January 2015

EFFECTS OF COLLABORATION AND ISOMORPHIC MODELS ON TRANSFER: AN L2 WRITING INVESTIGATION

Dennis Mitchell Koyama
Purdue University

PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance

This is to certify that the thesis/dissertation prepared

By Dennis Mitchell Koyama

Entitled
Effects of Collaboration and Isomorphic Models on Transfer: An L2 Writing Investigation

For the degree of Doctor of Philosophy

Is approved by the final examining committee:

Tony Silva
Chair

Margie Berns

David A. Sears

Shelley Staples

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy of Integrity in Research" and the use of copyright material.

Approved by Major Professor(s): Tony Silva

Approved by: P. Ryan Schneider 8/17/2015
Head of the Departmental Graduate Program Date

EFFECTS OF COLLABORATION AND ISOMORPHIC MODELS ON TRANSFER: AN L2 ENGLISH WRITING INVESTIGATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Dennis Mitchell Koyama

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

December 2015

Purdue University

West Lafayette, Indiana

里恵子、僕の人生コラボレーター

For Rieko, my lifelong collaborator

ACKNOWLEDGEMENTS

Completing a long-term project, such as a dissertation, requires a lot of support. There are many people that have helped me achieve this goal, but the usual suspects are indeed the ones I'd like to acknowledge here. I begin with my thanks to the students who participated in this study and to the teachers who allowed me to disrupt their semester calendars with my project—thank you! I am also thankful for Dr. April Ginther, who encouraged me to apply for this program and saw to it that I got everything I needed to begin my career at Purdue. I would like to thank Jill Quirk for helping make every university process happen as quickly and smoothly as possible. I also am grateful to my friends and colleagues who helped me transition to Purdue, and the Midwest winters. I thank Marshall Klassen and HeeJung Kwon for making my first winter warm with conversation and for the many carpools after evening classes. I also thank Scott Partridge for being my home brewing teacher and for sharing his family dinner table on countless occasions. Scott, you and your family will always have a seat at our *kotatsu*. I also extend my thanks to Ghada Gherwash. Ghada, you are an amazing friend. Meeting you and having you as a colleague and friend is a blessing to me and to my family—thank you for your advice and friendship.

I would also like to thank my committee, starting with Dr. Tony Silva, my advisor and chair. Tony, I hope to someday master the art of one-liners as you have. You somehow can compress a very complex idea into an anecdote that seems intuitive, and, in a sentence or two, you can put a smile on any face within earshot. I aspire to be as concise and clear in my advice to others as you have been with me—I imagine anyone who has asked me for advice waits for that day as well. I thank Dr. Margie Berns for her support and suggestions for my project. Margie, your keen eye for sharpening the details of a theory while keeping the big picture in mind is a skill many seek to master, myself included. Thank you for always having an open door, great conversations, and kind words. I would also like to thank Dr. Shelley Staples for joining my committee. Shelley, you have provided me with useful feedback, expert advice, and part-time work all within the first semester of having met you. Your affable nature is infectious, and your advice is genuine. Thank you for being my friend and colleague. I would like to express my sincere gratitude and thanks to Dr. David Sears. David, your research is inspirational, and your insightful comments are equally powerful. It is not possible to summarize the affect your course, *Collaborative Learning,* had on my research. I can say without hesitation, though, that this dissertation (and my current research program) would not exist had I never taken your course. Thank you for taking time from your work schedule and your beautiful, growing family to spend time with me and my endless string of questions. Finally, I would like to thank, Rieko Koyama. 里恵子、君に

出会えて本当に良かった。君のサポートは無限であること、言葉にならないほど感謝しています … 愛してます。

TABLE OF CONTENTS

Page

ABSTRACT

Koyama, Dennis Mitchell. Ph.D., Purdue University, December 2015. Effects of Collaboration and Isomorphic Models on Transfer: An L2 English Writing Investigation. Major Professor: Tony Silva.

How can feedback become a productive resource for students? Much of the research investigating the role of feedback in second language (L2) writing has set out to find an answer to this question. Based on the principle that feedback is given to students as a means of providing useful information to improve their writing (Bitchener, 2008, 2009; Hanaoka & Izumi, 2012), the discussion on feedback includes the idea that learners will transfer knowledge from feedback to improve subsequent writing (Hyland, 1998; Storch & Wigglesworth, 2010). When learners apply feedback to their subsequent writing, they are using collected knowledge, which is the essence of learning transfer (Schwartz, Bransford, & Sears, 2005). Unfortunately, no method of writing feedback has been deemed the frontrunner for improving learner texts (Ferris & Roberts, 2001; Hyland & Hyland, 2006; Storch & Wigglesworth, 2010) or for helping learners transfer writing knowledge across writing situations (James, 2006a,b, 2008, 2009, 2010). While this outlook may seem bleak for writing instructors, recent research provides evidence for presenting learners with expert models as a fruitful way of offering feedback.

