Exploring Construct and Consequential Validity of Collaborative Learning Assessment in Higher Education

Hajo Meijer¹, Jasperina Brouwer¹, Rink Hoekstra¹, and Jan-Willem Strijbos¹

Abstract
Despite the advance of collaborative learning in higher education, there is a tension between its assessment and the orientation on students' individual domain-specific abilities (e.g., knowledge and/or skills) of most higher education curricula. We examined the construct and consequential validity of group assessment, individual assessment, and combined assessment of collaborative learning. Findings showed that the construct and consequential validity of these assessment methods can vary widely within and across cohorts. In view of these findings and considering pragmatic and didactic considerations, combined assessment of collaborative learning might be better suited than group assessment and individual assessment of collaborative learning.

Keywords
group grades, assessment, construct validity, consequential validity, higher education

¹University of Groningen, The Netherlands

Corresponding Author:
Hajo Meijer, Faculty of Behavioural and Social Sciences, Department of Education, University of Groningen, Grote Rozenstraat 3, Groningen 9712 TG, The Netherlands.
Email: hajomeijer@gmail.com
Assessment in higher education typically focusses on the construct of students’ individual domain-specific abilities (e.g., knowledge and/or skills). Assessment can have (a) a formative (or assessment for learning) purpose, in which assessment outcomes are used to aid students’ learning without it contributing to high-stakes decisions, or (b) a summative purpose (or assessment of learning), in which assessment is used to determine what students have learned with its assessment outcomes contributing to high-stakes decisions (Scriven, 1967; Sluijsmans & Segers, 2018). Summative assessment outcomes are accumulated into students’ individual transcript of records and subsequently used to determine whether students can be awarded with an individual degree (i.e., certification). Simultaneously, assessed outcomes as a result of student collaboration seemed to have increased in the last decades (Abernethy & Lett, 2005; Bacon, 2005; Bourner et al., 2001; De Hei et al., 2015; Flores et al., 2015; Lejk et al., 1997). These outcomes are often group assessments of products, such as reports, essays, or presentations, resulting from two or more students who collaborated (Dijkstra et al., 2016; Frykedal & Chiriac, 2011; Nicolay, 2002; Pitt, 2000). Although few would deny the value of collaborative learning in higher education, researchers, teachers, and students express concerns regarding its assessment, specifically about its construct validity and consequential validity (Ahern, 2007; Almond, 2009; De Hei, Sjoer et al., 2016; De Hei, Strijbos et al., 2016; Forsell et al., 2020; Kagan, 1995; Lejk et al., 1997; Meijer et al., 2020; Strijbos, 2011, 2016).

Construct validity refers to the extent to which a test measures a given unobservable psychological construct (Cronbach & Meehl, 1955), which Messick (1995) refers to as the content aspect of construct validity. Construct validity is typically explored by correlations between two tests that supposedly measure the same or a similar psychological construct (Panadero et al., 2013). The construct of interest that higher education assessment typically emphasizes, and therefore the construct under investigation in the current study, is students’ individual domain-specific abilities (e.g., knowledge and/or skills) (Boud et al., 1999; Plastow et al., 2010; Sharp, 2006). Since students differ in their individual domain-specific abilities and the effort they invest into collaboration, within and between group differences arise (Brookhart, 2013). These differences might lead to a misalignment between (a) the individual domain-specific abilities that are typically the construct of interest in higher education assessment and (b) the construct that is typically measured with assessment of collaborative learning (i.e., a group performance) (Meijer et al., 2020).

Consequential validity was introduced by Messick (1989) and refers to the actual and potential consequences of test use (cf. Reckase, 1998). Test outcomes typically result in decisions that carry intended and unintended positive and negative consequences (Kane, 2013; Slomp et al., 2014). An intended
positive consequence of an educational assessment test is, for example, a positive effect on the retention of knowledge compared to non-testing (Roediger & Karpicke, 2006), whereas an unintended negative consequence is, for example, adverse impact on minority groups (Shepard, 1993). According to Messick (1989), traditional forms of validity measures (i.e., construct, content, and criterion-related validity) ignore the consequences of test interpretation and test use. He argues that “we must inquire whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values” (p. 8). In addition, we agree with Slomp et al. (2014) that “those who design and use tests have an obligation to examine both the intended and unintended consequences that accrue as a result of their decision-making process and, where warranted, to remedy negative unintended consequences” (p. 279). Therefore, we regard consequential validity as an important aspect of validity, and accordingly, it is included in the current study. When assessment methods of collaborative learning measure a different construct than what is assumed to be the construct of interest in higher education—that is, students’ individual domain-specific abilities—this might lead to unintended negative consequences, such as students’ passing a course while not meeting the required domain-specific abilities, as specified by a curriculum (Plastow et al., 2010).

Construct and consequential validity of collaborative learning assessment are determined by multiple variables, such as group composition (e.g., similar or mixed ability groups) (Lejk et al., 1999), group dynamics (e.g., the process of collaboration) (Ohaja et al., 2013), and purpose of assessment (e.g., formative or summative) (Sridharan et al., 2019). In addition, the chosen method to assess collaborative learning influences the construct and consequential validity as well (Meijer et al., 2020). Assessment of collaborative learning can be roughly done using three methods: group assessment, individual assessment, and group assessment combined with individual assessment (henceforth combined assessment) (Forsell et al., 2020; Lejk et al., 1997; Strijbos, 2011, 2016). In the present study, the construct and consequential validity of these three assessment methods will be examined, under the assumption that the construct of interest in higher education assessment is typically students’ individual domain-specific abilities.

**Construct and Consequential Validity of Group Assessment of Collaborative Learning**

Group assessment of collaborative learning takes place when all students in a group receive the same assessment (e.g., group score, grade, or comments).
Since group assessment of collaborative learning eliminates variation in grading (Dijkstra et al., 2016; Nordberg, 2008), the average of students’ individual domain-specific abilities in a group turned out to be the best predictor of a group grade (Bacon et al., 1999). Across a variety of domains (e.g., science, marketing, occupational therapy) and school levels (e.g., primary, secondary, and higher education), it was found that group assessment of collaborative learning measures a different construct than merely students’ individual domain-specific abilities (Almond, 2009; Harvey et al., 2019; Lejk et al., 1999; Plastow et al., 2010; Saner et al., 1994; Webb, 1993).

Hence, when the construct of interest is student’s individual domain-specific abilities, which is typically the case in higher education assessment, group assessment’s construct validity might be compromised. For example, Plastow et al. (2010) reported low correlations of .09 and .15 between group assessment grades and individual assessment grades. In contrast, however, Lejk et al. (1999) found relatively high correlations of about .40. Furthermore, Lejk et al. (1999) found that the correlation between group assessment of collaborative learning and individual assessment of individual learning increased from about .40 to .65 when students worked in similar-ability groups (i.e., students were matched based on the average grade on two tests at the beginning of the course) as opposed to mixed-ability groups. Overall, existing evidence indicates group assessment of collaborative learning might compromise construct validity. Group assessment may not accurately reflect students’ individual domain-specific abilities, which is problematic given the current individual focus in higher education. Simultaneously, when combined, the studies by Lejk et al. (1999) and Plastow et al. (2010) illustrate that the degree of construct validity may vary widely.

In terms of consequential validity, Almond (2009) found that students with relatively higher individual grades received lower group grades (on average −9%) and students with relatively lower individual grades received higher group grades (on average +35%). Hence, group assessment of collaborative learning might lead to unintended negative consequences such as adverse impact on students with relatively higher individual grades, beneficial impact on students with relatively lower individual grades, and students passing/failing a course while individually (not) meeting the domain-specific abilities as specified by a curriculum (Plastow et al., 2010).

**Construct and Consequential Validity of Individual Assessment of Collaborative Learning**

Individual assessment of collaborative learning is when students in a group receive a personalized assessment (e.g., individual score, grade, or
comments) based on, for example, (a) the part they were responsible for within a group assignment or (b) an individual examination or assignment after students’ collaboration. Individual assessment of collaborative learning is expected to have a higher construct validity (i.e., more aligned with measuring individual domain-specific abilities as assumed by the higher education curricula) than group assessment of collaborative learning, since students have to perform the assessment task solely by themselves (Meijer et al., 2020). However, theoretically, it can be questioned whether the individual assessment of collaborative learning can be entirely detached from the collaborative setting prior to the assessment. This results in the assumption that individual assessment of collaborative learning can be at best qualified as a collaboration-moderated individual assessment of collaborative learning (Strijbos, 2011).

