construction of a curriculum.

There are clear indications in the findings of the ASSIST-ME project that the participating teachers found the process of operationalising competences overly time-consuming (see, Dolin et al., Chap. 3; Harrison et al., Chap. 4; Dolin et al., Chap. 5; Evans et al., Chap. 9). Moreover, many teachers in the project found it fundamentally difficult to operationalise general inquiry competences through learning progressions – even when teachers are being assisted by researchers. As reported by Dolin et al. (Chap. 5), the teachers who implemented structured assessment dialogues tended to formulate rubrics that essentially not had the structure of a progression but rather consisted of unstructured or non-taxonomically ordered signs of student learning. Such operationalisations could be called *non-hierarchical* operationalisations of competences.

In the narratives of the participating teachers (for an exposition, see Dolin 2016; Nielsen and Dolin 2016), there are indications that some of the teachers felt that detailed learning progressions could lead to some form of instrumentalist assessment paradigm (see also Torrance 2007) involving teachers and students in following rudimentary learning checklist. As such some of the teachers were opposed to using what could be called *specific* operationalisations of competences. This issue harks back to a discussion by Rönnebeck et al. (Chap. 1) about the extent to which a generic competence can or should be deconstructed into a myriad of smaller constructs or whether a more holistic approach is preferable.

These findings indicate to us the benefit of thinking about teachers' operationalisation of competences as an activity that leads to a product or outcome that can be analysed along two continuums or dimensions (see Fig. 11.2). First, the outcome of a concrete act of operationalising a competence can be more or less *specific*. At one end of the continuum, the sub-constructs and/or assessment guides that are formulated in order to make the competence operational can be very precise and minute,



General

The operationalised learning goal is made explicit in loose terms with a minimal translation into sub-constructs that can inform teaching and assessment – e.g. as a general way of thinking about what the learning goal means.

Non-hierarchical

The operationalised learning goal is made explicit in terms of sub-constructs, criteria, and signs of learning that are not in a hierarchical or taxonomical order.

Hierarchical

The operationalised learning goal is made explicit in terms of sub-constructs, criteria, and signs of learning that clearly organized in hierarchical or taxonomical order

Specific

The operationalised learning goal is made explicit in great detail with a heavy translation of the competence into precise sub-constructs that can inform teaching and assessment – e.g. involving exhaustive learning progression schemas.

Fig. 11.2 A diagram of two possible dimensions of operationalising competences

e.g. by detailing potential signs of learning vis-à-vis multifarious sub-competences to be used in a single lesson. Alternatively, at the other end of the continuum, the operationalisation can result in a more general explication of the competence, e.g. by stipulating one or a few general signs of learning that can guide the assessment of the development of the competence over a year. Second, the outcome of a concrete act of operationalising a competence can be more or less *hierarchical.* At one end of this continuum, the sub-constructs and/or assessment guides that are formulated in order to make the competence operational can be structured in a hierarchical fashion, e.g. projecting potential learning trajectories in a learning progression format. At the other end of the continuum, the operationalisation can result in an array of sub-constructs and/or assessment guides that are not related or hierarchically ordered.

It is important to note that there will surely be more dimensions that are salient for the analysis of competence operationalisation. For example, teacher intentions seem to be an obvious candidate dimension. Further, we think that our model with the two dimensions should not be used normatively. Different contexts may call for different operationalisation strategies and operationalisation aims. Our primary aim with the two-dimensional model is to propose a terminological and analytical framework for analysing, and talking about, a key activity in education that we feel demands more explication.

The two-dimensional model of competence operationalisation hopefully has the potential to support future research vistas into classroom assessment research. One such area could be research into teacher professional development. As argued by Andrade (2012), there is a need for research that focusses on professional development vis-à-vis developing pedagogy based on learning progressions. The two-dimensional model may be used for both conducting professional development activities and analysing teachers' professional development in, e.g. action-research projects. Based on the findings from the preceding chapters, such professional development ought to be implemented over significant periods of time with ample possibility for teachers to negotiate meanings of learning goals together with educators and other teachers.

