Chapter 16

Assessment in Collaborative Learning

Jan van Aalst, The University of Hong Kong

Van Aalst, J. (2013). Assessment in collaborative learning. In C. E. Hmelo-Silver, C. A. Chin, C. K., K. Chan, & A. O'Donnell (Eds.), *The International handbook of collaborative learning* (pp. 280-296). New York: Routledge.

Introduction

Learning and assessment are mutually dependent because both students and teachers tend to pay greater attention to learning objectives that are assessed (Biggs, 1996; Shepard, 2000). This relationship has profound implications for the large-scale uptake of *collaborative learning*, which is defined for the purpose of this chapter as any educational approach in which students work toward a shared learning goal. Examples include learning in small groups, learning from online discussions, and learning in communities, which are discussed in other chapters in this handbook.

This chapter considers four issues with assessment in collaborative learning:

1. If assessment is based on a *group product*, then it is difficult, if not impossible, to ascertain what individual students have learned. Grades are often based on participation in the group process, but such participation is also difficult to ascertain, and often confuses learning with effort. This kind of assessment is frequently regarded as unfair by students, parents, and other stakeholders.

2. If students are assessed *individually* after learning in a small group, then what they know is measured correctly, but is attributed incorrectly to their personal achievement; at least some of their learning is a shared accomplishment. A well-functioning collaborative group can solve more difficult problems than any single student in the group. Stahl's (2010) theory of group cognition refers to learning effects that are *irreducible* to individual learning effects.

3. Assessment practices treat collaboration as a *method* for accomplishing learning, but it can be argued that it should be seen as a *human capability* worth assessing in its own right. Collaboration distributes the learning process over students (e.g., via other-regulation and shared reflections), and there is a potentially powerful role for assessment in the development of such practices. The development of competence has been identified as an important 21st century skill (Assessment of 21st Century Skills [ACT21] Project; see www.act21.org).

4. Situations in which collaborative learning is *most necessary*, in the sense that it would be impossible to achieve the learning goals without the cognitive benefits of collaboration referenced in Issue 2, all involve novelty, problem solving, and creativity. In these situations, there are qualitative differences in the outcomes generated by different teams, rendering objective and reliable assessment difficult.

Unless practical solutions to these issues are found, the widespread use of collaboration in formal education seems unlikely. This chapter presents a review of the research on assessment in collaborative learning, primarily of the cognitively oriented studies published between 1994 and 2009, to examine the foregoing issues. The next section places this review in the context of the historical trends in research

on assessment in general, and the subsequent sections discuss the major themes that emerge from the literature review: formative assessment, assessment of learning in small groups, assessment of online learning, and peer- and self-assessment. These themes draw from research in K-12 education, nursing and medical education, engineering education, and teacher education. The final section returns to the four aforementioned assessment issues and outlines opportunities for further research.

Historical Context

In the first half of the 20th century, assessment was greatly influenced by intelligence testing and psychometric test theory. As Gipps (1994) explained, intelligence testing assumes that aptitude is a stable, universal, and one-dimensional construct. Glaser (1963, cited in Gipps [1994]) noted a corresponding preoccupation with aptitude, selection, and prediction, and proposed criterion-referenced testing as an alternative to norm-referenced testing. Criterion-referenced tests focus on whether a student reaches a given standard regardless of how many other students do so.

Subsequent decades saw a trend away from norm-referenced government examinations toward school- and criterion-based assessment. In the 1970s, many Canadian and U.S. jurisdictions abolished government examinations in favor of teacher-designed tests. In the Netherlands, a school examination was introduced in addition to the government examination, and a Teacher Assessment Scheme for science education was introduced in Hong Kong. Since the mid-1980s, there has been a major effort to develop national benchmarks and standards in a variety of school subjects (AAAS, 1993; NCTM, 2000; NRC, 1996). In recent assessment reforms in Hong Kong, the government has placed greater emphasis on school-based assessment to assess practical skills in science and conversational skills in English (CDC/HKEAA, 2007), arguing that a student's own school and teacher provide a

better context for eliciting his or her best performance. An important aspect of the school-based assessment movement is the introduction of formative assessment (Scriven, 1967)—assessment that is carried out while learning is still occurring and that is used to *improve* the learning process. Such assessment utilizes formative feedback (Ramaprasad, 1983), which is based on the gap between a student's current performance and the desired standard.

Effort has also been devoted to providing a conceptual foundation for school-based assessment. Referring to the "one-dimensional" and "universal" nature of aptitude implied by norm-referenced testing, Gipps (1994) proposed a comprehensive framework for *educational* assessment that recognizes that domains of knowledge are complex, emphasizes standards, encourages students to think rather than regurgitate facts, elicits students' best performance, and involves grading by the teacher, possibly subject to moderation. Shepard (2000, p. 7) proposed a framework that aligns a contemporary view of curricula (e.g., addressing challenging content and higher-order thinking, establishing an authentic relationship between in-school and out-of-school learning, and fostering important habits of mind), cognitive and constructivist learning theories, and school-based assessment.

The foregoing developments, sketched in brief here, indicate progress toward a view of learning that acknowledges its complexity, attempts to specify the most important dimensions of learning via educational standards, and increased emphasis on school-based assessment. However, this general literature pays little attention to collective outcomes (Issue 2) or collaboration as a human competence (Issue 3); it is almost exclusively concerned with the learning outcomes of individual students.

Themes in Research on Assessment

This section describes the main themes emerging from a review of the literature on assessment relevant to collaborative learning: formative assessment, assessment of learning in small groups, assessment of online discussions, and self- and peer-assessment. Section 24.4 then connects the main findings of the research on these themes to the four issues outlined in Section 24.1.

