Athens Journal of Education - Volume 10, Issue 4, November 2023 – Pages 575-592

Process Indicators for Grading Group Essays: Learning Analytics of Assessment Data and Online Behaviour

By Mei-Shiu Chiu[*] & Ya Ping (Amy) Hsiao[±]

The aim of this study was to identify process-related indicators for grading group essays. The research participants were students registered in a teacher-training course using an instructional design with face-to-face and digital blended learning. The course required the students in small collaborative groups to design, implement, and evaluate a teaching program using creative pedagogical designs, which were documented using group essays. Four indicators relating to group essays along the course process were collected: (A) group essay grades assessed by different agents, (B) students' other course grades or behaviours (i.e., multiple assessments) as well as (C) comment behaviours and (D) version history behaviours through an online co-editing system (i.e., Google Docs). Statistical analysis results indicated that the instructor's group essay grades were related to the group essay grades assessed by out-group peers (i.e. peers from other groups), online group comment frequencies, and online group comment interaction density.

Keywords: assessment methods and tools, collaborative learning, essay grading, learning analytics, higher education

Introduction

Learning, communication, and collaboration skills are three essential skills in the 21st century (Messersmith, 2015). Students can acquire and exercise these skills simultaneously through appropriate pedagogies focusing on social interaction (e.g., group work). Despite the benefits of group work for students to practice these skills, assessing group work remains challenging for instructors in higher education. One of these challenges is assigning fair grades to individual group members (King & Behnke, 2005), especially when there are free riders in the group, which is a well-known drawback of using a collective grade for group work (Maiden & Perry, 2011). Even though it is reasonable to assess the group process and take it into account for grading, monitoring and assessing the group process is technically difficult. It is like grading a black box (Davies, 2009).

Grading group essays is especially challenging for instructors when group work takes place outside the classroom, lasts for a long time, and consists of several stages. A recent development in real-time group editing techniques (e.g., Google Docs) can help document the group process (Woodrich & Fan, 2017; Zhou, Simpson, & Domizi, 2012) and increase essay performance and collaborative learning (Suwantarathip & Wichadee, 2014).

[*]Professor, Department of Education, National Chengchi University, Taiwan.
[±]Assessment Specialist, Tilburg University, The Netherlands.

https://doi.org/10.30958/aje.10-4-2 doi=10.30958/aje.10-4-2

Ideally, classroom assessment on student products generated from real educational settings should be considered as part of a course design based on research-based pedagogical and learning theories; that is, classroom assessment is of, for and as learning (Black & Wiliam, 2018). This paper focuses on group essays, as one of the major learning outcomes (assessments) of a course studied by this study. The course was designed on the basis of sociocultural learning and creativity theories. With the design, students' group-essay grades may relate to other student assessments and behavioural outcomes, which suggest effective indicators for grading group essays. The following literature review addresses the course's pedagogical design and related student assessment/behavioural outcomes in more detail.

Theories for Group Essays as Assessment to Address Course Objectives and Pedagogy

Classroom assessment needs to be grounded in learning theories to support and optimize student learning outcomes. The teacher-training course described in this study was developed on the basis of sociocultural learning theory (SLT) and 4P (Person, Product, Place, and Process) creativity theory (4PCT), which was transformed into the course' major objective: "Students are able to develop a creative pedagogical design by collaborating with others". Group essays are one of the most appropriate and most valid assessment methods for addressing this course's objective and pedagogy.

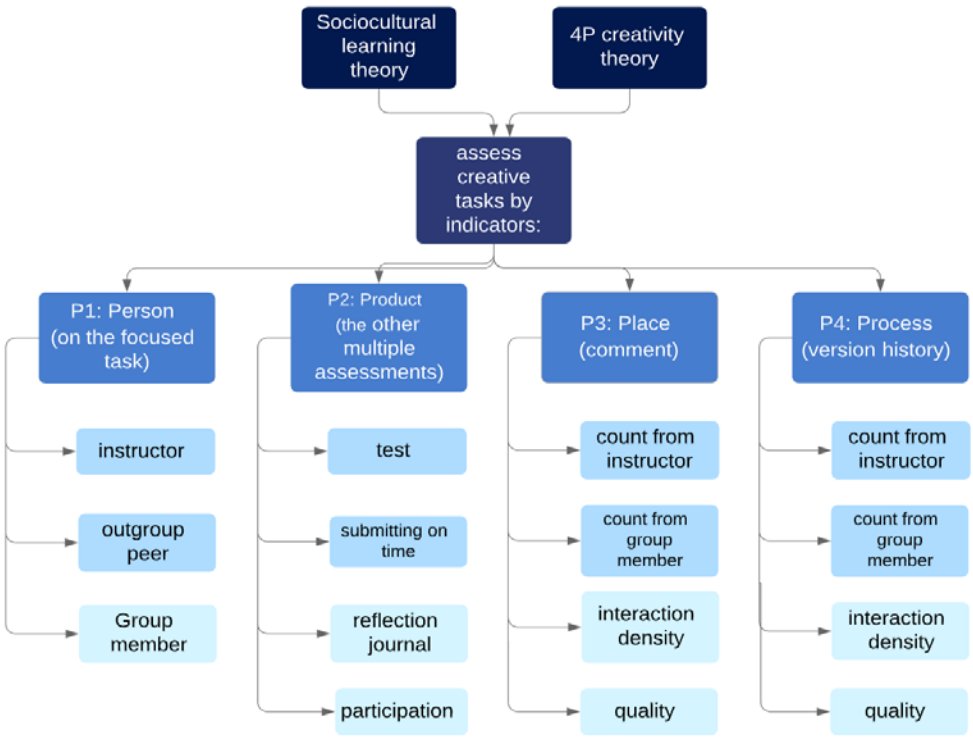
SLT emphasizes learners' participation in a learning community to develop higher-order cognitive, emotional, and social abilities (Zeidler, 2016). SLT is, therefore, suitable for instructional designs aiming to promote collaborative learning (e.g., writing group essays) that emphasizes sharing knowledge, norms, and practices in a learning community. SLT can also serve as the theoretical basis for a teaching, assessment and research design in a complex off- and on-line blended, collaborative learning environment (Shepard, Penuel, & Pellegrino, 2018).