The two-dimensional model could also support efforts to meet the need for more knowledge about whether and how teachers design instruction on the basis of the cognitive constructs that are tested for in large-scale testing systems (for an argument for this need, see McMillan 2012). In general, it must be important for the field of classroom assessment research to study the efficacy of more organised operationalisation as compared to less organised operationalisations. Studies of this kind could become an important theme in the further investigation of the potency of pedagogy based on learning progressions that many scholars call for (see, e.g. Andrade 2012).

Schneider and Andrade (2013) argued that the questions of whether 'teachers have sufficient skill to analyse student work' and of how 'teachers use evidence of student learning to adapt instruction on the intended learning target' (p. 159) are among some of the key research questions for the future of classroom assessment research. The two-dimensional model proposed here could offer an interpretive framework for analysing observations, narratives and other data collected in order to elaborate on research questions such as these. Further, as Randel and Clark (2012) argued, there is a growing need for the development of instruments that can be used to measure teachers' assessment practices. The two-dimensional model may provide us with an outline for formulating items that pertain to the specificity and level of organisation of teachers' operationalisation of competences. In relation to this, the model may assist future research into teacher assessment preparation – a research focus that some argue needs to be systematically pursued (see, e.g. Campbell 2012).

Conclusion

In this chapter, we have pointed to several aspects that seem to cut across the multifarious research findings from the different contexts and studies involved in the ASSIST-ME project. In particular, we identified an underlying theme in the findings that pertains to how teachers interpret complex learning goals and make them more operational in order to be instructive for designing and implementing teaching and assessment activities. By beginning to talk about a process of operationalising learning goals, we hope that the fields of classroom assessment and science education will gain a more explicit nomenclature for approaching some of the perennial issues that emerge from studying teachers' assessment practice.

References

Andrade, H. L. (2012). Classroom assessment in the context of learning theory and research. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 17–34). Thousand Oaks: SAGE Publications.

Belova, N., Dittmar, J., Hansson, L., Hofstein, A., Nielsen, J. A., Sjöström, J., & Eilks, I. (in press). Cross-curricular goals and raising the relevance of science education. In K. Hahl, K. Juuti, J. Lampiselkä, J. Lavonen, & A. Uitto (Eds.), *Cognitive and affective aspects in science education research: Selected papers from the ESERA 2015 conference*. Rotterdam: Springer.

Bernholt, S., Rönnebeck, S., Ropohl, M., Köller, O., & Parchmann, I. (2013). Report on current state of the art in formative and summative assessment in IBE in STM-Part 1. *ASSIST-ME Report Series, 1*.

Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The SOLO taxonomy, Structure of the observed learning outcome*. New York: Academic.

Brookhart, S. M. (2012). Grading. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 257–272). Thousand Oaks: SAGE Publications.

Bungum, B. (2006). Transferring and transforming technology education: A study of Norwegian teachers' perceptions of ideas from design & technology. *International Journal of Technology and Design Education, 16*(1), 31–52.

Campbell, C. (2012). Research on teacher competency in classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 71–84). Thousand Oaks: SAGE Publications.

Chevallard, Y. (1991). *La transposition didactique – du savoir savant au savoir enseigné*. Grenoble: La Pensée Sauvage.

Cowie, B. (2012). Assessment in the science classroom: Priorities, practices. And prospects. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 473–488). Thousand Oaks: SAGE Publications.

Danish Ministry of Education. (2013). *Bekendtgørelse om uddannelsen til studentereksamen. Nr 776 af 26/06/2013* (Executive Order nr. 776 of 26/06/2013). Copenhagen: Danish Ministry of Education.

Desurmont, A., Forsthuber, B., & Oberheidt, S. (2008). *Levels of autonomy and responsibilities of teachers in Europe*. Eurydice. Available from: EU Bookshop.

Dolin, J. (2012). *ASSIST-ME project proposal*. Copenhagen.

Dolin, J. (2016). Idealer og realiteter i målorienteret undervisning. [Ideals and realities in goal-oriented teaching]. *Cursiv, 19*(1), 67–87.