Formative assessment

Formative assessment, students' use of feedback to improve learning while the learning process is ongoing, provides a potentially powerful resource for collaborative learning. For example, having received feedback from the teacher, students can help one another to understand that feedback and then devise and monitor a plan together for making use of it to improve their learning performance. Such collaborative interactions can lead to improvements in understanding of the nature and standard of performance desired. If a group of students receives feedback on a project from their teacher, then collaboration is necessary for developing a shared understanding of that feedback and making use of it to improve the project. Students can also peer-assess one another's performance rather than rely on the teacher to provide feedback, and can then help one another to understand how the performances can be improved. This kind of collaborative interaction can be a precursor to a more self-directed approach to learning that involves self-assessment. This section discusses the literature on formative assessment published since the well-known review of Black and Wiliam (1998), with the aims of determining the extent to which these practices are already occurring and identifying the theoretical and pedagogical development that is still needed.

Black and Wiliam (1998) discussed learning effects, existing practice, the role and nature of feedback, students' reception to feedback, and systemic considerations such

as the influence of external and summative assessment systems. They reviewed 250 papers published between 1988 and 1998, including approximately 20 rigorous quantitative studies that involved comparisons of learning effects. Their major finding was that formative assessment has a consistently positive impact on student learning outcomes across educational settings ranging from the early elementary school years to undergraduate study at university. However, they also found classroom assessment practices to be generally underdeveloped, finding few examples of formative assessment initiated by students or approaches that involve collaboration; the role of the student was limited primarily to receiving feedback from the teacher and acting on it individually.

Since Black and Wiliam's (1998) review, significant effort has been invested in the development of classroom formative practices. For example, Black, Harrison, Lee, Marshall, and Wiliam (2003) collaborated with teachers of science and mathematics at six secondary schools to enhance teachers' questioning, feedback by marking, peer- and self-assessment, and formative uses of summative tests, and reported that questioning became more focused on student thinking over time, and feedback more specific. Ruiz-Primo and Furtak (2007) proposed a modification to the IRE model of classroom discourse (teacher Initiates, student Responds, and teacher Evaluates) by adding a fourth step, in which the teacher uses students' responses to modify his or her teaching plan. However, in most approaches, classroom discourses appear to have remained dialectical, with the teacher controlling most of the talk. Yorke (2003) questioned whether, if a student employs feedback on a draft from the teacher to improve the final version of the assignment, then he or she can also perform adequately when such feedback is *not* available. Nevertheless, in a society in which collaboration is pervasive, it is usually possible to find someone who can provide

feedback on a work-in-progress. From this perspective, one of the competencies that students should be developing is the ability to seek and make use of feedback from peers.

A number of theoretical points have also been raised about formative assessment. Taras (2009) argued that the dichotomy between formative and summative assessment that has emerged in the literature was not intended by Scriven (1967), and called for better integration of the two concepts. She explained that the concept of formative feedback proposed by Ramaprasad (1983) involves information about the gap between actual performance and a reference level, and therefore has a summative aspect. Assessment, whether its function is formative or summative, involves a judgment about quality relative to some criterion. Yorke (2003) pointed out the need for a theory of formative assessment that aligns it with a constructivist theory of learning, takes into account epistemological models that are relevant to the subject that students are studying, provides cognitive models for learning from feedback, and takes into account such factors as readiness to learn and the impact of feedback on student characteristics such as motivation and self-esteem. However, in the model he proposed (Yorke, 2003, p. 487), the assessor sets the assessment task and grading criteria, which are then modified on the basis of student performance. Students' interpretation of feedback from the assessor influences their long-term development (i.e., performance on the next task), but not their performance on the task at hand. In other words, Yorke's model does not make use of feedback to enhance learning while it is in progress, limits the role of students as the agents of their learning, and is not collaborative. Perrenoud (1998) pointed out that Black and Wiliam (1998) missed an important body of literature published in French, which develops the *regulation of learning* as a central concept integrating formative assessment, the didactic content of

the disciplines in question, and differentiation in teaching. This French literature has strong foundations in a cognitive theory of learning that involves scaffolding; however, Perrenoud's (1998) concern was with regulation by the *teacher*, not by the students, and did not include collaboration.

More recently, Black and Wiliam (2009) proposed a theory of formative assessment that partly responds to the foregoing criticisms and developments, and attempts to incorporate self-regulated learning and limited collaboration. Here, they began from Ramaprasad's (1983) three key processes of teaching and learning (establishing where students are in their learning, where they need to be going, and what needs to be done to get them there), and pointed out that although the teacher has been primarily responsible for all three, students also have a role to play. Black and Wiliam's (2009) theory refers to five strategies for formative assessment: (a) clarifying and sharing learning intentions and criteria for success, (b) engineering effective classroom discussions and other learning tasks that elicit evidence of student understanding, (c) providing feedback that moves students forward, (d) activating students as instructional resources for one another (a collaborative strategy), and (e) activating students as the owners of their own learning. Following a suggestion by Hattie and Timperley (2007), Black and Wiliam conceptualized feedback at three levels: task, processes needed to understanding the task, and self-regulation (self-monitoring, directing, and regulating actions).

In summary, formative assessment has received considerable attention in the review period, but most of this research focuses on a teacher-directed activity in which students make use of feedback from the teacher. Nevertheless, the theory recently proposed by Black and Wiliam (2009) pays more attention to empowering

students as the agents of their own learning through self-regulated learning and includes collaboration among students in making use of the teacher's feedback.

Assessment of learning in small groups

Learning in small groups is the most prevalent form of collaborative learning in formal education. It occurs in science learning laboratories, in which available resources such as equipment and physical space do not allow students to carry out experiments individually, and in project-based learning, in which it is infeasible for the teacher to guide and provide feedback on projects by individual students. In these situations, the use of small group arrangements is often motivated by practical constraints rather than the potential cognitive benefits of collaboration, and, in the formation of these groups, teachers do not always pay adequate attention to findings from social psychology and the socio-cognitive dynamics of learning in small groups. As a result, learning in small groups is often fraught with inequities (Issue 1). Individual learner variables such as prior knowledge, motivation, interest in the task, and social skills all influence group performance, and it is quite common for high-performing students to learn less individually in groups than they would have done solo, even though the group as a whole benefits from having such students as collaborators. This section discusses several important studies that have investigated these phenomena.