Language researchers have shown that using models can provide learners with opportunities to engage with the language in the model, encouraging them extract useful language from it (Hanaoka & Izumi, 2012). Naturally, the more language the learners notice in the model, the more likely they will recall that language and content elements at a later time (Schmidt, 1990; Schmidt and Frota, 1986). To increase this chance for noticing language, Watson (1982) suggests learners discuss the model in pairs or small groups. Other L2 researchers share Watson's interest in collaboration in L2 writing classrooms and advocate for engaging learner collaboration in all stages of composition. This collaborative approach identifies learners as a language resource via spontaneous peer feedback (Fernández Dobao & Blum, 2013; Storch, 2005, 2013; Watanabe & Swain, 2007; Wigglesworth & Storch, 2010). Collectively, these perspectives suggest feedback can be crafted by moderating the written form of feedback (e.g., presenting expert models) and by incorporating learner-to-learner interaction (e.g., collaborative tasks). Investigating factors such as these might provide theoretical and practical insight into how learners transfer feedback.

Educational researchers have consistently shown the benefits of collaboration on learning outcomes (Dillenbourg, 1999; Johnson & Johnson, 1999; Slavin, 1983) and transfer (Pai, Sears, & Maeda, 2014; Sears, 2006). However, these positive outcomes do not come automatically, and numerous researchers caution that learning gains are only found when appropriate scaffolds are in place (Janssen, Erkens, Kirschner, & Kanselaar, 2010; Johnson & Johnson; O'Donnell, 2006; Slavin, 1996; Webb et al., 2008; Weinberger, Ertl, Fischer, & Mandl, 2005). Recent research adds to this caveat by positing that task complexity and demonstrability (more is said on this later) may also play a role in the efficacy of collaborative learning tasks (Sears & Reagin, 2013).

Learning collaboratively has several features that distinguish it from learning independently. For example, learning in groups or pairs (also called *dyads*) has shown that the ways in which learners interact are key to promoting learning (Storch, 2002a, 2013; Webb, 1992; Webb & Farivar, 1994). Researchers have shown how the cognitive processes of explanation, elaboration, and clarification evident in student interaction play a role in supporting durable learning (Webb; Webb & Farivar). Researchers have also shown that learners gain a deeper understanding of the material through learners giving and receiving explanations (Storch, 2013; Webb), and from building on the ideas of others in a group (Berkowitz & Gibbs, 1983; Kruger, 1992). The resiliency of learning in groups is important to education because it prepares learners to apply that knowledge to other situations (i.e., transfer their knowledge). Despite this potential boon of learning transfer, however, some researchers (e.g., Detterman, 1993) claim the transfer of knowledge is rare if ever possible in education castigating it as, at best, a moving target, and at worst, nonexistent.

A reflection of this latter sentiment can be seen in a call for taking an alternative perspective to learning transfer (e.g., Bransford & Schwartz, 1999; Larsen-Freeman, 2013). Specifically, Bransford and Schwartz call for a reconsideration of a “sequestered problem solving” (SPS) view of transfer, often associated with teaching and assessments that require direct applications of rote learning. Instead, their research petitions for a renewed understanding of transfer that extends learning assessments and activities to a forward-looking definition learning, called, "preparation for future learning" (PFL; see also Schwartz et al., 2005).

In the PFL framework, collaborative interactions engage learners in *innovation tasks* that push learners to reach beyond their individual dispositions and to build knowledge and understanding through each other's contributions. Specifically, innovation tasks necessitate the use of prior knowledge in learner attempts to construct, or at a minimum offer, solutions to problems that are unfamiliar or unknown to them (Sears, 2006). A primary objective for having students innovate their initial solution is to prepare them to perceive and appreciate an expert solution. This approach is opposed to a more traditional *efficiency* model of learning tasks that embroils canonical solutions that can be rote and practice for mastery without generating novel ideas or solutions (Sears). The PFL framework posits that the innovation of ideas and solutions through dyadic interaction, with each other and with materials, can be a valuable catalyst for learning transfer of the deep structures of knowledge that apply across multiple contexts. Evidence of transfer for deep structures within the PFL framework can be seen by assessing learner performance on complex tasks (often called, *target transfer tasks*) that were not contained in the learning materials of a task, and not a focus of the individual or collaborative tasks (Schwartz et al.). This change in perspective from SPS to PFL has important implications for how language classrooms might design, employ, and measure effective learning tasks.