In terms of consequential validity, individual assessment of collaborative learning is expected to mitigate challenges faced by group assessment, including (a) advantaging students with relatively lower individual grades and disadvantaging students’ with relatively higher individual grades (Almond, 2009) and (b) students passing a course while individually not meeting the domain-specific abilities as specified by a curriculum (Plastow et al., 2010).

**Construct and Consequential Validity of Combined Assessment of Collaborative Learning**

Group or individual assessment of collaborative learning (e.g., scores, grades, or comments) can be exclusively used to determine students’ grades on collaborative learning. However, some teachers prefer to combine group assessment and individual assessment of collaborative learning because they assume that it ensures a more accurate grade (i.e., more aligned with measuring students’ individual domain-specific abilities as assumed by the higher education curricula) (Ahern, 2007; Augar et al., 2016). Combining individual and group assessment of collaborative learning can be done in at least two ways: (a) moderating a group assessment using intra-group peer-assessment of student’s contribution to the group process and/or group product (Forsell et al., 2020; Strijbos, 2011, 2016), and/or (b) adding an individual assessment to a group assessment (e.g., an individual task such as an examination that is unrelated to the group task or by specifying an individual part within a group assignment; Strijbos, 2011, 2016). Although combining individual and group assessment to assess collaborative learning is common practice in higher education (Ahern, 2007; Augar et al., 2016; Boud et al., 1999; Nordberg, 2008; Plastow et al., 2010), little is known about its construct
and consequential validity (Strijbos, 2011, 2016). This is in contrast to other combined assessment approaches, such as combining group assessment with intra-group peer-assessment (e.g., Bushell, 2006; Carson & Glaser, 2010; Carver & Stickley, 2012; Diprose et al., 1997; Jin, 2012; Ohaja et al., 2013; Sridharan et al., 2019; Willcoxson, 2006).

When combined assessment of collaborative learning consists of a separate individual and group task and/or an individual and group part of an assignment, the issue of weighting becomes important. Although we are not aware of common or recommended practices on how to weight the individual and group assessment in a combined assessment (Strijbos, 2011, 2016), the selected weighting can impact two core principles: (a) individual accountability and (b) positive interdependence in collaborative learning. Individual accountability refers to the degree to which individual students are held accountable for their contribution to collaborative learning (Slavin, 1980), whereas positive interdependence refers to the extent to which the performance of single group members depends on the performances of other group members (Johnson, 1981). Group assessment of collaborative learning fosters positive interdependence but not individual accountability, and the opposite holds for individual assessment of collaborative learning (Meijer et al., 2020). Combined assessment of collaborative learning has the potential to foster both positive interdependence and individual accountability.

The pivotal question with respect to weighting is how the individual assessment and group assessment could be weighted in such a way that it fosters both positive interdependence and individual accountability—or, at the very least, does not compromise either one or both. For example, if the group assessment contributes little to a combined grade (e.g., 15% group and 85% individual), positive interdependence is minimally fostered relative to individual accountability. Therefore, collaboration might be devaluated and students might give it little consideration (Strijbos, 2011, 2016). On the other hand, when the group assessment contributes heavily to a combined grade (e.g., 85% group and 15% individual), individual accountability is minimally fostered relative to positive interdependence, which might elicit free-riding and higher-able students taking over tasks of other students (Kagan, 1995; Meijer et al., 2020; Pitt, 2000; Strijbos, 2011, 2016). Weighting distributions that are reported in the literature vary widely, such as 30% group grade and 70% individual grade (Plastow et al., 2010), 10% group grade and 90% individual grade (Ahern, 2007), and 50% group grade and 50% individual grade (Nordberg, 2008).

Harvey et al. (2019) found that the construct validity of combined assessment (i.e., group 70%, individual 30%) of collaborative learning was higher
compared to group assessment of collaborative learning, but it was still off from the construct validity of individual assessment of collaborative learning. Likewise, Plastow et al. (2010) found that the more the individual assessment contributed to the combined assessment, the better it reflected student’s individual domain-specific abilities. In terms of consequential validity, it was found that the higher the share of the individual grade in the combined grade, the more the unintended consequences of group assessment of collaborative learning were mitigated, such as students passing/failing a course while individually (not) meeting the domain-specific abilities as specified by a curriculum (Plastow et al., 2010).

Current Study

The current study investigates the construct validity of group assessment, individual assessment, and combined assessment of collaborative learning, as well as the consequential validity of group assessment and combined assessment of collaborative learning in higher education. The construct under investigation is students’ individual domain-specific abilities (e.g., knowledge and/or skills), since this is the construct higher education assessment typically emphasizes (Boud et al., 1999; Plastow et al., 2010; Sharp, 2006). Hence, we examine to what extent individual domain-specific abilities of students are reflected in group assessment, individual assessment, and combined assessment of collaborative learning.

To our knowledge, thus far studies in higher education only examined the validity of collaborative learning assessment methods by comparing two different types of assessment tasks. For example, grades of (a) an individual written examination with (b) group assessment grades of a group assignment (e.g., Lejk et al., 1999), or (a) group assessment grades for the group product with (b) individual assessment grades for contributions to the collaborative process (e.g., Almond, 2009). To extend the current literature, we examine the validity of group assessment of collaborative learning (i.e., written assignment) not only by comparing it with an individual assessment on a different assessment task (i.e., written examination), but also with an individual assessment on a near identical task (i.e., written assignment) to make a more accurate comparison. That is, we compared the assessment of (a) the group part of a written assignment with the assessment of (b) the individual part within the same assignment. Regarding the comparison of the group assessment of the written assignment with the individual written examination; the content and question format we used in the individual examination were similar in terms of constructs measured in the assignment, thus providing sufficient ground for comparing the grade of the individual written examination with the three
types of assessments of collaborative learning. In addition, we examined the validity of combined assessment of collaborative learning, in which group assessment and individual assessment are combined (weighted 50/50). We could not examine the consequential validity of the individual assessment of collaborative learning, as the specific course and study design did not allow for meaningful comparison with an individual assessment on a near identical task. Our explorative research questions were:

1. What is the (a) construct validity and (b) consequential validity of group assessment of collaborative learning in a fourth and final-year undergraduate university course?
2. What is the construct validity of individual assessment of collaborative learning in a fourth and final-year undergraduate university course?
3. What is the (a) construct validity and (b) consequential validity of combined assessment of collaborative learning (individual and group assessment weighted 50/50) in a fourth and final-year undergraduate university course?
4. What is the (a) construct validity and (b) consequential validity of 19 potential compositions of combined assessment of collaborative learning—including the combination used in the current course—in a fourth and final-year undergraduate university course?

**Method**

**Design and Participants**

This exploratory study was conducted in a fourth and final-year undergraduate course of an academic teacher training program at a university in The Netherlands and comprised a course evaluation design (post-test only). We collected data from this course in two consecutive years: 2019/2020 (Cohort 1) and 2020/2021 (Cohort 2). The study design was approved by the ethics committee of the program’s department. Cohort 1 consisted of 35 students (34 female, 1 male) of whom 29 (83%, 28 females, 1 male) participated and gave their active informed consent to collect data (including their current course grades). In the end, one student did not partake in several assessed parts of the course and was excluded, resulting in a sample of 28 students. Their ages ranged from 21 to 26 ($M=22.17, SD=1.17$). The 28 participants were divided over 11 groups; ten groups consisted of three students and one group consisted of two students. Of the ten groups of three students; we got
data from seven complete groups (i.e., with all three students agreeing to share their scoring and grades) and three incomplete groups (i.e., with two out of three students agreeing to share their scoring and grades). Of the group of two students, one student agreed to share his/her scoring and grades and one student did not. To be clear, “incomplete” groups were only incomplete from a research standpoint due to the ethical permission. That is, these groups of collaborating students were “incomplete” because one student did not consent to share scoring and grades for the current study. The “incomplete” groups were retained in the study and followed the same procedures as the “complete” groups in the course.