Essay writing is a creative task, which calls for a course design that considers the four elements of creativity, named 4P creativity theory (4PCT) in this study (Hasirci & Demirkan, 2003). Group essay grades are a summative assessment result of students' collective creations. As suggested by SLT, grading group essays should not only be based on the summative assessment results of the group product, but also consider the continuous assessment results of the group process (Zeidler, 2016; Shepard, Penuel, & Pellegrino, 2018). An SLT-based continuous assessment of a collective creative task (e.g., the group essay in this study) needs to include assessment indicators of active interactions between people, product, place (or environment), and process (i.e., the 4PCT).

Using the SLT and the 4PCT as theoretical framework, potential indicators for group essays need to emphasize interactions in four different aspects: the actions taken by various agents (person), multiple assessments that reflect students' diverse abilities (product), and interactions between in-group peers as they are working on their task, as revealed by their dialogues (e.g., comments) in the

environment (place) and their essay version histories (process). Figure 1 presents the Framework for Indicators of Creative group tasks (the FIC) using SLT and 4PCT as the theoretical basis for the pedagogical, assessment, and research design of this study. The FIC suggests four indicator categories as follows.

Figure 1. The Framework for Indicators of Creative Group Tasks (FIC)



"Person" Indicators: Group Essay Grades Assessed by Different Agents

Essay tasks can be assessed by different agents in the learning environment. The most frequent agents are instructors, out-group peers, and in-group peers.

Instructors. Instructor grading is often the main or only criterion of students' grades or the summative assessment. Despite their lack of reliability, school teachers' grades remain the basis for educational decision-making about students' learning (Guskey & Link, 2019).

Out-Group Peers. Out-group peers are classmates who do not work on the same group essays. Out-group peers are the second type of agent. They have been extensively researched and are recommended for use in formative assessment activities. A peer feedback or review activity with learners' reflections based on a social constructivist design can benefit students' higher-order learning outcomes, improve task outcome quality, and reduce academic staff's workload (Taylor, Ryan, & Pierce, 2015). There is, however, some doubt about out-group peer review quality, due to the often large difference between instructors' and out-group peers' grades (ArchMiller, Fieberg, Walker, & Holm, 2017).

In-Group Peers. For group work, there is a need to distinguish the grading behaviour between in-group peers and out-group peers. It is because that 'in-group favoritism' and 'out-group hostility' are two prevalent phenomena that entail bias or discrimination in human society (Perry et al., 2018, p. 89).

"Product" Indicators: Multiple Assessment Grades related to Group Essay Grades

Multiple different assessment methods should be used to ensure that students with different abilities are assessed appropriately. The assessment measures should be aligned with the course objectives clearly and precisely.

Traditional Cognitive Test Results. Traditional tests examining student acquired knowledge (e.g., by multiple choice items) are the most frequently used assessment measure to address a course' objectives. However, traditional tests are mostly appropriate for assessing lower cognitive skills such as knowledge, comprehension and application skills, although it is an inevitable or necessary measure in national and international large-scale assessment programs (Shepard, Penuel, & Pellegrino, 2018). To ensure validity of creative group work, other measures are more appropriate than traditional tests.

Self-Regulation Behaviours. It is mostly agreed that students' self-regulation relates to positive learning outcomes (Seker, 2016). Formative assessment that focuses on self-assessment is a common technique to measure student self-regulation (Meusen-Beekman, Joosten-ten Brinke, & Boshuizen, 2016). For example, students self-assess their performance or behaviour and then reflect on their learning based on the course or personalized learning objectives (Hsu & Ching, 2013).