Webb, Nemer, Chizhik, and Sugrue (1998) investigated the impact of group composition on group performance, the quality of group discussions, and individual achievement in a study of 445 seventh and eighth grade students drawn from 21 classes studying three-week instructional units on electricity and electric circuits. The students came from a range of socio-economic backgrounds. Prior to teaching, students completed three pre-tests measuring vocabulary, verbal reasoning, and non-

verbal reasoning. After teaching, they completed two tests individually: a practical test in which they were required to assemble simple electric circuits, draw diagrams of their circuits, and answer questions about them; and a paper-and-pencil test involving similar circuits. A month later, 80% of the students repeated the practical test collaboratively in triads, and the remaining 20% completed them individually as a control; all of the students repeated the paper-and-pencil test. Group interactions were videotaped and coded for collaborative assessment. Multilevel covariate analysis was then used to analyze the results, with students nested within groups (or as individuals) and groups within types of group composition.

This study produced a number of important findings. Regression analysis of three group composition variables (the highest, lowest, and average ability levels in each group) on achievement showed that (a) the *highest ability level* in the group was the only significant predictor of achievement for students in the bottom three quarters of the sample, whereas (b) the *lowest ability level* in the group was the only significant predictor of achievement for those in the top quarter. Thus, most students would benefit from working in a group with a high-ability student. Further analysis showed this effect to be especially strong for students in the bottom two quarters: working with high-ability students enhanced the performance of these students on both the group test and the individual test, and the video analysis showed that they had learned from hearing more high-quality explanations than other students. In contrast, high-ability students performed better when they worked in homogeneous groups than in heterogeneous groups. However, their performance did not suffer when they worked with low-ability students, relative to their performance when they worked with students of medium- to high-ability (i.e., the third quarter). Webb et al. (1998) pointed out that their findings raised important questions about the *fairness* of group

assessment: “If the purpose of collaboration on an assessment is to measure students’ performance after they have an opportunity to learn from others, then, to give all students the same advantage, all groups must have a high-achieving student” (p. 643). However, the results for high-achieving students indicated that this would not be the optimal configuration for them.

To replicate this study and clarify the impact of group composition on the achievement of high-ability students, Webb, Nemer, and Zuniga (2002) employed similar methods to examine the co-construction of task solutions, helping behavior such as responses to questions and corrections of one another’s statements, and the socio-emotional processes in operation when students work in groups. The authors drew three main conclusions: (1) high-ability students perform well in homogeneous groups, as well as in some, but not all, heterogeneous groups; (2) the types of group interaction that occur during group work strongly influence performance; and (3) *group interaction* predicts performance more strongly than either student ability or the overall ability composition of the group. The average level of help that the high-ability students both gave and received in this study (Webb, Nemer, and Zuniga, 2002) was significantly related to these students’ delayed post-test scores, and how frequently high-ability students heard other students verbalize fully correct answers was positively correlated with their delayed post-test performance. Negative socio-emotional behavior such as domineering, insulting, and off-task behavior predicted the frequency with which high-ability students handed in work less complete than their immediate posttest answers at the delayed post-test. The researchers called for strategies that *maximize group functioning* to activate the intellectual resources of all groups.

In another study focusing on high-ability students, Barron (2003) investigated collaborative problem solving in 12 triads of sixth graders. She controlled for ability by including only students who scored above the 75th percentile on a national standardized achievement test on mathematics. The participants viewed a 15-minute video adventure from the *Adventures of Jasper Woodbury* series as a class, and they then solved a challenge related to the adventure in triads. Next, they completed two tests individually: a repeat of the initial challenge and a second, structurally equivalent challenge. The triads were divided into less successful and more successful, using a score of 50% for the triad's solution as the dividing line. Barron (2003) found that the more successful triads accepted or discussed correct proposals more often and produced a higher proportion of proposals that were directly related to the challenge than the less successful groups. She presented four case studies—designated as follows—to illustrate archetypes of interaction that rendered the triads less or more successful in the co-construction of solutions. In *Competing to know* (less successful), the participants were unable to co-regulate problem solving; instead, they competed with one another to have their solutions heard. In *Two's company* and *Wait, listen, and watch*, the interactions primarily involved only two students; in the former case, the third student made three attempts to contribute to the solution construction, all of which were ignored by the other two, whereas, in the latter case, the third student was more assertive and insisted that his contribution be heard. Finally, the fourth (more successful) case, *Coordinated co-construction*, illustrated effective collaborative problem solving. Barron's (2003) study shows that even heterogeneous groups with students of high ability can be inequitable in terms of assessment (Issue 1) because groups differ in their ability to co-regulate the collaborative problem-solving process, a similar conclusion to that reached by Webb

et al. (2002). Barron therefore called for a shift from an instrumental view of collaboration to one that treats it as an important human capability in its own right (Issue 3).

Learning in groups is widespread in higher education, particularly in professional fields such as healthcare and engineering, in which the ability to work in teams is an important professional area of competence. In these contexts, concerns about fairness and the uneven contributions of group members are well known. To mention just one study, Colbeck, Campbell, and Bjorklund (2000) interviewed 65 undergraduate students from seven college campuses who had just completed, or were completing, engineering courses that involved a group design project. Many of the students mentioned their concerns about "slackers" and "leaders." Slackers did not contribute their fair share of effort to the project, and leaders frequently took up the leadership role because other group members did not keep their commitments or make quality contributions, often ending up doing most of the work themselves. However, high-performing students were also found to assume leadership roles because other students in the group were not performing to their standards. Students were also found to complain if the leader did not accept their ideas or let them carry out part of the project.