Cohort 2 consisted of 44 students (40 female, 4 male) and since the scoring and grades are archive data, the ethics committee of the respective department allowed us to use the pseudonymized data without permission of the students. Hence, the scoring and grades of all forty-five students were retrieved. Since two students did not partake in several assessment parts, the final sample consisted of 42 students (39 female, 3 male). The ages of these participants were similar to those of Cohort 1. The 42 students were divided over 14 groups, all consisting of three students.

**Course Design, Learning Materials and Assignment, and Assessment**

All information discussed in this section is the same for Cohort 1 and Cohort 2.

**Course design.** The eight-week course on “instructional design” was delivered by two teachers with ample teaching experience and expertise in higher education as well as for this specific course (teacher 1, second author; teacher 2, fourth author). The course consisted of four mandatory 90-minute interactive online lectures and five mandatory practice assignments to prepare students for the assessed assignment. These five practice assignments were not graded, but formative feedback was provided. The practice assignments were performed in the same groups as the ones students worked in for the assessed assignment. The summative assessment consisted of (a) an individual, open book, essay examination and (b) an assignment, consisting of a near identical group and individual part (see next paragraph for more information on what was specifically assessed). In the first week, students formed self-selected groups of three students for the assessed assignment (there was one group of two students). The course was initially planned as a face-to-face course, but partly redesigned as an online-only course due to the Covid-19 pandemic.
Therefore, the lectures, student collaboration, and the individual essay examination took place online.

**Learning materials and their assessment formats**

*Textbook on the “Kemp Model” and its assessment.* The students studied ten core chapters of a textbook on the “Kemp model” for instructional design (Morrison et al., 2013). The individual, online open book, essay examination consisted of six questions in which students’ (application of) knowledge about the Kemp model was assessed (for more detail on the essay examination questions see Supplemental Material Part A.

*Textbook on the “4CID Model” and its assessment.* The students studied all chapters of a textbook on the “4CID model” (Hoogveld et al., 2017) and they completed an educational design assignment, consisting of a near identical group and individual part, to assess their application of knowledge about the model. Each group was required to develop four so-called “complexity levels” (but only three levels for the group of two students) for a complex skill using the 4CID model. Students worked as a group at the first level (i.e., the group part of the assignment), but they were each in charge of one of the remaining levels individually (i.e., the individual part of the assignment). Groups first constructed a draft design of all complexity levels and lessons within them for their fifth practice assignment, on which they received written feedback as input for a subsequent 15-minute online consultation session with the second author (teacher 1). Afterwards, they worked on the group and individual parts in parallel. Each student had to hand in the group and individual part at the same time at the end of the course. Since they were allowed to revise both parts of the assignment during the course, it is likely that students used information from the group part in their individual part and vice versa to improve the assignment. Students worked on both the group and individual part of the assignment using the same structure, that is, they developed three learning tasks, provided an argumentation for the design decisions, and added a reference list. The only differences were: (a) the group part required an introduction of the complex skill and the four (or three) complexity levels and (b) the four (or three) complexity levels and the associated learning tasks differed slightly due to their increasing complexity. Figure 1 shows an example of the group and individual parts within one group’s assignment, with “Drawing a scaled map of the school with an accurate legend” being the group’s complex skill of choice. Students mainly worked on the assignment in their own time without the direct supervision of teachers. This means that students may have worked together, within or between groups, on either part of the assignment; for example, students could have informally peer-reviewed (between and within groups) each other on both pieces.
Course assessment and grading. Summative assessment in this course consisted of an individual essay examination grade and an assignment grade. The assignment grade was a composite of (a) the group part of the assignment and (b) the individual part of the assignment, both making up 50% of the assignment grade. The course grade was a composite of the individual essay examination which accounted for 30%, and the assignment grade which accounted for 70% (of which 50% group and 50% individual) (see Figure 2). We collected students’ grades for the individual essay examination and the assignment, including the composite and partial grades of the group and individual part of the assignment. All students were allowed to re-sit the individual essay examination and the assignment. When students decided to re-sit, we took that grade as the final grade in our dataset.

Scoring and grading of assessment formats. The individual essay examination consisted of six questions. In total 100 points could be gained (see Supplemental Material Part A for distribution) and was graded by the second author (teacher 1) using a template per (sub)question (e.g., see Supplementary Material Part B). The group and individual part of the assignment for Cohort 1 were all scored by the fourth author (teacher 2) and for Cohort 2, the fourth author (teacher 2) scored eight groups and the remaining six were scored by a teaching assistant (trained by the fourth author; teacher 2). The group and individual part were scored on an anchored visual analog scale (Svensson, 2000). An anchored visual analog scale consists of a range of possible values between the ends of a straight line representing a continuum, with anchors to

Figure 1. Example outline of the group and individual part of the assignment.
aid the assessor in determining the position on the scale. In this study, the scale consisted of a continuum of 10 cm, corresponding to the Dutch grading system that ranges from 1 to 10, with the anchors “fail” (grades ranging from 1 to 5), “sufficient” (grades ranging from 5.5 to 7.5) and “good” (grades ranging from 8 to 10). Figure 3 shows the anchored visual analog scale for the criterion “complexity level” and the score—that is, position of the slider on the continuum—assigned by the teacher.

| Complexity level | Assessment |
|------------------|------------|
| Description of complexity level | Fail | Sufficient | Good |
| Description of supporting information for complexity level | | | |
| Description of procedural information for complexity level | | | |

**Figure 2.** Overview of summative assessment in the course on “instructional design.”

**Figure 3.** Example of a visual analog scale used in this course (criterion “complexity level”).
For pragmatic reasons (i.e., available time), teacher 2 decided not to measure the number of millimeters for every criterion, but holistically assigned a grade based on the “general impression” of all scored criterions for the individual and group part of the assignment. For the purpose of the current study, however, it was decided to use the most precise method to assign grades and therefore we used score-anchored grades. That is, the first author used a ruler to measure the number of millimeters for each anchored visual analog scale leading to a score between 0 and 100 based on the length measured in millimeters on each assessment criterium. Subsequently, an average score was computed for the group and individual part of the assignment to determine scoring-anchored grades for the group and individual part. The score-anchored grades were converted to grades on a 1 to 10 grading scale with increments of 0.1 according to the Dutch grading system. Both sets of grades—holistic and score-anchored—were compared, revealing a high correlation between them: For Cohort 1, \( r(28)_{\text{group part}} = 0.91 \) and \( r(28)_{\text{individual part}} = 0.94 \), and for Cohort 2, \( r(42)_{\text{group part}} = 0.95 \) and \( r(42)_{\text{individual part}} = 0.92 \).

**Analyses**

The analyses to determine construct and consequential validity were performed using four grades: (a) *examination grade* for the individual essay examination, reflecting individual assessment of individual learning, (b) *combined assignment grade* for the combination of the group and individual part of the assignment, reflecting combined assessment of collaborative learning, (c) *group grade* for the group part of the assignment reflecting group assessment of collaborative learning, and (d) *individual grade* for the individual part of the assignment reflecting individual assessment of collaborative learning. The analyses were performed under the assumption that curricula in higher education typically emphasize the measurement of students’ individual domain-specific abilities. Although the assignment and individual essay examination were different kinds of tasks, they also shared similarities: both tasks were in the context of the same course on instructional design, both tasks mainly assessed the declarative knowledge of students, in both tasks students had to demonstrate writing skills, and students were allowed to use reference works in both tasks. Hence, the tasks were sufficiently similar for meaningful comparison of construct and consequential validity (for an overview of the analyses see Table 1).