Other measures used to assess students' self-regulation are, for example, disciplined student behaviour. Instructors can assess students' submission or management behaviour by asking students to submit their work on time. In higher education, student participation or engagement in the learning environment can also serve as a criterion for self-regulation (Haladyna, 1999). The three aforementioned behaviours (keeping reflection journals, on time submission, and active participation in the class) are signs of online and offline self-regulation behaviours for students higher in education (You, 2016; Thibodeaux, Deutsch, Kitsantas, & Winsler, 2017).

"Place" Indicators: Comments in the Online Group Essay Writing Process

Peer review and comments as a form of learning presence are likely to facilitate the group work process (Shea et al., 2013). However, whether peer review and comments can serve as a process-related indicator for quality of group work products remains unknown in the literature so far. In the context of essay writing or in the academic world of paper writing, negative or one-direction comments may be destructive (Lu & Bol, 2007), whereas peer reviews with dialogues (working as in-group peers) can be effective for improving the quality of essays or papers. It is because student peers have the opportunity to exchange

perspectives and justify their thinking (Nicol, 2010), which help generate good outcomes (Southavilay, Yacef, & Callvo, 2010).

"Process" Indicators: Version History in the Online Group Essay Writing Process

The ability to see the complete version history of an essay is a specific feature in Google Docs. Essay version histories indicate the process of essay writing, which can be fully documented given the present development of technology. Group essay versions documented in Google Docs can be used to study the student writing process and can serve as an important measure (continuous assessment) for grading group essays.

Past research, however, does not mention the effect of the life cycle or history of essay versions. Some likely behaviours of Google Document version history related to final essay grades include: quickly responding to peer feedback, completing the essay long before the due date, and changing text frequently (Southavilay, Yacef, & Callvo, 2010).

Research Question

This study is based on a course designed using SLT and 4PCT, given that a group essay is a collective, creative task. The instructor's group essay grades serve as the basis for detecting process-related indicators. Figure 1 presents the theoretical framework of the course, followed by the related person, product, place, and process indicators. The statistical methods to identify process-related indicators were comparisons between groups with different performances. Concretely speaking, this study aimed to answer the following RQ:

What are the person, product, place, and process (4Ps) indicators that distinguish between high- and low- performing groups, as assessed by the instructor's group essay grade?

Method

Participants

The research participants were 18 undergraduate students registered in a teacher-training course on pedagogies at an academic university in Taiwan. Among the students, 14 students were female and four were male; 16 students were Taiwanese and two students were international students from South Asia; 16 students were third-year, and two students were fourth-year students; 14 students studied education and the other four students studied science, commerce, social sciences, and language, respectively. The students were divided into five groups of 3–4 members. The grouping method was negotiated between the course design and the students' preference.

The Pedagogical and Assessment Design of the Course

Course Objectives

The course objectives (CO) aimed to develop students' competences of (A) knowing and understanding pedagogical theories, design methods, and implementation principles, (B) analysing and applying the pedagogies to related cases, and (C) creating, implementing, evaluating, and reflecting in small groups on the pedagogies used to teach any domain of knowledge or any educationally worthwhile topic. CO (C) invited a higher-order pedagogy to apply the knowledge and skills learned by using pedagogies to address CO (A) and CO (B).

Five Phases

The course was organized into five phases as follows.

Phase 1 aimed at fulfilling CO (A), lasting around 3 weeks. Students learned basic knowledge of pedagogical designs. The teaching methods included the instructor giving lectures and demonstrating likely ways to use the knowledge. The students asked and answered questions.

Phase 2 aimed at fulfilling CO (B), lasting around 5 weeks. The students explained knowledge and/or demonstrated likely ways to use the knowledge. The instructor reported related pedagogical cases to demonstrate how to apply the knowledge.

Phase 3 aimed at fulfilling the creating part of CO (C), creating new pedagogies. Students working in groups reported on their initial creative pedagogical design and wrote the title, literature review, research questions, and method of the group essay as midterm presentations.

Phase 4 aimed at fulfilling the implementing, evaluating, and reflecting part of CO (C). Students implemented the pedagogical design they created in Phase 3 in class, collected peers' learning process or outcome data, analysed the data, and completed the remaining part of the group essay, focusing on the results and discussion sections.

Phase 5 fulfilled the three COs by letting students complete their group essays in three steps. In Step 1, students drafted their initial completed group essay. In Step 2, out-group peers, in-group peers, and the instructor graded the drafts and provided comments. In Step 3, students completed their final group essays based on the grades and comments obtained in Step 2.

Data Collection and Ethical Considerations

The data were collected during regular didactic practices. The course mainly used face-to-face class teaching (14 weeks) but partially used various digital LMSs (e.g., Google Drive during the whole course process, including 4 weeks fully online and 14 weeks blended with the face-to-face teaching).