There is relatively little research on assessment practices themselves in the literature reviewed here, but one exception is in medical education, where teams are widely used in problem-based learning (PBL). Willis et al. (2002) investigated the assessment preferences of medical students who used PBL at the University of Manchester, and developed a rubric for assessing group interaction and activity. Their rubric emphasizes efficiency and engagement with the task: task-oriented inputs, the posing of questions that are salient to the task, and task completion. On the basis of a

1998 survey of registered nurses who graduated from a PBL program at McMasters University in Canada, Ladouceur et al. (2004) reported that assessment was "among the three worst aspects of the program" (p. 447). These authors suggested that a common challenge is finding a way to standardize assessments when there is a common curriculum. In this respect, one problem with collaborative learning is that although the curriculum implemented in multiple versions of a course or tutorial group may differ considerably, the intended curriculum remains the same. In professional fields, in which it is often important to certify that students have attained certain outcomes, objective and consistent assessment is regarded as being of paramount importance. Pauli, Mohiyeddini, Bray, Michie, and Street (2008) developed a questionnaire designed to measure individual differences in negative experiences of learning in groups, which can be used to examine the extent to which students feel such problems exist in their group work. This questionnaire has four subscales: lack of group commitment, task disorganization, storming group, and fractional group. A storming group is characterized by arguments, rows, and gossip, whereas a fractional group is characterized by feelings of isolation, the development of factions, and difficulties in deciding roles.

In summary, the literature discussed in this section reveals substantial problems in the assessment of learning in small groups. The major problem is that collaborative learning is rarely developed to the point that the potential cognitive benefits of collaboration are realized in all groups (Issues 1 and 3). The reasons for this under-development include variations in prior achievement, ability to coordinate contributions from multiple group members, interaction style, and motivation and interest in the task. There is also relatively little research on the quantitative measurement of group processes, and there are concerns about objectivity and

reliability in courses with multiple tutorial sections, in which the nature of learning outcomes can vary substantially (Issue 4).

Assessment of learning in asynchronous online discussion forums

The use of asynchronous discussion forums has become widespread, and provides an opportunity for collaborative learning. Online discussions are employed by students in secondary and postsecondary education to prepare for or extend in-class learning (e.g., Guzdial & Turns, 2000; Hsi & Hoadley, 1997), and it has been noted frequently that online discussions provide greater opportunities for students to share ideas and information than classroom discussions. Many instructors look to online discussions to promote critical thinking, argumentation, and knowledge construction. However, the majority of studies on online discussions reveal low levels of participation and interactivity (Guzdial & Turns, 2000; Hewitt, 2005), and there is little consensus on how such discussions should be assessed. This section reviews research on participation rates, portfolio assessments, and the emerging uses of server-log data.

Instructors often treat online discussions as an *addition* to classroom learning, something incidental to it, and they are therefore not generally assessed; deep integration between online discussions and classroom learning is rare. When online discussions *are* assessed, such assessment is often limited to measures of participation, such as the number of notes that are created and read by individual students. Furthermore, although some studies suggest a positive correlation between participation rates and measures of conceptual knowledge (e.g., Lee, Chan, & van Aalst, 2006) the research is mixed, and a model of how participation in online discussions contributes to learning is lacking. The research literature nevertheless suggests that the nature of the discourse—e.g., interaction and focus on concepts

rather than facts—plays an important role in learning, in conjunction with quantitative indicators (Hakkarainen, 2003).

There also are conceptual difficulties with current uses of participation rates for assessment because they treat individual events such as note creation as independent of one another, but they clearly are related. For example, when a class is summarizing online what has been learned over a certain period of time, students who are late in submitting their contributions may find it unnecessary to contribute at all because the points they wished to make have already been made several times. As numerous authors have noted, a discursive act is always a response to an earlier act (Wells, 1999). Hence, treating note creations as independent events removes the *collaborative* aspect of learning in online discussions (Stahl, 2002). Although it is useful to examine whether all students in a class are, over time, creating and reading notes, posing questions, and generating ideas, it is important to assess online discourse at additional levels to understand what groups of students, and the class as a whole, are doing and accomplishing together.

A focus on individual learning outcomes dominates the research on the educational uses of online discussions, but there are more powerful models in which the students constitute a *community*. In these models, individual students work to achieve collective goals, and are appreciated for the unique contributions they make to this joint effort (Bielaczyc & Collins, 1999). A prominent example is the *knowledge-building community* model, in which the community's goals are focused on extending the frontier of knowledge, as the community understands it (see Chan, this volume; Scardamalia & Bereiter, 2006). In knowledge building, ideas are regarded as epistemic objects: after they are introduced into a public space, the community works to improve them. The discourse of knowledge building involves much more than

sharing ideas; it requires substantial interaction to improve ideas and extensive synthesis and rise-above to reach new levels of conceptualization (Scardamalia & Bereiter, 2006; van Aalst, 2009). Whereas most examples of online discussions span from only a few days to weeks, knowledge-building discourse can last for many weeks or even months, making the need for rise above and synthesis particularly important. Thus, the technology generally used to support knowledge building, Knowledge Forum®, is designed to support synthesis and rise-above (Chan, this volume). There is a particular need for assessment to scaffold the development of knowledge-building discourse. If assessment focuses on this kind of work, then it reduces the need for teachers to read and evaluate a large number of initial contributions and raises the standard of online work. Students not only need to contribute notes, but they must also demonstrate that their collective efforts lead to shared knowledge advances.

Along these lines, van Aalst and Chan (2007) employed electronic portfolio notes in Knowledge Forum®. Students were provided with principles that describe knowledge building, and were asked to assess the extent to which their class's work on Knowledge Forum fit these principles. Although students completed this task individually, the principles represented both individual and collective aspects of knowledge building. Interviews showed that the task helped twelfth grade and university students to understand better how they should contribute to online discourse to enhance knowledge building. Lee et al. (2006) employed this approach with ninth grade students, and found that portfolio scores made a significant contribution to conceptual understanding scores after controlling for measures of participation, depth of inquiry, and depth of explanation obtained from analysis of the Knowledge Forum® database content. However, although this approach can provide

students and teachers with useful information about the efficacy of their discourse, the gathering of evidence by students is highly labor-intensive. Thus, tools to simplify this aspect of the assessment process are needed. Many researchers are therefore attempting to develop computer-based assessment tools that can provide semi-automatic analyses of online discussions.