This dissertation explored the usefulness of an expert model and a structured task in an L2 writing classroom. Two interaction levels—individual and collaborative—

were examined for their facility of descriptive language related to data integration of graphical information from model feedback in a controlled pre/posttest experiment with international university students enrolled in an L2 English composition course. Two approaches to coding the data were taken. The first approach employed a coding scheme that provided a percentage of content overlap with the expert model—an indicator of factual recall and transfer. This was done by a line-by-line coding scheme (Glaser, 1978). The second approach considered how well the essays "fit" the expected data integrations provided in the model—an indicator of transfer of deep writing structure based on the relative balance of global versus local integrations. This was calculated with a Chi-square test of fit. The transfer of deep structures was further measured through an analysis of if students could identify a data interaction that did not exist in the model description. The results showed that learners in the dyad condition significantly outperformed learners in the individual and control conditions on content overlap and expected data integrations. The dyad condition also surpassed a truth-wins comparison, which provides a comparison of actual dyads to the theoretical pooling of knowledge individuals (Lorge & Solomon, 1955), and dyads were the only condition to include the target transfer item in their posttest revisions, indicating dyads were able to understanding complex data integrations in ways not available to learners in the individual and control conditions.

CHAPTER 1. INTRODUCTION

1.1 Introduction

Integrating additional sources of information into writing has been identified as an essential skill for the academic success of university students (Campbell, 1990; Leki & Carson, 1997). Emphasizing the importance of this skill for language learners, Horowitz (1986) documented how multiple sources of information of language input (e.g., lecture notes) or visual input (e.g., graphs) are commonplace in university lectures and assignments. This means teaching L2 learners how to write about graphs and data is essential to helping them achieve academic success in their course work and in their higher education careers.

Carswell, Emery, and Lonon (1993) found that learners interpreting graphs consistently used language to identify *global* integrations of data (e.g., an overall trend) and *local* integrations of data (e.g., identifying sub-groups of data). This aspect of graph descriptions is one reason direct corrections to learner essays may not be productive for learning how to write about graphs, as corrections may include both global and local integrations. For example, if a teacher encounters the statement, "data is more up" does this mean there was an increase in one data set or an increase in a trend across data sets. Specifically addressing an error when the source of the error is not clear might be more confusing than helpful. Direct corrections have other repercussions, too. Ferris (2006) noted, for example, that direct corrections to learner essays are often uncritically incorporated by learners who do not reflect on their original mistake. This deprives learners of the learning process found in the cyclical stages common to writing assignments in school work (e.g., rough draft, revision from feedback, and final draft). This suggests that writing tasks investigating feedback should be designed into multiple stages (herein, *multistage writing tasks*) because they mirror what learners do in classroom activities (Leki, 1998; Zamel, 1983). Furthermore, research on collaborative learning notes the common use of group and pair work across classrooms and curriculums, identifying that students learn more and remember more when they share ideas in order to understand new information (Johnson & Johnson, 1999). Therefore, a fruitful line of inquiry might blend these benefits through an investigation of feedback through multistage writing tasks with collaborative conditions.

1.2 Feedback studies in second language writing

The field of second language writing (SLW) is flourishing with feedback studies, yet no method has been deemed the frontrunner for improving learner texts (Ferris & Roberts, 2001; Hyland & Hyland, 2006; Storch & Wigglesworth, 2010), or for helping learners transfer writing knowledge across writing situations (James, 2006a,b, 2008, 2009, 2010). For some researchers, the trouble with receiving feedback is not in the content of the feedback, but in the possible negative motivational outcomes of feedback (Hedgecock & Lefkowitz, 1994; Hyland, 1998; Truscott, 1999). These researchers note how negative feedback can shortcut learning because learners adopt avoidance strategies such as deleting instead of revising errors, or by blindly accepting feedback without considering the cause of the original error (Ferris, 2006; Kepner, 1991; Semke, 1984; Sheppard, 1992). These unfortunate potential outcomes of feedback run counterintuitively to the assumption that students use feedback for constructive evidence to help them improve their writing (Adams, 2003; Hanaoka & Izumi, 2012). Despite these possibilities, however, many L2 researchers deem feedback a worthwhile enterprise.