This study was a teaching evaluation study, with an aim to improve learning and teaching, using data naturally generated from regular teaching practices and

adult students (> 20 years of age), which met the ethical criterion of teaching research and therefore did not need to obtain approval from the ethical committee (Official Document 1040003540 issued by the Ministry of Science and Technology, Taiwan). In addition, the participants' identities and the course name were protected during the entire research process and are not reported in this paper.

Indicators

Detailed scale ranges and descriptive statistics of the indicators are presented in Table 1. Each indicator is explained further as follows.

"Person" Indicators: Different Agents' Grades

Group essay grades were assessed by the instructor, out-group peers, and in-group peers. Students rated each group essay on seven items using a 4-point Likert-type scale, ranging from 1 = *Strongly Disagree* to 4 = *Strongly Agree*. Mean scores were taken for the ratings from out-group peers (i.e., peer assessment) and in-group peers (i.e., group self-assessment), respectively. The seven items were

(1) creative research topic,
(2) educationally beneficial research topic,
(3) clear motivation and literature to address the research topic,
(4) concrete research hypotheses or questions that can be inferred from literature review,
(5) clear description of research methods (on research participants, teaching designs, measures, and data analysis methods),
(6) clear description of results for each research hypothesis or questions including proper tables and figures, and
(7) in-depth discussion of the meaning of each result, including implications for educational practices.

"Reason and/or suggestions?" served as an open-ended question for each of the above seven Likert-type items and the whole group essay. This allowed students to provide comments and suggestions for further revision of the group essay.

"Product" Indicators: Multiple Assessment Grades

Multiple assessment grades and student behaviour records were naturally generated from the course process. These indicators included

(1) students' results of tests on the course content (i.e., pedagogical knowledge) from completing 10 multiple-choice items l (scores 0-10);
(2) submitting on time: This reflected group essay grading management behaviours. The scores combined the degree to which the students completed the task of grading five groups' essays and how punctual the

students were in completing this task (scores 0-10 = 5 group essays * 2 points * 1 being punctual (to 0 = no submission);
(3) reflection journals: students' weekly journals kept for each week (0-18 weeks), and
(4) participation (0-42 hours).

"Place" Indicators: Comments in Online Processes

In-group peers and the instructors gave comments when the group essays were processed online. The Google Docs were downloaded as Word files, the comments on which were extracted to separate Excel files and subsequently coded and analysed (https://www.thedoctools.com/word-macros-tips/word-macros/extract-comments-to-new-document/). The indicators pertaining to comments included

(1) the instructor's total comment count for each group essay;
(2) in-group peers' total comment counts;
(3) groups' comment interaction density, which was calculated as the average of the number of replies to each comment (including the comment itself). To give an example, if a comment received no replies, its reply count was 1; if a comment received 2 replies, then its count was 3; and
(4) the mean (average) of student comment quality. All comments were coded as 1 = *not inviting further action* (e.g., "Marked as resolved"), 2 = *compliments or acknowledge* (e.g., "good job", "thank you", and "OK, I understand now."), or 3 = *substantial opinions for improving essay contents, responding to previous comments, or inviting further actions* (e.g., "Yes, I do have an assumption…"). The coding results were divided by the total comment count of each group. The second author of this paper learned the above coding rule and some examples, based on portions of the comment items that were coded without knowing the first author's coding. Inter-rater reliability was calculated using the formula "items with same code/total items". We obtained a 94% agreement.

"Process" Indicators: Version History in Online Processes

The instructor created an essay template for each group to work on Google Docs. The digital system automatically recorded when and which parts of the group essays were changed. The process generated the group essays' version histories, which were copied and pasted into Excel files and then analysed. The related indicators that emerged included:

(1) counts of the group essay versions generated by the instructor;
(2) counts of the group essay versions generated by the in-group peers;
(3) group essay version interaction density (or weighted simultaneous writing frequencies), which were the numbers of in-group peers who worked at the same time divided by the group sizes. This indicator aimed to detect the density of students working simultaneously; and

(4) student essay version quality, which was graded by the instructor for each student, using the criterion of substantial contribution provided by one student at least once by the midterm and once on the final essay (1 = *no contribution* to 3 = *substantial contribution*).

Data Analysis

The RQ was answered using the Kruskal-Wallis one-way analysis of variance by ranks with the R FSA package. If the Kruskal-Wallis chi-squares were significant, Dunn's multiple comparison tests with p-values adjusted by the false discovery rate method were used by using the R dunn.test package (Dinno, 2017).

Results

Person Indicators

The criterion was the instructor's grades on students' group essay performance, because the instructor was normally the 'major' person or agent in implementing a course and assigning grades for student learning outcomes. As indicated by Kruskal-Wallis test results, there were significant differences between the five groups in the instructor's group essay grades (Kruskal-Wallis chi-squared value (KWx^2) = 17.000; degree of freedom (df) = 4, p = 0.002; Table 1). Dunn's multiple comparison test results further indicated that Group A had a higher grade than Group D and Group E (A>D, E).