The construct validity of group assessment of collaborative learning (research question 1) was explored by calculating the correlation between *group grade* and (a) *individual grade* and (b) *examination grade* as proxies for students’ individual domain-specific abilities. The consequential validity of group assessment of collaborative learning was explored by (a) visualizing
| Research question | Validity type | Grade | Contrast Grade | Percentages/points advantaged/disadvantaged | Fail/pass and distinction/not distinction comparison |
|-------------------|--------------|-------|---------------|---------------------------------------------|-----------------------------------------------|
| 1                 | Construct    | Group Grade | Individual grade | Examination grade | Between group grade and individual grade | Between group grade and individual grade |
|                   |              | Consequential |               | | | |
| 2                 | Construct    | Individual grade | Examination grade | | | Between combined assignment grade (50/50) and individual grade |
| 3                 | Construct    | Combined assignment grade (50/50) | Individual grade | Examination grade | Between combined assignment grade (50/50) and individual grade | Between combined assignment grade (50/50) and individual grade |
|                   |              | Consequential |               | | | |
| 4                 | Construct    | 19 potential compositions of the combined assignment grade (incl. (50/50)) | Individual grade | Examination grade | Between 19 potential compositions of the combined assignment grade (incl. (50/50)) and individual grade | Between 19 potential compositions of the combined assignment grade (incl. (50/50)) and individual grade |
students’ grades per group for the group and individual parts of the assignment, (b) calculating the percentages and points by which students with a lower individual grade than group grade would have been advantaged (on average) and by which students with a higher individual grade than group grade would have been disadvantaged (on average), and (c) comparing whether the individual grade and group grade would have been above 5.5 (the passing threshold) and above 8.0 (the distinction threshold). We preferred the individual grade over the examination grade as proxy for students’ individual domain-specific abilities for exploring consequential validity of group assessment, because individual grade was more comparable with group grade, given the individual and group part of the assignment were near identical. The construct validity of individual assessment of collaborative learning (research question 2) was explored by calculating the correlation between individual grade and examination grade as proxy for students’ individual domain-specific abilities.

The construct validity of combined assessment of collaborative learning, with the individual and group part of the assignment weighted 50/50 (research question 3), was explored by calculating the correlation between combined assignment grade and individual grade as a proxy for students’ individual domain-specific abilities. As we are aware that this comparison is not ideal, since the combined assignment grade consists for 50% of the individual grade, we also compared the combined assignment grade with the examination grade as a proxy for students’ individual domain-specific abilities. The consequential validity of combined assessment of collaborative learning was explored by (a) calculating the percentages and points by which students with a lower individual grade than assignment grade would have been advantaged and students with a higher individual grade than combined assignment grade would have been disadvantaged, and (b) comparing whether the individual grade or the combined assignment grade would have been above 5.5 (the passing threshold) and above 8.0 (the distinction threshold).

Finally, research question 4 explored 19 potential compositions of weighting the individual grade and group grade as part of the combined assignment grade (i.e., incl. 50/50 weighting). The construct validity of these 19 potential compositions was explored by calculating the correlation between 19 potential compositions of combined assignment grade—consisting of individual grade (weighted from 5% to 95%, in 5% increment) and group grade (weighted from 95% to 5%, in 5% increment)—and (a) individual grade and (b) examination grade as proxy for students’ individual domain-specific abilities. The consequential validity of these 19 potential compositions of combined assignment grade was explored by calculating the percentages and points by which
students with a lower *individual grade* than *combined assignment grade* would have been advantaged and students with a higher *individual grade* than *combined assignment grade* would have been disadvantaged.

Concerning the interpretation of the correlation coefficients, Cohen (1988) provided general guidelines of .10 (small effect size), .30 (medium effect size), and .50 (large effect size). However, Cohen (1988) also explicitly noted that “The terms ‘small’, ‘medium’, and ‘large’ are relative, not only to each other, but to the area of behavioral science or even more particularly to the specific content and research method being employed in any given investigation” (p. 25). Hence, we consulted earlier research examining correlations between measures of students’ individual domain-specific abilities and found that most correlations were between .40 and .70 (Lejk et al., 1999; Smith, 1994). Based on these findings and following Cohen’s (1988) advice on tailoring the guidelines to the area of behavioral science, we used the following guidelines when interpreting the correlations: .10 as reflecting low construct validity, .30 as reflecting medium construct validity, and .40 and above as reflecting high construct validity.

**Results**

**Descriptives for the Grades**

Figure 4 displays boxplots for *examination grade*, *combined assignment grade*, *group grade*, and *individual grade* for Cohort 1 (C1 in Figure 4) and 2 (C2 in Figure 4). In Cohort 1, the median for the *combined assignment grade*, *group grade*, and *individual grade* is very similar (about 7.0), with the *group grade* showing the least variation (range = 3.00, *SD* = 0.77), compared to the *combined assignment grade* (range = 4.00, *SD* = 0.91) and *individual grade* (range = 6.10, *SD* = 1.41). Second, the median for the *examination grade* (8.10) is higher than the *group grade*, *individual grade*, and *combined assignment grades*. In addition, the variation in *examination grade* (range = 3.50, *SD* = 0.82) is similar to that of *group grade* (range = 3.00, *SD* = 0.77) and *combined assignment grade* (range = 4.00, *SD* = 0.91), but lower than *individual grade* (range = 6.10, *SD* = 1.41).

In contrast, in Cohort 2, the median for *combined assignment grade*, *group grade*, and *individual grade* ranges between about 6.8 and 7.4. In addition, the variation within the *group grade* (range = 3.6, *SD* = 1.15), *combined grade* (range = 3.6, *SD* = 1.01), and *individual grade* (range = 3.9, *SD* = 1.03) is more similar compared to Cohort 1. The *examination grade* has a higher median (8.10) and more variation (range = 4.6, *SD* = 0.99) than the other assessment methods.
Construct validity. For Cohort 1, the correlation (a) between group grade and individual grade was $r(28) = .33$, $p = .083$, and (b) between group grade and examination grade was $r(28) = .20$, $p = .301$. For Cohort 2, the correlation (a) between group grade and individual grade was $r(42) = .74$, $p = .000$, and (b) between group grade and examination grade was $r(42) = .15$, $p = .343$. The scatterplots in Figure 5 provide more details about the nature of the distributions for both cohorts.

Consequential validity. Figure 6 displays for Cohort 1’s seven complete groups of three students and four incomplete groups (Group 1, 2, 7, and 10), including three groups of which two out of three students agreed to participate, and one group of which one out of two students was included in the study. In Group 3, two students were awarded an eight as individual grade and are therefore displayed as a single black square with the number two in it. Figure 6 displays substantial individual differences in individual grade within groups (black squares) compared to the group grade (circles), indicating within-group differences in students’ domain-specific abilities. For example, the group with the largest range (Group 4) has a range of about 5.5 points and the groups with lowest range (Group 3, 5, and 8) have a range of about two points. On average, students with a higher individual grade than group grade scored 12.41% lower (1.07 points) on their group grade, whereas
students with a lower individual grade than group grade scored 25.94% higher (1.38 points) on their group grade. In addition, in terms of the difference between what students achieved individually and as a group, Figure 6 shows that two students individually scored lower than a 5.5 (pass grade), whereas their group grade was higher. Additionally, nine students individually scored an eight or higher (distinction grade), but only one achieved this on the group grade. Concerning between-group differences, Figure 6 indicates that students with similar individual grades in different groups can have different group grades. For example, two students in Groups 6 and eleven have a very similar individual grade (7.1 for a student in Group 11 vs. 7.0 for a student in Group 6), but the student in Group eleven received an 8.6 for the group grade, whereas the student in Group 6 a 7.0.
Concerning Cohort 2, Figure 7 shows 14 complete groups of three students. In group 14, two students were awarded an eight as individual grade and are therefore displayed as a single black square with the number two in it. Similar to Cohort 1, Figure 7 shows substantial within and between-group differences in students’ domain-specific abilities. On average, students with a higher individual grade than group grade scored 8.47% lower (−0.57 points) on their group grade, whereas students with a lower individual grade than group grade scored 12.79% higher (0.79 points) on their group grade. Furthermore, in terms of the difference between what students achieved individually and as a group, Figure 7 shows that eight students individually scored lower than a 5.5 (pass grade), whereas six students did so for their group grade. Additionally, four students individually scored an eight or higher (distinction grade), and nine achieved this on the group grade.

**Individual Assessment of Collaborative Learning**

**Construct validity.** For Cohort 1, the correlation between the individual grade and examination grade was $r(28) = .49, p = .008$. For Cohort 2, the correlation between the individual grade and examination grade was $r(42) = .30, p = .050$. The scatterplots in Figure 8 provide more details as to the nature of the distributions for both cohorts.
**Figure 7.** Distribution for individual grade and group grade for all groups in cohort 2.