In this respect, the information stored in online discussion forums—e.g., note content, keyword use, vocabulary, participation, and interactivity—provides a large database that can, in principle, be analyzed by a computer. If such analysis provides useful feedback to teachers and students, then it could become a valuable resource for formative assessment. One kind of analysis that has been employed widely is social network analysis (de Laat, Lally, & Lipponen, 2007; Haythornthwaite, 2002), a set of techniques for analyzing aspects of the social structure of discourse, such as the centrality of certain participants, the emergence of sub-groups, and the extent to which all participants have co-participants who follow or use their contributions. These techniques can provide useful information to teachers about *interactivity*, but they lack information on *content*, such as the concepts that participants are using. Nevertheless, in computer science, substantial progress has been made in data mining and text mining techniques that open up new possibilities for formative assessment. For example, latent semantic indexing, which was initially developed to improve a web search engine, is a set of techniques designed to discover the semantic structure of a large corpus of texts (Foltz, 1997). It has been successfully applied to such assessments as the machine grading of essays (Landauer, Laham, & Foltz, 2003). Chen and Chen (2009) provided an initial framework for the application of data mining techniques in general to formative assessment. They used five computational schemes—correlational analysis, gray relational analysis, k-means clustering, fuzzy

clustering algorithms, and fuzzy association—and argued that the results obtained from such techniques can be used for precise formative assessments based on the learning portfolios of individual learners collected from a web-based learning system. The teacher-side formative assessment tool in their study is able to provide information on individual students, including their degree of concentration, question and answer responses, and comments.

In sum, both the educational uses and assessment of asynchronous online discussions require substantial development; online discussions often lack sustained participation and interaction and are dominated by sharing practices. Whereas equity issues are important in small-group learning, here the conceptual difficulties arising from an emphasis on learning by individual learners is more important (Issue 2). There is a mismatch between the learning models that underpin the use of online discussions and most assessments of these discussions (Chan & van Aalst, 2004). The formative use of assessments based on the content of online discussions can be an important resource for improving the nature and quality of these discussions, but work in this arena is still in an early stage of development (Issue 3).

Peer- and self-assessment

If collaboration is oriented toward improving learning outcomes, then it must involve the use of criteria for commenting on and raising questions about a collaborator's work. Peer-assessment is one structured approach to this kind of collaboration. Indeed, peer-assessment can be a method for learning how to collaborate. It can also serve as a first step toward self-assessment, as it is often more difficult to see the limitations in one's own work than those in that of others. This section briefly discusses peer- and self-assessment as the fourth theme in research on assessment.

The literature search, which emphasized cognitively oriented research, revealed relatively few studies in this arena, but it is clear that they these studies are just the tip of the an iceberg. Much of the impetus for peer- and self-assessment comes from the literature emphasizing critical thinking and authentic assessment tasks, in which students have a role in defining criteria and standards and then apply these criteria/standards to peer- and self-assessments of performance. Presumably, this kind of involvement leads to deeper student understanding of the expected learning outcomes. The literature reveals mixed evidence on whether students are capable of peer- and self-assessment, and raises a number of questions about the validity and reliability of such assessments (Issue 4).

White and Frederiksen (1998) found that reflective assessments carried out by middle-school students had a positive impact on their performance on a science inquiry test and physics test and that the process was particularly beneficial to low-achieving students. In a study by Ross and Starling (2008), secondary school students were involved in setting criteria, were taught to use those criteria, were given feedback on their self-assessments, and used assessment data to develop action plans. These researchers found self-assessment to be a valid and reliable method of assessing student performance, particularly when it is used for formative rather than summative purposes. In the medical education arena, Dannefer et al. (2005) also found peer assessment to be a reliable and valid method of assessing both cognitive and interpersonal aspects of medical students' performance. Tiwari and Tang (2003) examined portfolio assessment in healthcare education, and found that students enjoyed preparing their portfolios and that doing so improved their cognitive outcomes and affect toward their course of study. One of the things that these adult students appreciated was the ability to choose material for their portfolios to

demonstrate their learning. However, in their study of peer assessment in a computer-supported collaborative learning (CSCL) environment, Prins, Sluijsmans, Kirschner, and Strijbos (2005) raised concerns about the completion rates of these assessments and the quality of the assessment performance. Davis, Kumtepe, and Aydeniz (2007) concluded from their review of numerous studies in this arena that although peer assessment is conducive to the improvement of learning, its validity and reliability are open to question, mainly because students lack sufficient subject knowledge.

These studies suggest that self- and peer-assessment may be more suitable for formative than summative assessment purposes, although further research is necessary to determine the impact of incorrect formative assessments on eventual performance. As mentioned earlier, Webb et al. (2002) found that the eventual performance (i.e., delayed post-test results) of high-ability students was negatively influenced by the earlier receipt of partially incorrect or incomplete explanations. Nevertheless, the "authenticity" of assessments that students have co-developed with the teacher may make students more ready to learn from assessment results—an important key to formative assessment.

Conclusions and Implications

This chapter has reviewed collaborative learning-related, cognitively oriented research on assessment published between 1994 and 2009. The beginning of this period coincides with the publication of Gipps' (1994) *Beyond Testing: Towards a Theory of Educational Assessment*, which argues for a shift from psychometric views of assessment to views that honor the multi-dimesnional and context-dependent nature of educational performance. All of the studies discussed herein are relevant to an understanding of the intersection of collaborative learning and assessment, although they do not all deal with both topics. The assessment strategies that have

been discussed include formative assessment involving individual students; assessment of individual students' science knowledge and inquiry and problem-solving abilities after working in small groups; quantitative measures of individual students' participation in online discussions; assessment of class knowledge building using portfolios created by individual students; and a variety of self- and peer-assessments focusing on learning by individual students after they have worked in small groups. This section elaborates upon why—despite the substantial difficulties in assessment—collaborative learning is a necessity, and then returns to the four issues highlighted in the Introduction to ouline a research agenda.