Table 1. Indicator Descriptive Statistics and Group Difference Test results for all and Group Samples

Samples		All						Group A*	Group B	Group C	Group D	Group E			
Indicators	scale range	Min.	Max.	Mean	SD	median		median	median	median	median	median	KWx^2	p-value	DunnMC
1.Essay grades by the instructor (the criterion)	7 ~ 28 points	10.000	27.000	22.830	6.071	26.000		27.000	26.000	26.000	23.000	10.000	17.000	0.002	A>D,E
2.Essay grades by outgroup peers	7 ~ 28 points	18.067	23.400	21.673	1.826	22.571		23.400	22.710	22.570	21.140	18.070	17.000	0.002	A>D,E; B>E
3.Essay grades by in-group peers	7 ~ 28 points	18.000	25.250	22.222	2.719	22.667		22.670	24.000	25.250	18.000	21.000	17.000	0.002	C>D
4.Knowledge test grades	0 ~ 10 points	1.000	7.000	3.390	1.852	3.500		5.000	1.500	3.000	5.000	3.000	8.945	0.062	
5.Essay grading management	0 ~ 10 points	6.000	10.000	9.110	1.158	10.000		10.000	9.000	10.000	8.000	8.000	6.053	0.195	
6.Weekly journal	0 ~ 18 times	8.000	20.000	15.833	3.400	17.500		18.000	16.500	17.500	15.000	13.000	4.005	0.405	
7.Participation	0 ~ 42 hours	31.400	42.000	37.694	3.441	38.450		42.000	38.100	38.050	35.650	38.400	2.656	0.617	
8.Instructor comment count	0 ~ N times	14.000	27.000	18.556	5.159	15.000		15.000	20.000	14.000	27.000	15.000	17.000	0.002	D>C
9.Student comment count	0 ~ N times	0.000	101.000	19.278	37.722	2.000		101.000	8.000	2.000	1.000	0.000	17.000	0.002	A>D,E; B>E

10.comment interaction density	0 ~ N times	0.000	3.000	1.541	1.023	1.880		2.340	2.000	1.880	1.640	1.530	17.000	0.002	A>D,E; B>E
11.student comment quality	0 ~ 3 points	1.530	2.340	1.872	0.275	1.375		2.079	1.375	3.000	1.000	0.000	17.000	0.002	C>D,E
12.Instructor essay version count	0 ~ N times	4.000	8.000	5.444	1.464	5.000		5.000	8.000	4.000	5.000	5.000	17.000	0.002	B>C
13. Student essay version count	0 ~ N times	18.000	62.000	39.944	14.550	35.000		35.000	35.000	43.000	62.000	18.000	17.000	0.002	D>E
14.Essay version interaction density	0 ~ 1 points	0.316	0.389	0.341	0.025	0.337		0.389	0.337	0.326	0.316	0.353	17.000	0.002	A,E>C
15.Student essay version quality	0 ~ 3 points	2.000	3.000	2.944	0.236	3.000		3.000	3.000	3.000	3.000	3.000	3.500	0.478	

Note. * The group names are ordered from high to low instructor's essay grades to facilitate readability. KWx^2 = Kruskal-Wallis chi-squared with a degree of freedom = 4. DunnMC = Dunn multiple comparison tests.

The five groups were also different in their essay grades by out-group peers (KWx^2(df) = 17.000(4), p = 0.002; Table 1). Out-group peers' grades not only repeated the pattern of the instructor's group essay grades (A>D, E) but also indicated a detailed difference (B>E). Note that we have ordered the group names in Table 1 from high to low group essay grades (i.e., A>B>C>D>E) to facilitate readability. The ordering mainly used the instructor's essay grades and partially used out-group peers' grades for the two groups whose instructor's grade was the same.

The test results on in-group peers' essay grades also revealed a significant group difference (KWx^2(df) = 17.000(4), p = 0.002; Table 1). However, Dunn's multiple comparison test results showed the result as C>D.

Product Indicators

There was no significant difference between the five groups in students' knowledge test grades (KWx^2(df) = 8.945(4), p = 0.062), essay grading management behaviours (6.053(4), 0.195), weekly journal performances (4.005(4), 0.405), and participation rates (2.656(4), 0.617; Table 1). The results implied that different assessments tended to measure distinct student abilities.

Place Indicators

The students' comment counts and comment interaction density, like the out-group peers' essay grades, not only replicated the pattern of the instructor's grades (A>D, E) but also indicated a more detailed difference (B>E; both KWx^2(df) = 17.000(4), p = 0.002). The instructor comment count showed D>C (17.000(4), 0.002), revealing a sign of an undesirable indicator. The students' comment quality revealed C>D, E (17.000(4), 0.002), showing a correct value order but failing to meet the criterion (i.e. the instructor's grades; A>D, E).

Process Indicators

None of the four version-history sub-indicators replicated the pattern of group difference in the criterion (i.e., the instructor's grade; A>D, E; Table 1). However, there were two group differences revealing a different picture: the instructor's version counts (B>C; KWx^2(df) = 17.000(4), p = 0.002) and the students' essay version counts (D>E; 17.000(4), 0.002).