**Figure 8.** Scatterplots for Pearson’s correlation coefficient comparisons for construct validity of individual grade.

**Combined Assessment of Collaborative Learning (Weighted 50/50)**

**Construct validity.** For Cohort 1, the correlation between combined assignment grade (group and individual grade weighted 50/50) and individual grade was high, $r(28) = .92, p = .000$ as expected because the individual grade made up 50% of the combined assignment grade. Therefore, the correlation
between combined assignment grade (weighted 50/50) and examination grade was calculated as well, resulting in \( r(28) = .48, p = .010 \). For Cohort 2, the correlation between combined assignment grade (group and individual grade weighted 50/50) and individual grade was \( r(42) = .92, p = .000 \). The correlation between combined assignment grade (weighted 50/50) and examination grade was \( r(42) = .24, p = .120 \). The scatterplots in Figure 9 provide more details as to the nature of the distributions for both cohorts.

![Figure 9](image.png)

**Figure 9.** Scatterplots for Pearson’s correlation coefficient comparisons for construct validity of combined assignment grade.

**Consequential validity.** For Cohort 1, students with a higher individual grade than combined assignment grade scored, on average, 6.21% lower (0.53 points) on their combined assignment grade, whereas students with a lower individual grade than group grade scored, on average, 12.97% higher (0.69 points) on their combined assignment grade. With respect to a “pass”
decision (5.5 or higher), Table 2 displays two students in Cohort 1 (S1 and S2) who individually scored lower than a 5.5 on the individual grade, with the combined assignment grade for S1 being lower than a 5.5 but for S2 higher than a 5.5. Additionally, with respect to a “distinction” decision (8.0 or higher), nine students (S3–S9) individually scored an eight or higher, of which three achieved this on the combined assignment grade.

For Cohort 2, students with a higher individual grade than combined assignment grade scored, on average, 4.23% lower (−0.28 points) on their combined assignment grade, whereas students with a lower individual grade than group grade scored, on average, 6.4% higher (0.41 points) on their combined assignment grade. With respect to a “pass” decision (5.5 or higher), Table 2 shows that eight students in Cohort 2 (S1–S8) individually scored lower than a 5.5 (pass grade), with the combined assignment grade being lower than a 5.5 for five students. With respect to the distinction grade (8.0 or higher), four students (S9–S12) individually scored an eight or higher, of which three achieved this on the combined assignment grade.

### Nineteen Potential Compositions of Combined Assessment of Collaborative Learning

**Construct validity.** Whereas the group and individual part were weighted 50/50 in the course, there are 18 additional potential compositions of weighting. In both cohorts, we found that, naturally, the higher the share of the individual grade in the combined assignment grade, the higher the correlation between them (see Figure 10 for Cohort 1, Figure 11 for Cohort 2). We also explored the construct validity of the additional 18 compositions of the combined assignment grade with the examination grade (see Figure 10 for Cohort 1 and 11 for Cohort 2). In Cohort 1, increasing the weight of individual grade
relative to group grade steeply increases the correlation with each increment of 5% in the beginning, but decreases from about 50% and becomes relatively stable, between .47 and .50. This pattern does not occur in Cohort 2.

Consequential validity. Table 3 shows for both cohorts the consequences in terms of percentage (points) change for 19 potential compositions of combined assignment grade (i.e., including 50/50 weighting), by which students with a higher individual grade than combined assignment grade would have been disadvantaged and students with a lower individual grade than combined assignment grade would have been advantaged.
Table 3. Percentage (points) change for 19 compositions of weighting (incl. 50/50) by which students would have been advantaged or disadvantaged.

| Composition assignment grade | Cohort 1 Percentage (points) change for students with higher individual grades | Cohort 1 Percentage (points) change for students with lower individual grades | Cohort 2 Percentage (points) change for students with higher individual grades | Cohort 2 Percentage (points) change for students with lower individual grades |
|-----------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|--------------------------------------------------------------------------------|
| Group 5%—Individual 95%     | -0.62% (0.05)                                                                 | +1.30% (+0.07)                                                                | -0.39 (-0.09)                                                                | +0.64 (+0.07)                                                                  |
| Group 10%—Individual 90%   | -1.24% (-0.11)                                                                | +2.59% (+0.14)                                                                | -0.85 (-0.10)                                                                | +1.28 (+0.11)                                                                  |
| Group 15%—Individual 85%   | -1.86% (-0.16)                                                                | +3.89% (+0.21)                                                                | -1.27 (-0.12)                                                                | +1.92 (+0.15)                                                                  |
| Group 20%—Individual 80%   | -2.48% (-0.21)                                                                | +5.19% (+0.28)                                                                | -1.69 (-0.13)                                                                | +2.47 (+0.20)                                                                  |
| Group 25%—Individual 75%   | -3.10% (-0.27)                                                                | +6.49% (+0.34)                                                                | -2.12 (-0.15)                                                                | +3.20 (+0.24)                                                                  |
| Group 30%—Individual 70%   | -3.72% (-0.32)                                                                | +7.78% (+0.41)                                                                | -2.36 (-0.17)                                                                | +3.96 (+0.25)                                                                  |
| Group 35%—Individual 65%   | -4.34% (-0.37)                                                                | +9.08% (+0.48)                                                                | -2.96 (-0.20)                                                                | +4.48 (+0.29)                                                                  |
| Group 40%—Individual 60%   | -4.96% (-0.43)                                                                | +10.38% (+0.55)                                                               | -3.39 (-0.23)                                                                | +5.12 (+0.33)                                                                  |
| Group 45%—Individual 55%   | -5.59% (-0.48)                                                                | +11.67% (+0.62)                                                               | -3.81 (-0.26)                                                                | +5.76 (+0.38)                                                                  |
| Group 50%—Individual 50%   | -6.21% (-0.53)                                                                | +12.97% (+0.69)                                                               | -4.23 (-0.28)                                                                | +6.40 (+0.41)                                                                  |
| Group 55%—Individual 45%   | -6.83% (-0.59)                                                                | +14.27% (+0.76)                                                               | -4.66 (-0.31)                                                                | +7.04 (+0.44)                                                                  |
| Group 60%—Individual 40%   | -7.54% (-0.64)                                                                | +15.57% (+0.83)                                                               | -5.08 (-0.34)                                                                | +7.68 (+0.48)                                                                  |
| Group 65%—Individual 35%   | -8.07% (-0.69)                                                                | +16.86% (+0.90)                                                               | -5.50 (-0.37)                                                                | +8.32 (+0.52)                                                                  |
| Group 70%—Individual 30%   | -8.69% (-0.75)                                                                | +18.16% (+0.97)                                                               | -5.50 (-0.40)                                                                | +8.95 (+0.56)                                                                  |
| Group 75%—Individual 25%   | -9.31% (-0.80)                                                                | +19.46% (+1.03)                                                               | -6.35 (-0.43)                                                                | +9.59 (+0.59)                                                                  |
| Group 80%—Individual 20%   | -9.93% (-0.85)                                                                | +20.75% (+1.10)                                                               | -6.77 (-0.46)                                                                | +9.88 (+0.63)                                                                  |
| Group 85%—Individual 15%   | -10.55% (-0.91)                                                               | +22.05% (+1.17)                                                               | -7.20 (-0.48)                                                                | +10.87 (+0.67)                                                                  |
| Group 90%—Individual 10%   | -11.17% (-0.96)                                                               | +23.35% (+1.24)                                                               | -7.62 (-0.51)                                                                | +11.51 (+0.71)                                                                  |
| Group 95%—Individual 5%    | -11.79% (-1.02)                                                               | +24.65% (+1.31)                                                               | -7.47 (-0.54)                                                                | +12.15 (+0.75)                                                                  |
Discussion

Over the last decades, assessment of collaborative learning seemed to increasingly impact students’ assessment and thus their individual transcript of records (Bacon, 2005; Bourner et al., 2001; De Hei et al., 2015; Flores et al., 2015; Lejk et al., 1997). However, curricula in higher education typically emphasize the necessity of measuring students’ individual domain-specific abilities (e.g., knowledge and/or skills). Collaborative learning assessment methods might not align with this orientation, leading to tensions with respect to construct and consequential validity of different assessment methods of collaborative learning (Almond, 2009; Harvey et al., 2019; Lejk et al., 1999; Meijer et al., 2020; Plastow et al., 2010; Saner et al., 1994; Webb, 1993). Therefore, we explored the construct validity of group assessment, individual assessment, and combined assessment of collaborative learning, as well as the consequential validity of group assessment and combined assessment of collaborative learning.