Drost, E. A. (2011). Validity and reliability in social science research. *Education Research and Perspectives, 38*(1), 105.

Dysthe, O., Engelsen, K. S., Madsen, T., & Wittek, L. (2008). A theory-based discussion of assessment criteria – The balance between explicitness and negotiation. In A. Havnes & L. McDowell (Eds.), *Balancing dilemmas in assessment and learning in contemporary education* (pp. 121–131). New York: Routledge.

Kattmann, U., Duit, R., Gropengiesser, H., & Komorek, M. (1996). Educational reconstruction – Bringing together issues of scientific clarification and students' conceptions. *NARST, 1996*, 22.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory Into Practice, 41*(4), 212–218.

McMillan, J. H. (2012). Why we need research on classroom assessment. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 3–16). Thousand Oaks: SAGE Publications.

Moss, P. A. (2008). Sociocultural implications for assessment I: Classroom assessment. In P. A. Moss, D. C. Pullin, J. P. Gee, E. H. Haertel, & L. J. Young (Eds.), *Assessment, equity, and opportunity to learn* (pp. 222–258). Cambridge: Cambridge University Press.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London: SAGE Publications.

Nielsen, J. A. (2015). Assessment of innovation competency: A thematic analysis of upper secondary school teachers' talk. *The Journal of Educational Research, 108*(4), 318–330. doi:10.1080/00220671.2014.886178.

Nielsen, J. A., & Dolin, J. (2016). Evaluering mellem mestring og præstation. [Assessment between mastery and performance]. *MONA, 2016*(1), 51–62.

Randel, B., & Clark, T. (2012). Measuring classroom assessment practices. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 145–164). Thousand Oaks: SAGE Publications.

Rönnebeck, S., Bernholt, S., & Ropohl, M. (2016). Searching for a common ground –A literature review of empirical research on scientific inquiry activities. *Studies in Science Education, 52*(2), 161–197.

Ruiz-Primo, M., & Li, M. (2012). Examining formative feedback in the classroom context: New research perspectives. In J. H. McMillan (Ed.), *SAGE handbook of research on classroom assessment* (pp. 215–232). Thousand Oaks: SAGE publications.

Schneider, M. C., & Andrade, H. (2013). Teachers' and administrators' use of evidence of student learning to take action: Conclusions drawn from a special issue on formative assessment. *Applied Measurement in Education, 26*(3), 159–162. doi:10.1080/08957347.2013.793189.

Sjöström, J., & Eilks, I. (2017). Reconsidering different visions of scientific literacy and science education based on the concept of Bildung. In Y. J. Dori, Z. Mevarech, & D. Bake (Eds.), *Cognition, metacognition, and culture in STEM education*. Dordrecht: Springer.

Star, S. L., & Griesemer, J. R. (1989). Institutional ecology, 'Translations' and boundary objects: Amateurs and professionals in Berkeley's Museum of Vertebrate Zoology, 1907-39. *Social Studies of Science, 19*(3), 387–420. doi:10.1177/030631289019003001.

Torrance, H. (2007). Assessment as learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. 1. *Assessment in Education: Principles, Policy & Practice, 14*(3), 281–294. doi:10.1080/09695940701591867.

Wiek, A., Bernstein, M., Foley, R., Cohen, M., Forrest, N., Kuzdas, C., … Keeler, L. (2015). Operationalising competencies in higher education for sustainable development. In M. Barth, G. Michelsen, M. Rieckmann, & I. Thomas (Eds.), *Handbook of higher education for sustainable development. Routledge* (pp. 241–260). London: Routledge.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching, 46*(6), 716–730.

Winsløw, C. (2011). Anthropological theory of didactic phenomena: Some examples and principles of its use in the study of mathematics education. In M. Bosch (Ed.), *Un panorama de la TAD. An overview of ATD*, *CRM documents* (Vol. 10, pp. 117–138). Barcelona: Centre de Recerca Matemàtica.

Index

J. Dolin, R. Evans (eds.), *Transforming Assessment*, Contributions from Science Education Research 4, DOI 10.1007/978-3-319-63248-3