The essay version interaction density (A, E>C; 17.000(4), 0.002) slightly mismatched the order of grade values assessed by the instructor (i.e. A>B= C>D>E). There was no group difference in the students' essay version quality (3.500(4), 0.478). This might be because all the students contributed to their essays in both midterm and final-term essays. The small class allowed the instructor to make each student a successful completion of the assignment.

Discussion

This study provides theoretical and practical contributions for pedagogy, assessment, and learning analytics: The FIC for theory and the identified effective and ineffective indicators for practice by learning analytics.

The Theoretical Basis: From Course- to Ecology-Focused

SLT (Zeidler, 2016) and 4PCT (Hasirci & Demirkan, 2003) form the FIC (Figure 1), guiding the pedagogical, assessment and research design of this study. The use of the FIC in this study suggests that the FIC contributes to a specific course's pedagogical, assessment, and research design and emphasizes using collaborative learning to generate creative products. Future courses and studies may use the FIC as a base for their pedagogical, assessment, and research purposes, given that this study is a case study, with a small sample size, and its result cannot be generalized in nature.

Limitations and Suggestions. The FIC (Figure 1) has incorporated some indicators proposed by the SLT and the 4PCT. All the indicators, however, are situated in or constrained by the ecological support in relation to a course. For example, process refers to students' behaviours in the process of completing the tasks situated in or constrained by the Google Docs platform. Further, small classes (below or equal to around 30 students) is a typical practice of teacher-training courses in Taiwan. The issue of broad or whole ecological systems to support pedagogical, and assessment design should be addressed by ecological theories relating to information and communication technology (ICT) (Chiu, 2019; Johnson, 2010; Johnson & Puplampu, 2008). Future research may elaborate the FIC by adding and examining a broader scope of ecological indicators.

Effective Indicators

Out-group Peers' Assessment as a Proxy Measure of Instructors' Grades

Out-group peers' group essay grades state the criterion (the instructor's grading) reliably and even more precisely. The results appear to suggest the role of the out-group peers' grades as a proxy, efficient, and even accurate indicator of group essay grades (Taylor, Ryan, & Pearce, 2015) although there exists research indicating that out-group peers' assessment may be ineffective (ArchMiller, Fieberg, Walker, & Holm, 2017).

The reason for the positive role of the out-group peers' assessment in this study may be that this study uses an SLT-based teaching design. Further, the small class allows the instructor an opportunity to fully implement SLT-based teaching (the FIG or ESPA; Figures 1 and 2) by monitoring, scaffolding and catering students' progress and needs thoroughly. The SLT-based design emphasizes social interaction in the course process, similar to a design using social constructivism (Taylor, Ryan, & Pearce, 2015). The SLT designs a pedagogy where out-group peers have active involvement in the other groups' essays as research participants

(i.e., out-group peers serving as the students in the group essays on pedagogies) and conference attendees (i.e., midterm presentations), which may reduce out-group bias (Perry et al., 2018), increase mutual understanding, and generate accurate grading.

Peer Comments as a Golden Rule for Improvement

Peer review has long been an effective practice for advancing academic knowledge, but is embedded with passive and negative effects (Lu & Bol, 2007). This study finds that comment counts and comment interaction density relate to final grades. The results suggest that in-group peers' comments or dialogues serving as a learning presence can have beneficial effects on group essay outcomes (Nicol, 2010; Shea et al., 2013).

Instructors need to have a pedagogical and assessment design inviting students to give more high-quality comments to group tasks. This study focused on two quantitative indicators: comment counts and comment interaction density (indicating active responses to comments), with the latter slightly, but still not completely, addressing the issue of high-quality comments. Although an interactive co-writing digital environment (e.g., Google Docs) makes commenting a convenient practice, the key may still be instructors' pedagogies (including assessment designs). For example, the course of this study was based on sociocultural learning theory (Zeidler, 2016), aiming to cultivate students' deep thinking and engage in interaction for improving their group essays. In addition, the pedagogies require instructors' deep and working content knowledge, pedagogical knowledge, and technological knowledge (Scherer, Tondeur, Siddiq, & Baran, 2018).

Ineffective Indicators

Multiple Assessments for Different Student Abilities

This study corresponds to past research findings that different assessment measures reflect different aspects of students' abilities (Grossman, Cohen, Ronfeldt, & Brown, 2014). Many teachers use group work as only learning activity or formative assessment (without grading) and they tend to use a heavily weighted knowledge test as the final grading method.

Our findings show that the knowledge tests and group works assess different cognitive skills and suggest that group work should be graded to show the attainment of course objectives. The results suggest a need for instructors to identify the course objectives from the perspective of cultivating learners' diverse abilities, design pedagogies to address the objectives and incorporate assessment measures for students to demonstrate their diverse abilities.