The current study was conducted in two consecutive cohorts of a fourth and final-year undergraduate course of an academic teacher training program at a university in The Netherlands. Cohort 1 (2019/2020) consisted of 28 students divided over 11 groups, and Cohort 2 (2020/2021) consisted of 42 students, divided over 14 groups. We explored our research questions under the assumption that the construct of interest in higher education assessment is typically students’ individual domain-specific abilities. Considering Cohen’s (1988) general guidelines and earlier research examining correlations between individual measures of individual learning (Lejk et al., 1999; Smith, 1994), we used the following guidelines to interpret correlations to determine construct validity: 0.10 as low construct validity, 0.30 as medium construct validity, and 0.40 and above as high construct validity.

Construct Validity of Group, Individual, and Combined Assessment of Collaborative Learning

The construct validity of group assessment of collaborative learning was low to medium in Cohort 1 and low to high in Cohort 2. In both cohorts, the comparison between a group assignment grade and an individual examination grade resulted in low construct validity. In our view, the most informative comparison was between the group part of the assignment and the individual part of the assignment, which resulted in medium construct validity in Cohort 1 and high construct validity in Cohort 2. The difference between the correlation coefficients across the cohorts was rather considerable, which is in line with findings in existing literature. For example, Lejk et al. (1999) found high
construct validity for group assessment, whereas Plastow et al. (2010) found low construct validity. One possible explanation for this difference is that students’ individual domain-specific abilities for the assignment, as measured with the individual grade for the individual part of the assignment, showed larger variance for students in Cohort 1 compared to students in Cohort 2. That is, there were more similar-ability groups in Cohort 2 than in Cohort 1. As was shown in Lejk et al. (1999), similar-ability groups yield higher construct validity than mixed-ability groups. Similarly, we found that the construct validity of individual assessment of collaborative learning was high in Cohort 1 and medium in Cohort 2. Hence, as with group assessment, the construct validity differs across cohorts.

The construct validity of combined assessment of collaborative learning (weighted 50/50), as shown by comparing the combined assessment of the assignment with the individual grade for the individual part of the assignment, was high in both cohorts. However, since the combined assessment consisted for 50% of the individual grade for the individual part of the assignment, our finding of high construct validity should be interpreted with caution. Instead, the comparison of the combined assessment with the examination grade—resulting in high construct validity in Cohort 1 and low construct validity in Cohort 2—might be more informative. With respect to the other 18 potential compositions of combined assessment of collaborative learning, we found that, generally, the construct validity increases as the weight of the individual assessment increases. Since the combined assessment consisted partly of the individual grade for the individual part of the assignment, this was not surprising.

In all, our findings show (a) that construct validity of the assessment methods of collaborative learning can differ across cohorts and (b) that construct validity is not a fixed characteristic of the specific assessment methods. Rather, these assessment methods’ construct validity are affected by multiple variables (Meijer et al., 2020) such as group composition (e.g., similar or mixed ability groups; Lejk et al., 1999), group dynamics (e.g., the process of collaboration; Ohaja et al., 2013), and purpose of assessment (e.g., formative or summative assessment) (Sridharan et al., 2019). Furthermore, regardless of (a) the criterion used and (b) the cohort, our findings show that combined assessment of collaborative learning yielded greater construct validity compared to group assessment and was similar to, or higher, than individual assessment of collaborative learning. Hence, when students’ individual-domain specific abilities are at stake, combined assessment (weighted 50/50) might be better suited for the assessment of collaborative learning compared to group assessment of collaborative learning and just as construct valid as individual assessment of collaborative learning. Finally, our
findings also show that the criterion measure (i.e., the proxy of students’ individual domain-specific abilities) clearly matters when investigating the validity of collaborative learning assessment. Although it signals that (near) identical tasks should be preferred, it is simultaneously too early to tell, and this should be further explored in future studies.

**Consequential Validity of Group and Combined Assessment of Collaborative Learning**

We found that group assessment of collaborative learning could lead to unintended negative consequences, such as advantaging students with relatively lower individual grades and disadvantaging students with relatively higher individual grades. This effect was found in both cohorts, although the effects were larger in Cohort 1. These consequences were similar to those reported by Almond (2009), who found that students with relatively high individual grades received lower group grades (on average −9%) and students with relatively low individual grades received higher group grades (on average +35%). In both our cohorts as well as in Almond’s (2009) study, the advantage for students with relatively lower individual grades is more significant than the disadvantage for students with relatively higher individual grades. We found that combined assessment of collaborative learning mitigated these negative consequences, leading to smaller percentages by which students could have been (dis)advantaged.

In accordance with Plastow et al. (2010), we found in both cohorts that the assessment method can affect pass/fail and distinction decisions. More specifically, group assessment could lead to students passing a course merely based on their group assessment, whereas individually they would not have passed the course. Additionally, students could receive a lower than distinction grade based on their group assessment, but individually they would meet the criterion for a distinction grade. Combined assessment mitigated both issues.

**Practical Implications**

When policy makers, educational advisors, and teachers include assessed collaborative learning in their curriculum and when their construct of interest is students’ individual domain-specific abilities, our results have two implications. First, although the construct and consequential validity of assessment methods differs across cohorts, our results showed a consistent pattern within cohorts. That is, when assuming students’ individual domain-specific abilities as the standard for assessment, regardless of the comparison, we observed
that: (a) the construct validity increases when combined assessment of collaborative learning is used, compared to group assessment, and (b) negative consequences—such as adverse impact on students with relatively higher individual grades, beneficial impact on students with relatively lower individual grades, and students passing/failing a course while individually (not) meeting the domain-specific abilities as specified by a curriculum—are less severe with combined assessment, compared to group assessment of collaborative learning. Hence, when policy makers, educational advisors, and teachers want to assess students’ individual domain-specific abilities, combined assessment of collaborative learning might be the most appropriate assessment method for collaborative learning. That is, it provides relatively high construct validity (i.e., more accurately reflects students’ individual domain-specific abilities than group assessment of collaborative learning) and therefore mitigates the challenges of group assessment’s consequential validity (i.e., adverse or beneficial impact on students). Nevertheless, we advise to combine a group and individual assessment only when they measure similar domain-specific abilities (knowledge and/or skills), as was the case in the course that we examined. By doing so, it is more likely that students would receive a grade that reflect their individual domain-specific abilities and unintended consequences such as students’ passing a course while individually not meeting the domain-specific abilities are mitigated. In contrast, in the case where individual assessment and group assessment are combined but assess different kinds of individual domain-specific abilities, an individual fail grade on a given domain-specific ability might be compensated by a higher group grade on a different domain-specific ability. If so, a combined assessment might lead to students passing a course while not meeting the individual domain-specific abilities.

Second, other factors than construct and consequential validity, such as pragmatic considerations (e.g., grading time; cf. Harden, 1979) and didactic considerations (e.g., adherence to individual accountability and positive interdependence) influence the choices of policy makers, educational advisors, and teachers for an assessment method of collaborative learning. In terms of these factors, group assessment of collaborative learning ensures less grading time, but might also, for example, elicit free-riding because individual accountability is not fostered (Kagan, 1995; Pitt, 2000). Furthermore, individual assessment of collaborative learning might increase grading time and it is questionable whether students engage in genuine collaboration because positive interdependence is not fostered (Meijer et al., 2020). Therefore, combined assessment of collaborative learning might be appealing because, for example, it fosters both individual accountability and positive interdependence. In addition, our results suggest that a combined assessment with a weighting of about
50% for a group performance and 50% for an individual performance has relatively high construct validity. That is, it is similar to individual assessment of collaborative learning and more suitable than group assessment of collaborative learning. Moreover, a 50/50 weighting assures that both individual accountability and positive interdependence are fostered and that students will not prioritize one part over the other (Strijbos, 2011, 2016). Hence, from a perspective in which construct and consequential validity as well as pragmatic and didactic considerations are incorporated, combined assessment of collaborative learning appears better suited for the assessment of collaborative learning, compared to group and individual assessment.