The necessity of collaborative learning

The author contends that in a 21st century educational worldview, collaborative learning is no longer an *instructional choice* but a necessity. Learning and working in teams, for example, software design teams, executive teams, and restaurant staff, is pervasive in the world of work. In work situations, collective goals and achievements are the *raison d'être* for collaborative teams, but individual contributions and achievements are also important. The world of work requires that individuals are able to work with others in a variety of situations, and the need for work-place collaboration has increased substantially in recent decades. Technological developments throughout the 20th century led to phenomenal increases in access to travel, information, and communications, and rendered the world more globally competitive and faster-paced. The need for up-to-date knowledge that goes beyond the boundaries of what is known and "disruptive innovation" (Hagel III & Seely Brown, 2005) has increased dramatically. Knowledge work is significantly more difficult than traditional learning, and relies on teamwork. Equally important, the new work-related conditions are mirrorred by the experience of students, who also have

unprecedented access to information and social networks (e.g., Google and other Internet search engines, Facebook, and Twitter) and now use them in their everyday lives. However, although students can make use of online communities for learning (Gee, 2007), their information and digital literacy and inquiry and collaboration skills still require substantial development. Thus, educational priorities must be altered to provide a better balance between domain knowledge and development of the capacity for new learning. Collaboration is necessary because it can lead to better results than individual efforts and because it can render development of the capacity for new learning more feasible.

Issues 1 and 2: Assessment of learning in small groups

Collaborative learning strategies are frequently employed for practical reasons, for example, because they are effective, involve mutual engagement with learning goals, distribute effort, and lead to collective learning outcomes that surpass what students can accomplish solo. However, many difficulties arise from individual differences in pior knowledge, motivation, interest in the task, effort, and ability to coordinate the contributions of multiple students (e.g., Barron, 2003; Colbeck et al., 2000; Webb et al., 1998), rendering the assessment of collaborative learning in small groups inequitable. Some groups clearly learn more than others for reasons that have little to do with the degree of effort invested in the learning process. Webb and colleagues (1998, 2002) suggested that each group must have at least one high-achieving student.

Another strategy for addressing these problems is to have students work with many different collaborative partners over time, to ensure that they are not always part of an advantaged or disadvantaged team and that they have opportunities to learn to work with students who vary in such variables as prior knowledge and interaction style. Learning communities may provide a better configuration than fixed small groups

because they place more emphasis on collective goals, articulated in part by the commnity members, and allow for expertise to be distributed to a greater extent than is possible in small groups (Bielaczyc & Collins, 1999). Thus, more students can benefit from the various competencies that different students bring to the community's learning. In a study of three successive knowledge-building communities in Grade 4 classes taught by the same teacher, Zhang, Scardamalia, Reeve, and Messina (2009) found that the diffusion of ideas was greatest in the "opportunistic group" configuration, in which students formed collaborative groups to deal with emerging inquiry questions and, over time, worked in many different groups, compared with the "fixed group" and "interacting group" configurations, in which they worked with the same students throughout the school year. Strategies for diffusing learning across small groups, such as gallery walks and presentations, can also be helpful (Kolodner et al., 2003). Another important variable is the *nature of the task*; a strong commitment to shared goals is more likely to occur when all collaborating students grasp the educational value of the task in question. Almost all of the studies examined in this review involved relatively short-term collaborations in small groups. More research examining issues of equity and learning longitudinally and at different levels of analysis (individual, group, and community) would be useful.

There also are problems with the *validity* of assessments of collaborative learning, both in small groups and in general. Because, as Stahl (2010) argues, the major cognitive benefit of collaborative learning is that it leads to collective learning outcomes that are irreducible to individual learning outcomes, it seems important to report learning outcomes at both levels. Reporting a grade for the group-level outcome, such as the overall quality of the project, as well as that reflecting a

particular student's own knowledge relative to that of other group members, would provide a more complete picture of learning performance. Doing so could show, for example, that the group had developed a high-quality project, but that the particular student in question had learned less from the project than his or her peers, that is, less than the group-level grade would suggest. This kind of assessment approach would allow us to learn more about whether individual students are benefiting cognitively from collaboration. Current practice neglects one or the other of the two levels of learning performance: it measures individual learning outcomes but neglects group-level outcomes or measures group-level outcomes and incorrectly infers individual learning, often confusing it with effort. The more widespread uptake of collaborative learning, it seems, requires assessments that provide clear evidence of the benefits of collaboration.

Issue 3: Developing collaborative learning as a human competence

Following Barron (2003), it is proposed that collaborative learning is not merely a method for learning but a *human competence* that is difficult to achieve and requires effort to develop. As indicated earlier, it is also an important 21st century competence. Elsewhere, van Aalst (in progress) argues that the "collaborative" in collaborative learning should refer to an aspect of learning that, if well-developed, empowers students to achieve learning that would be impossible without collaboration. Learning how to learn collaboratively then becomes an aspect of learning how to learn, a skill emphasized in many recent curriculum reforms (e.g., CDC, 2000). Although collaboration is more difficult than solo learning in some respects (i.e., coordination, social skills), it can scaffold learning how to learn by distributing cognitively difficult processes such as regulation and reflection.

Formative assessment and self- and peer-assessment have important roles to play in this arena.