Essay Version History as an Ineffective or Uncertain Indicator

Essay version history fails to relate to the instructors' group essay grades. The reasons for this result may be that different individuals have different writing

processes. For example, students may use MS Word more than Google Docs when writing their essays and this in turn reduces essay version history counts (Southavilay, Yacef, & Callvo, 2010).

However, the question remains whether group essay version interaction density (i.e., weighted simultaneous writing frequencies) relate to group essay grades, because this study appears to be the first in the literature to indicate this phenomenon. Future research needs to address and investigate this phenomenon further.

Contributions, Limitations, and Suggestions for Future Research

Contributions

Educational Theories. The FIC combines two theories (SLT and 4PCT) and generates a meaningful pedagogical, assessment and research design, which in turn leads to some successful findings in this study.

Methodology and educational practices. This study successfully identifies three major process-related indicators for group essay grades: out-group peers' assessment, group peers' comment counts, and group peers' comment interaction density.

Limitations and Suggestions

Although it is common that the 4Ps are interwoven, a more comprehensive and elaborated framework still needs to be established in future research. This framework may direct a more detailed pedagogical design. Further, this study used nonparametric statistics due to the small sample size. Future research needs to validate the results using a larger sample size in order to extrapolate the results to similar processes of grading group essays and teaching practices. Thirdly, culture can affect grading behaviours. For example, Taiwanese students are more likely to keep weekly journals than students from other countries (Chiu, 2016). Finally, version history sub-indicators may vary with different stages of group essay writing (Southavilay Yacef, & Callvo, 2010). Future studies may resolve these issues, especially when they are able to use large sample sizes or incorporate qualitative research methods to supplement quantitative ones.

Conclusion

This study evidences that a sociocultural learning theory (SLT) and 4P (Person, Product, Place, and Process) creativity theory (4PCT) combined framework can support the design of effective pedagogies for incorporating diverse assessments into teaching. Out-group peers' assessment is the most proxy measure for instructors' grades of students' group essays. Peer comments are a measure for improving student group essay quality (or grades). Knowledge tests and group works assess different cognitive skills.

For educational practice and policy, implications inferred from the results of this study include that the SLT-based pedagogy allows for out-group peers to have active involvement in the other groups' essays as research participants and conference attendees. This may reduce out-group bias and increase mutual understanding, and generate accurate grading. Instructors need to have a pedagogical and assessment design inviting students to give more high-quality comments to group tasks and actively respond to the comments. Instructors need to identify the course objectives from the perspective of cultivating learners' diverse abilities. Group work should be graded to show the attainment of course objectives.

Acknowledgments

This work was supported by National Chengchi University (DZ15-B4). The funder only provides financial support and does not substantially influence the entire research process, from study design to submission. The authors are fully responsible for the content of the paper.

References

ArchMiller, A., Fieberg, J., Walker, J. D., & Holm, N. (2017). Group Peer Assessment for Summative Evaluation in a graduate-Level Statistics Course for Ecologists. *Assessment & Evaluation in Higher Education*, *42*(8), 1208-1220.

Black, P., & Wiliam, D. (2018). Classroom Assessment and Pedagogy. *Assessment in Education: Principles, Policy & Practice*, *25*(6), 551-575.

Chiu, M.-S. (2016). Engaging Internationally Diverse Students by Integrating the Teaching of Reading and Writing and Using Writing via ICT Tools for Assessment. *Scholars Bulletin*, *2*(11), 625-636.

Chiu, M.-S. (2019). Exploring Models for Increasing the Effects of School Information and Communication Technology Use on Learning Outcomes Through Outside-School Use and Socioeconomic Status Mediation: The Ecological Techno-Process. *Educational Technology Research and Development*, *68*(Aug), 413-436.

Davies, W. M. (2009). Groupwork as a Form of Assessment: Common Problems and Recommended Solutions. *Higher Education*, *58*(4), 563-584.

Dinno, A. (2017). *Package "dunn.test."* Available at: http://cran.stat.unipd.it/web/packages/dunn.test/dunn.test.pdf.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The Test Matters: The Relationship Between Classroom Observation Scores and Teacher Value Added on Multiple Types of Assessment. *Educational Researcher*, *43*(6), 293-303.

Guskey, T. R., & Link, L. J. (2019). Exploring the Factors Teachers Consider in Determining Students' Grades. *Assessment in Education: Principles, Policy & Practice*, *26*, 303-320.

Haladyna, T. M. (1999). *A Complete Guide to Student Grading*. Allyn and Bacon.

Hasirci, D., & Demirkan, H. (2003). Creativity in Learning Environments: The Case of Two Sixth Grade Art-Rooms. *The Journal of Creative Behaviour*, *37*(1), 17-41.

Hsu, Y.-C., & Ching, Y.-H. (2013). Mobile App Design for Teaching and Learning: Educators' Experiences in an Online Graduate Course. *The International Review of Research in Open and Distributed Learning*, *14*(4), 117-139.