Limitations

A first limitation of our study is the generalizability of our findings. We studied assessment of collaborative learning in two cohorts of the same course, in one educational program, and at one university. Despite these limitations, our findings were in part similar to previous studies exploring construct and consequential validity of assessment methods of collaborative learning with larger student samples across various courses and programs in higher education (e.g., Almond, 2009; Bacon et al., 1999; Lejk et al., 1999; Plastow et al., 2010).

A second limitation is that the current study only explored the products (i.e., grades) of collaborative learning and did not include information on the process of collaborative learning (e.g., task distribution, time investment of group members, or how often groups met). Insights into the process of collaboration can provide information on how students worked together, which in turn can inform decisions about construct and consequential validity of collaborative learning assessment methods. For example, when findings indicate that one student contributed less than their fellow group members, group assessment might be invalid. Therefore, we encourage future research to study the construct and consequential validity of assessment of collaborative learning by additionally including process-related measures.

A third limitation concerns the interpretation of the grades and correlation coefficients. We assumed that the grades would be an accurate indicator of students’ individual domain-specific abilities; however, as is the case with any measurement, grades are subject to random measurement errors (Barnett et al., 2005). Hence, grades, correlation coefficients, and the percentage and point change used to determine whether students were (dis)advantaged could be different when repeating the assessment, for example due to the regression-to-the-mean effect (Barnett et al., 2005). That is, the changes in percentages and points and the subsequent (dis)advantage could have been
smaller after a repeated assessment. Nevertheless, although repeated assessment might change the exact grades, correlation coefficients, percentages and points, and the subsequent (dis)advantage, we do not expect them to change the general conclusions of our study concerning construct and consequential validity of the studied assessment methods of collaborative learning.

**Research Implications**

First, in order to advance our understanding of the relation between collaborative learning design choices (e.g., group composition, assessment method, and assessment purpose) and the construct and consequential validity of collaborative learning assessment methods, more empirical studies are needed. Such research could be specifically geared toward combined assessment of collaborative learning (i.e., combining a group and individual grade based on a group and individual part in an assignment) to extend the findings in the present study and those reported by Plastow et al. (2010) and Harvey et al. (2019). At the same time, other factors such as didactic and pragmatic considerations should be considered when determining the appropriateness of any assessment method of collaborative learning. In addition, this study showed that (a) the criterion and (b) cohorts of students, among other variables such as group dynamics and assessment purpose, affect the construct and consequential validity of assessment methods. We therefore encourage researchers to choose near-identical criterions, as was done in this study, and to examine the construct and consequential validity across different cohorts.

Second, although it is important to further explore the construct and consequential validity of assessment methods of collaborative learning, research should also focus on how teachers and policy makers could deal with the tension between the emphasis on students’ individual domain-specific abilities in their transcripts of records and the inclusion of assessment of collaborative learning as a student outcome. For example, it might be possible to represent and differentiate in students’ transcripts of records between grades that students achieved individually, collaboratively, or a combination of both (Meijer et al., 2020). It will also be productive to explore (a) whether it is legally possible to implement such differentiation and (b) how policy makers, educational advisors, teachers, and students would perceive and value such differentiation. This kind of knowledge might enable policy makers, educational advisors, and teachers to make informed decisions regarding which methods they want to use for the assessment of collaborative learning in their curricula and courses.
Declaration of Conflicting Interests
The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding
The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs
Hajo Meijer https://orcid.org/0000-0002-8563-5333
Jasperina Brouwer https://orcid.org/0000-0002-7332-9320

Supplemental Material
Supplemental material for this article is available online.

References
Abernethy, A. M., & Lett, W. L. (2005). You are fired! A method to control and sanction free riding in group assignments. *Marketing Education Review, 15*(1), 47–54. https://doi.org/10.1080/10528008.2005.11488891

Ahern, A. (2007). What are the perceptions of lecturers towards using cooperative learning in civil engineering? *European Journal of Engineering Education, 32*(5), 517–526. https://doi.org/10.1080/03043790701433152

Almond, R. J. (2009). Group assessment: Comparing group and individual undergraduate module marks. *Assessment & Evaluation in Higher Education, 34*(2), 141–148. https://doi.org/10.1080/02602930801956083

Augar, N., Woodley, C. J., Whitefield, D., & Winchester, M. (2016). Exploring academics’ approaches to managing team assessment. *The International Journal of Educational Management, 30*(6), 1150–1162. https://doi.org/10.1108/ijem-06-2015-0087

Bacon, D. R. (2005). The effect of group projects on content-related learning. *Journal of Management Education, 29*(2), 248–267. https://doi.org/10.1177/1052562904263729

Bacon, D. R., Stewart, K. A., & Silver, W. S. (1999). Lessons from the best and worst student team experiences: How a teacher can make the difference. *Journal of Management Education, 23*(5), 467–488. https://doi.org/10.1177/105256299902300503

Barnett, A. G., Van Der Pols, J. C., & Dobson, A. J. (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology, 34*(1), 215–220. https://doi.org/10.1093/ije/dyh299

Boud, D., Cohen, R., & Sampson, J. (1999). Peer learning and assessment. *Assessment & Evaluation in Higher Education, 24*(4), 413–426. https://doi.org/10.1080/0260293990240405
Bourner, J., Hughes, M., & Bourner, T. (2001). First-year undergraduate experiences of group project work. *Assessment & Evaluation in Higher Education, 26*(1), 19–39. https://doi.org/10.1080/02602930020022264

Brookhart, S. M. (2013). *Grading and group work: How do I assess individual learning when students work together?*. ASCD.

Bushell, G. (2006). Moderation of peer assessment in group projects. *Assessment & Evaluation in Higher Education, 31*(1), 91–108. https://doi.org/10.1080/02602930500262395

Carson, K. M., & Glaser, R. E. (2010). Chemistry is in the news: Assessing intra-group peer review. *Assessment & Evaluation in Higher Education, 35*(4), 381–402. https://doi.org/10.1080/02602930902862826

Carver, T. L., & Stickley, A. (2012). Teamwork in first year law units: Can it work? *Journal of University Teaching & Learning Practice, 9*(2), 35–66. https://doi.org/10.53761/1.9.2.4

Cohen, J. (1988). *Statistical power for the behavioural sciences*. Lawrence Erlbaum.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

de Hei, M., Strijbos, J. W., Sjoer, E., & Admiraal, W. (2016). Thematic review of approaches to design group learning activities in higher education: The development of a comprehensive framework. *Review of Educational Research, 18*, 33–45. https://doi.org/10.1016/j.edurev.2016.01.001

De Hei, M. S. A., Sjoer, E., Admiraal, W., & Strijbos, J. W. (2016). Teacher educators’ design and implementation of group learning activities. *Educational Studies, 42*(4), 394–409. https://doi.org/10.1080/03055698.2016.1206461

De Hei, M. S. A., Strijbos, J. W., Sjoer, E., & Admiraal, W. (2015). Collaborative learning in higher education: Lecturers’ practices and beliefs. *Research Papers in Education, 30*(2), 232–247. https://doi.org/10.1080/02671522.2014.908407

Dijkstra, J., Latijnhouwers, M., Norbart, A., & Tio, R. A. (2016). Assessing the “I” in group work assessment: State of the art and recommendations for practice. *Medical Teacher, 38*(7), 675–682. https://doi.org/10.3109/0142159x.2016.1170796

Diprose, M. F., Judd, P. L., & Morris, A. S. (1997). The Sheffield Industrial Project scheme: Team teaching for engineering students. *Teaching in Higher Education, 2*(1), 59–68. https://doi.org/10.1080/1356251970020105

Flores, M. A., Veiga Simão, A. M., Barros, A., & Pereira, D. (2015). Perceptions of effectiveness, fairness and feedback of assessment methods: A study in higher education. *Studies in Higher Education, 40*(9), 1523–1534. https://doi.org/10.1080/03075079.2014.881348

Forsell, J., Forslund Frykedal, K., & Hammar Chiriac, E. (2020). Group work assessment: Assessing social skills at group level. *Small Group Research, 51*(1), 87–124. https://doi.org/10.1177/1046496419878269

Frykedal, K. F., & Chiriac, E. H. (2011). Assessment of students’ learning when working in groups. *Educational Researcher, 53*(3), 331–345. https://doi.org/10.1080/00131881.2011.598661

Harden, R. M. (1979). Assess students: An overview. *Medical Teacher, 1*(2), 65–70. https://doi.org/10.3109/01421597909019394
Harvey, H., Keen, J., Robinson, C., Roff, J., & Gross, T. (2019). Quantitative analysis of approaches to group marking. Assessment & Evaluation in Higher Education, 44(8), 1133–1147. https://doi.org/10.1080/02602938.2019.1572069

Hoogveld, B., Janssen-Noordman, A., & van Merriënboer, J. (2017). Innovatief onderwijs ontwerpen: De ontwerpprincipes van het 4CID-model [Innovative educational design: The design principles of the 4CID-model]. Noordhoff.