However, although the current research on formative assessment is helpful in clarifying the possibility of using assessment as a learning resource, it falls short in demonstrating how students can be empowered as agents of their own learning; formative assessment is largely done to students by teachers. By comparison, self- and peer-assessments are performed by students and provide better opportunities for them to reflect and understand what constitutes quality performance, particularly on complex tasks. Further theoretical and empirical research is required to align formative assessment and self- and peer-assessment with a theory of agency in collaborative learning. Although agency is important in theories of self-regulated learning (Winne & Hadwin, 1998), these theories are primarily concerned with individual learning.

Issue 4: Objectivity and reliability in assessment

Objectivity and reliability are crucial to high-stakes assessments such as matriculation examinations and international evaluations of educational progress, which are beyond the scope of this chapter. However, as noted earlier, there has been substantial growth in the use of school-based assessments of complex performance, which sometimes involves collaboration (e.g., the ability to converse with others in a second language).

In some countries, teachers have considerable freedom in the conduct of school-based assessments, but, when external examinations also exist, these assessments often are designed to provide practice for the external examinations rather than an opportunity to assess more complex student performance than is amenable to large-

scale testing. Substantial work is necessary to establish a school-based assessment method that is suitable for assessing those types of performance that are unsuitable for large-scale testing. In many situations that require collaboration, the learning products may vary substantially between groups (for example, one group may solve a mathematics problem geometrically and another algebraically), and requiring solutions to be more uniform (e.g., only geometric) would undermine the goal of stimulating creativity. Standards-based approaches can be useful, but the dimensions of performance should be generative of future learning rather than merely descriptive of the content knowledge that is of immediate concern. In their work on portfolio assessment in knowledge building, van Aalst and Chan (2007) provided such dimensions (e.g., collaborative effort and progressive problem solving) and asked students to analyze the extent to which there was evidence of these dimensions in their class's Knowledge Forum® database.

Apart from these qualitative features of collaborative learning, research is also needed both to develop instruments that students and teachers can use to guide the improvement of collaborative practices and to determine *how* these instruments are used. As noted earlier, participation data such as the number of notes created by individual students are easily obtained, but they do not provide clear recommendations for improving online discussions. Advanced techniques such as social network analysis fail to address this problem. Of course, the way in which a class of students makes use of information depends on such contextual factors as students' previous experience and goals, but it would still be useful to know whether writing more notes is likely to lead to conceptual change or whether another strategy would be more beneficial. Willis et al. (2002) developed a rubric for assessing group interaction and activity among medical students engaged in PBL, but few studies

developing such instruments for use in other education fields were found in the literature review carried out for this chapter. Such development would thus constitute a fruitful direction for future research.

Acknowledgments

The preparation of this chapter was supported by a General Research Fund grant from the Research Grants Council of Hong Kong (Grant HKU 752508H). The author woud also like to thank Li Sha for useful discussions.

References

AAAS. (1993). *Benchmarks for science literacy*. New York, NY: Oxford University Press.

Barron, B. (2003). When smart groups fail. *The Journal of the Learning Sciences, 12*, 307-359.

Bielaczyc, K., & Collins, A. (1999). Learning communities in classrooms: A reconceptualization of educational practice. In C. M. Reigeluth (Ed.), *Instructional design theories and models, Vol II* (pp. 269-292). Mahwah, NJ: Lawrence Erlbaum Associates.

Biggs, J. (Ed.). (1996). *Testing: To educate or to select? Education in Hong Kong at the crossroads*. Hong Kong SAR, China: Hong Kong Educational Publishing Company.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: Putting it into practice*. New York, NY: Open University Press.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy, and Practice, 5*, 7-74.

Black, P., & Wiliam, D. (2009). Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability, 21*(1), 5-31.

CDC (2000). *Learning to learn: The way forward in curriculum development*. Hong Kong SAR, China: Author.

CDC/HKEAA. (2007). *Liberal studies curriculum and assessment guide (secondary 4-6)*. Hong Kong SAR, China: Curriculum Development Council and the Hong Kong Examinations and Assessment Authority.

Chan, C.K.K. (2013, this volume). Collaborative knowledge building: Towards a knowledge creation perspective. In C. E. Hmelo-Silver, C. A. Chin, C. K., K. Chan, & A. O'Donnell (Eds.), *The International handbook of collaborative learning* (pp. 437-461). New York: Routledge.

Chan, C. K. K., & van Aalst, J. (2004). Learning, assessment, and collaboration in computer-supported learning environments. In J. W. Strijbos, P. A. Kirschner & R. L. Martens (Eds.), *What we know about CSCL: And implementing it in higher education* (pp. 87-112). Dordrecht, the Netherlands: Kluwer Publishers.

Chen, C.-M., & Chen, M.-C. (2009). Mobile formative assessment tool based on data mining techniques for supporting web-based learning. *Computers & Education, 52*, 256-273.

Colbeck, C., Campbell, S., & Bjorklund, S. (2000). Grouping in the dark: What college students learn from group projects. *The Journal of Higher Education, 71*(1), 60-83.

Dannefer, E. F., Henson, L. C., Bierer, S. B., Grady-Weliky, T. A., Meldrum, S., Nofziger, A. C., . . . & Epstein, R. M. (2005). Peer assessment of professional competence. *Medical Education, 39*, 713-722.

Davis, N. T., Kumtepe, E. G., & Aydeniz, M. (2007). Fostering continuous improvement and learning through peer assessment: Part of an integral model of assessment. *Educational Assessment, 12*, 113-135.

de Laat, M., Lally, V., & Lipponen, L. (2007). Investigating patterns of interaction in networked learning and computer-supported collaborative learning: A role for social network analysis. *International Journal of Computer-Supported Collaborative Learning, 2*, 87-103.

Foltz, P. (1997). Latent semantic analysis for text-based research. *Behavior Research Methods, Instruments and Computers, 28*, 197-202.

Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. Abdingdon, UK: RoutledgeFalmer.

Guzdial, M., & Turns, J. (2000). Effective discussion through a computer-mediated anchored forum. *The Journal of the Learning Sciences, 9*, 437-469.