Johnson, G. (2010). Internet Use and Child Development: The Techno-Microsystem. *Australian Journal of Educational and Developmental Psychology, 10*, 32-43.

Johnson, G. M., & Puplampu, P. (2008). A Conceptual Framework for Understanding the Effect of the Internet on Child Development: The Ecological Techno-Subsystem. *Canadian Journal of Learning and Technology, 34*, 19-28.

King, P. E., & Behnke, R. R. (2005). Problems Associated with Evaluating Student Performance in Groups. *College Teaching, 53*(2), 57-61.

Lu, R., & Bol, L. (2007). A Comparison of Anonymous Versus Identifiable E-Peer Review on College Student Writing Performance and the Extent of Critical Feedback. *Journal of Interactive Online Learning, 6*(2), 100-115.

Maiden, B., & Perry, B. (2011). Dealing with Free-Riders in Assessed Group Work: Results from a Study at a UK University. *Assessment & Evaluation in Higher Education, 36*(4), 451-464.

Messersmith, A. S. (2015). Preparing Students for 21st Century Teamwork: Effective Collaboration in the online Group Communication Course. *Communication Teacher, 29*(4), 219-226.

Meusen-Beekman, K. D., Joosten-ten Brinke, D., & Boshuizen, H. P. A. (2016). Effects of Formative Assessments to Develop Self-Regulation Among Sixth Grade Students: Results from a Randomized Controlled Intervention. *Studies in Educational Evaluation, 51*, 126-136.

Nicol, D. (2010). From monologue to dialogue: Improving Written Feedback Processes in Mass Higher Education. *Assessment & Evaluation in Higher Education, 35*(5), 501-517.

Perry, R., Priest, N., Paradies, Y., Barlow, F. K., & Sibley, C. G. (2018). Barriers to Multiculturalism: In-Group Favoritism and Out-Group Hostility are Independently Associated with Policy Opposition. *Social Psychological and Personality Science*, *9*, 89–98.

Scherer, R., Tondeur, J., Siddiq, F., & Baran, E. (2018). The Importance of Attitudes Toward Technology for Pre-Service Teachers' Technological, Pedagogical, and Content Knowledge: Comparing Structural Equation Modeling Approaches. *Computers in Human Behavior, 80*, 67-80.

Seker, M. (2016). The Use of Self-Regulation Strategies by Foreign Language Learners and its Role in Language Achievement. *Language Teaching Research, 20*(5), 600-618.

Shea, P., Hayes, S., Smith, S. U., Vickers, J., Bidjerano, T., Gozza-Cohen, M., et al. (2013). Online Learner Self-Regulation: Learning Presence Viewed Through Quantitative Content- and Social Network Analysis. *The International Review of Research in Open and Distributed Learning, 14*(3), 427-461.

Shepard, L. A., Penuel, W. R., & Pellegrino, J. W. (2018). Using Learning and Motivation Theories to Coherently Link Formative Assessment, Grading Practices, and Large-Scale Assessment. *Educational Measurement: Issues and Practice, 37*(1), 21-34.

Southavilay, V., Yacef, K., & Callvo, R. A. (2010). Process Mining to Support Students' Collaborative Writing. In R. Baker, A. Merceron, & P. I. Pavlik Jr. (eds.), *Educational Data Mining 2010* (pp. 257-266). Available at: https://bit.ly/3jWYQIC.

Suwantarathip, O., & Wichadee, S. (2014). The Effects of Collaborative Writing Activity Using Google Docs on Students' Writing Abilities. *Turkish Online Journal of Educational Technology, 13*(2), 148-156.

Taylor, S., Ryan, M., & Pearce, J. (2015). Enhanced Student Learning in Accounting Utilising Web-Based Technology, Peer-Review Feedback and Reflective Practices: A Learning Community Approach to Assessment. *Higher Education Research & Development, 34*(6), 1251-1269.

Thibodeaux, J., Deutsch, A., Kitsantas, A., & Winsler, A. (2017). First-Year College Students' Time Use: Relations with Self-Regulation and GPA. *Journal of Advanced Academics*, *28*(1), 5-27.

Woodrich, M., & Fan, Y. (2017). Google Docs as a Tool for Collaborative Writing in the Middle School Classroom. *Journal of Information Technology Education: Research*, *16*, 391-410.

You, J. W. (2016). Identifying Significant Indicators Using LMS Data to Predict Course Achievement in Online Learning. *The Internet and Higher Education*, *29*, 23-30.

Zeidler, D. L. (2016). STEM Education: A Deficit Framework for the Twenty First Century? A Sociocultural Socioscientific Response. *Cultural Studies of Science Education*, *11*(1), 11-26.

Zhou, W., Simpson, E., & Domizi, D. P. (2012). Google Docs in an Out-of-Class Collaborative Writing Activity. *International Journal of Teaching and Learning in Higher Education*, *24*(3), 359-375.