Jin, X. H. (2012). A comparative study of effectiveness of peer assessment of individuals’ contributions to group projects in undergraduate construction management core units. Assessment & Evaluation in Higher Education, 37(5), 577–589. https://doi.org/10.1080/02602938.2011.557147

Johnson, D. W. (1981). Student-student interaction: The neglected variable in education. Educational Researcher, 10(1), 5–10. https://doi.org/10.3102/0013189x010001005

Kagan, S. (1995). Group grades miss the mark. Educational Leadership: Journal of the Department of Supervision and Curriculum Development, N.E.A, 52(8), 68–71.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. Journal of Educational Measurement, 50(8), 1–73. https://doi.org/10.1111/jedm.12000

Lejk, M., Wyvill, M., & Farrow, S. (1997). Group learning and group assessment on undergraduate computing courses in higher education in the UK: Results of a survey. Assessment & Evaluation in Higher Education, 22(1), 81–91. https://doi.org/10.1080/0260293970220106

Lejk, M., Wyvill, M., & Farrow, S. (1999). Group assessment in systems analysis and design: A comparison of the performance of streamed and mixed-ability groups. Assessment & Evaluation in Higher Education, 24(1), 5–14. https://doi.org/10.1080/0260293990240101

Meijer, H., Hoekstra, R., Brouwer, J., & Strijbos, J. W. (2020). Unfolding collaborative learning assessment literacy: A reflection on current assessment methods in higher education. Assessment & Evaluation in Higher Education, 45(8), 1222–1240. https://doi.org/10.1080/02602938.2020.1729696

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. Educational Researcher, 18(2), 5–11. https://doi.org/10.3102/0013189x018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons’ responses and performances as scientific inquiry into score meaning. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066x.50.9.741

Morrison, G. R., Ross, S. M., Kalman, H. K., & Kemp, J. E. (2013). Designing effective instruction (7th ed.). John Wiley.

Nicolay, J. A. (2002). Group assessment in the on-line learning environment. New Directions for Teaching and Learning, 2002(91), 43–52. https://doi.org/10.1002/tl.65

Nordberg, D. (2008). Group projects: More learning? Less fair? A conundrum in assessing postgraduate business education. Assessment & Evaluation in Higher Education, 33(5), 481–492. https://doi.org/10.1080/02602930701698835
Ohaja, M., Dunlea, M., & Muldoon, K. (2013). Group marking and peer assessment during a group poster presentation: The experiences and views of midwifery students. Nurse Education in Practice, 13(5), 466–470. https://doi.org/10.1016/j.nepr.2012.11.005

Panadero, E., Romero, M., & Strijbos, J. W. (2013). The impact of a rubric and friendship on peer assessment: Effects on construct validity, performance, and perceptions of fairness and comfort. Studies in Educational Evaluation, 39(4), 195–203. https://doi.org/10.1016/j.stueduc.2013.10.005

Pitt, M. J. (2000). The application of games theory to group project assessment. Teaching in Higher Education, 5(2), 233–241. https://doi.org/10.1080/135625100114876

Plastow, N., Spiliotopoulos, G., & Prior, S. (2010). Group assessment at first year and final degree level: A comparative evaluation. Innovations in Education and Teaching International, 47(4), 393–403. https://doi.org/10.1080/14703297.2010.518427

Reckase, M. D. (1998). Consequential validity from the test developer’s perspective. Educational Measurement Issues and Practice, 17(2), 13–16. https://doi.org/10.1111/j.1745-3992.1998.tb00827.x

Roediger, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Saner, H., McCaffrey, D., Stetcher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. Educational Assessment, 2(4), 325–338. https://doi.org/10.1207/s15326977ea0204_4

Scriven, M. (1967). The methodology of evaluation. In R. E. Stake (Ed.), Perspectives of curriculum evaluation (Vol. 1, pp. 39–55). Rand McNally.

Sharp, S. (2006). Deriving individual student marks from a tutor’s assessment of group work. Assessment & Evaluation in Higher Education, 31(3), 329–343. https://doi.org/10.1080/02602930500352956

Shepard, L. A. (1993). Chapter 9: Evaluating test validity. Review of Research in Education, 19(1), 405–450. https://doi.org/10.3102/0091732x019001405

Slavin, R. E. (1980). Cooperative learning in teams: State of the art. Educational Psychologist, 15(2), 93–111. https://doi.org/10.1080/00461528009529219

Slomp, D. H., Corrigan, J. A., & Sugimoto, T. (2014). A framework for using consequential validity evidence in evaluating large-scale writing assessments: A Canadian study. Research in the Teaching of English, 48(3), 276–302.

Sluijsmans, D., & Segers, M. (2018). Wat is nodig voor een toetsrevolutie in het hoger onderwijs? Vijf kernboodschappen voor de praktijk [What is required for a testing revolution in higher education? Five lessons for practice]. In D. Sluijsmans & M. Segers (Eds.), Toetsrevolutie: Naar een feedbackcultuur in het hoger onderwijs [Testing revolution: To a feedback culture in higher education] (pp. 216–232). Phronese.

Smith, H., McCaffrey, D., Stetcher, B., Klein, S., & Bell, R. (1994). The effects of working in pairs in science performance assessments. Educational Assessment, 2(4), 325–338. https://doi.org/10.1207/s15326977ea0204_4
Sridharan, B., Tai, J., & Boud, D. (2019). Does the use of summative peer assessment in collaborative group work inhibit good judgement? *Higher Education, 77*(5), 853–870. https://doi.org/10.1007/s10734-018-0305-7

Strijbos, J. W. (2011). Assessment of (computer-supported) collaborative learning. *IEEE Transactions on Learning Technologies, 4*(1), 59–73. https://doi.org/10.1109/tlt.2010.37

Strijbos, J. W. (2016). Assessment of collaborative learning. In G. T. L. Brown & L. Harris (Eds.), *Handbook of social and human conditions in assessment* (pp. 302–318). Taylor and Francis.

Svensson, E. (2000). Comparison of the quality of assessments using continuous and discrete ordinal rating scales. *Biometrical Journal, 42*(4), 417–434. https://doi.org/10.1002/1521-4036(200008)42:4<417::aid-bimj417>3.0.co;2-z

Webb, N. M. (1993). Collaborative group versus individual assessment in mathematics: Processes and outcomes. *Educational Assessment, 1*(2), 131–152. https://doi.org/10.1207/s15326977ea0102_3

Willcoxson, L. E. (2006). “It’s not fair!”: Assessing the dynamics and resourcing of teamwork. *Journal of Management Education, 30*(6), 798–808. https://doi.org/10.1177/1052562906287964

**Author Biographies**

**Hajo Meijer** received a B.A. and M.A. in psychology at the University of Groningen. He is a Ph.D. student in the Department of Educational Sciences; University of Groningen in the Netherlands. His Ph.D. research focuses on the assessment of collaborative learning in higher education.

**Jasperina Brouwer** is assistant professor at the Department of Educational Sciences; University of Groningen in the Netherlands. She investigates the impact of social networks and social capital on academic and professional development within the educational context and at the workplace.

**Rink Hoekstra** an assistant professor at the Department of Educational Sciences; University of Groningen in the Netherlands. His studies how researchers understand and use statistical outcomes, and how the scientific context influences the use of data. He has published regularly about use and understanding of significance testing and its alternatives.

**Jan-Willem Strijbos** is a Full Professor of Learning and Instruction at the Department of Educational Sciences, University of Groningen, the Netherlands. His research focuses on the design, implementation, and effectiveness of interactive learning practices—for example, collaborative learning, peer assessment and feedback, learning communities—in physical and virtual settings.