Hagel III, J., & Seely Brown, J. (2005). *The only sustainatble edge: Why business strategy depends on productive friction and dynamic specialization*. Boston, MA: Harvard Business School Press.

Hakkarainen, K. (2003). Emergence of progressive-inquiry culture in computer-supported collaborative learning. *Learning Environments Research, 6*, 199-220.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81-112.

Haythornthwaite, C. (2002). Building social networks via computer networks: Creating and sustaining distributed learning communities. In K. A. Renniger & W. Shumar (Eds.), *Building virtual communities: Learning and change in cyberspace* (pp. 159-190). New York, NY: Cambridge University Press.

Hewitt, J. (2005). Toward an understanding of how threads die in asynchronous computer conferences. *The Journal of the Learning Sciences, 14*, 567-589.

Hsi, S., & Hoadley, C. M. (1997). Productive discussion in science: Gender equity through electronic discourse. *Journal of Science Education and Technology, 6*, 23-36.

Kolodner, J. L., Camp, P. J., Crismond, D., Fasse, B., Gray, J., Holbrook, J., . . . & Ryan, M. (2003). Problem-based learning meets case-based reasoning in the middle-school science classroom: Putting Learning by Design into practice. *The Journal of the Learning Sciences, 12*, 495-547.

Ladouceur, M. G., Rideout, E. M., Black, M. E. A., Crooks, D. L., O'Mara, L. M., & Schmuck, M. L. (2004). Development of an instrument to assess individual student performance in small group tutorials. *Journal of Nursing Education, 43*, 447-455.

Landauer, T. K., Laham, D., & Foltz, P. (2003). Automatic essay assessment. *Assessment in Education: Principles, Policy & Practice, 10*, 295-308.

Lee, E. Y. C., Chan, C. K. K., & van Aalst, J. (2006). Students assessing their own collaborative knowlegde building. *International Journal of Computer-Supported Collaborative Learning, 1*, 277-307.

NCTM (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

NRC (1996). *National science education standards*. Washington, DC: National Academic Press.

Pauli, R., Mohiyeddini, C., Bray, D., Michie, F., & Street, B. (2008). Individual differences in negative group work experiences in collaborative student learning. *Educational Psychology, 28*, 47-58.

Perrenoud, P. (1998). From formative evaluation to a controlled regulation of learning processes. Towards a wider conceptual field. *Assessment in Education: Principles, Policy & Practice, 5*, 85-102.

Prins, F. J., Sluijsmans, D. M. A., Kirschner, P. A., & Strijbos, J.-W. (2005). Formative peer assessment in a CSCL environment: A case study. *Assessment and Evaluation in Higher Education, 30*, 417-444.

Ramaprasad, A. (1983). On the definition of feedback. *Behavioral Science, 28*, 4-13.

Ross, J. A., & Starling, M. (2008). Self-assessment in a technology-supported environment: The case of grade 9 geography. *Assessment in Education: Principles, Policy & Practice, 15*, 183-199.

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*, 57-84.

Scardamalia, M., & Bereiter, C. (2006). Knowledge building: Theory, pedagogy, and technology. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 97-115). New York, NY: Cambridge University Press.

Scriven, M. (1967). The methodology of evaluation. In R. Tyler, R. Gagne & M. Scriven (Eds.), *Perspectives on curriculum evaluation*. Chicago, IL: Rand McNally and Co.

Shepard, L. E. (2000). The role of assessment in a learning culture. *Educational Researcher, 29*(7), 1-14.

Stahl, G. (2002). Rediscovering CSCL. In T. Koschmann, R. Hall & N. Miyake (Eds.), *CSCL 2: Carrying forward the conversation* (pp. 169-181). Mahwah, NJ: Lawrence Erlbaum Associates.

Stahl, G. (2010). Group cognition as a foundation for the new science of learning. In M. S. Khine & I. M. Saleh (Eds.), *The new science of learning: Cognition, computers and collaboration in education* (pp. 23-44). Dordrecht, the Netherlands: Springer.

Taras, M. (2009). Summative assessment: The missing link for formative assessment. *Journal of Further and Higher Education, 33*, 57-69.

Tiwari, A., & Tang, C. (2003). From process to outcome: The effect of portfolio assessment on student learning. *Nurse Education Today, 23*(4), 269-277.

van Aalst, J. (2009). Distinguishing knowledge sharing, construction, and creation discourses. *International Journal of Computer-Supported Collaborative Learning, 4*, 259-288.

van Aalst, J., & Chan, C. K. K. (2007). Student-directed assessment of knowledge building using electronic portfolios. *The Journal of the Learning Sciences, 16*, 175-220.

Webb, N. M., Nemer, K. M., Chizhik, A. W., & Sugrue, B. (1998). Equity issues in collaborative group assessment: Group composition and performance. *American Educational Research Journal, 35*, 607-651.

Webb, N. M., Nemer, K. M., & Zuniga, S. (2002). Short circuits or superconductors? Effects of group composition on high-achieving students' science assessment performance. *American Educational Research Journal, 39*, 943-989.

Wells, G. (1999). *Dialogic inquiry: Toward a sociocultural practice and theory of education*. New York, NY: Cambridge University Press.

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

Willis, S. C., Jones, A., Bundy, C., Burdett, K., Whitehouse, C. R., & O'Neill, P. A. (2002). Small-group work and assessment in a PBL curriculum: A qualitative and quantitative evaluation of student perceptions of the process of working in small groups and its assessment. *Medical Teacher, 24*, 495-501.

Winne, P., & Hadwin, A. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277-304). Mahwah, NJ: Lawrence Erlbaum Associates.

Yorke, M. (2003). Formative assessment in higher education: Moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*, 477-501.

Zhang, J., Scardamalia, M., Reeve, R., & Messina, R. (2009). Designs for collective cognitive responsibility in knowledge-building communities. *The Journal of the Learning Sciences, 18*, 7-44.