One of the primary reasons teachers and researchers advocate for written feedback is it provides learners with the support to see how the writing can be improved (Hyland & Hyland, 2006; Ferris, 2011). This ability to examine language output, Williams (2012) suggests, allows learners to reflect and revisit feedback multiple times. This process of interpreting written feedback provides learners with opportunities for noticing areas for improvement that may lead to learning (Williams). Among the different modes of feedback, some L2 writing researchers (e.g., Hanaoka & Izumi, 2012; Yang & Zhang, 2010) have identified the use of expert *models* as a useful feedback method that scaffolds learners towards noticing and uptake of language and content. Before reviewing the L2 writing literature on expert models and collaborative learning, two SLA (second language acquisition) theories warrant some attention: The Noticing Hypothesis (Schmidt, 1990), and the Output Hypothesis (Swain, 1995). These two theories have played foundational roles in the underpinning of the design of several of the studies discussed in the coming sections. They are briefly touched upon here.

1.3 The Noticing Hypothesis, and the Output Hypothesis

The Noticing Hypothesis. For several years now, many language researchers (e.g, Robinson, 1995; Schmidt, 1990, 1995, 2001; Schmidt and Frota, 1986; Skehan, 1998) have come to the accord that noticing language features facilitates language learning. Schmidt (2001), defined noticing as "a very low level of abstraction . . . assuming that the objects of attention and noticing are elements of the surface structure of … the input – instances of language, rather than any abstract rules or principles" (p. 5). This means learners can be made aware of, or pushed to *notice*, specific language features—which has been called, "noticing the gap" (Robinson, 1995; Schmidt, 2001; Schmidt & Frota; Swain, 1995)—through interaction or collaboration with others (e.g., Long, 1996), or through enhanced input or scaffolding (e.g., Pica, 1994).

The Output Hypothesis. Evidence for the role of noticing in language acquisition has been enriched by Swain (1985, 1995, 1998, 2002) who investigated the role of language production (i.e., output) on a learner's ability to notice the gap. Swain offered the rationale that "output pushes learners to process language more deeply (with more mental effort) than does input" (p. 126). The Output Hypothesis fits well with one aim of written feedback, which is to draw learners' attention to (i.e., notice) issues so they can improve their subsequent writing (i.e., revised output). It also fits with the aim of innovative tasks in collaborative learning conditions, which ask learners to produce responses to complex problems (i.e., pushed output). While noticing can arise from internal self-reflection (e.g., Schmidt & Frota, 1986) or through external feedback (e.g., Hanaoka, 2007; Storch, 2013), paramount to the learning process is that learners are pushed to modify their output. This modified output includes both amendments for errors or style issues in current output, and the incorporation of new language forms and features learners notice from feedback and interaction. Through the process of outputting, noticing, and then modifying output, Swain and Lapkin (1995, among others) claim learners engage in mental processes that are said to play an integral role in the process of language learning..

CHAPTER 2. FEEDBACK

The use of feedback in the L2 writing classroom is a common instructional strategy (Biber, Nekrasova, & Horn, 2011; Hyland & Hyland, 2006; Ferris, 2011). In a comprehensive review of more than 200 L2 writing studies, Hyland and Hyland reported an unfortunate lack of consensus across study findings. They discussed the reality that several of the most important issues to L2 practitioners (e.g., the efficacy of various feedback methods) remain unanswered (Hyland & Hyland, p. 96). More recently, Biber et al. reviewed 306 L2 writing research articles and meta-analyzed a subset of 172 effect sizes from those studies. One facet they investigated was the source of feedback given to learners. They categorized studies by whether the feedback was provided by the teacher, by other students, or by computer-automated feedback (e.g., grammar errors indicated by word processing software). Their results showed that across pretest/posttest, and treatment/control comparisons, 37 studies (or, 79%) focused on feedback given by the teacher (Biber et al., p. 39). This finding identifies a conspicuous dearth in SLW research in terms of investigating how learners are affected by feedback generated by sources other than the teacher, such as learners interacting with model feedback.

2.1 Reformulations and models as feedback

While some researchers have investigated error correction as a positive means for students to improve their grammatical accuracy (e.g., Ferris, 2006, 2011), other researchers identify corrective feedback as discouraging (e.g., Hedgecock & Lefkowitz, 1994; Truscott, 1996, 1999). A few researchers have adopted reformulated texts of student writing because it avoids the error-covered markings that have been found to demotivate learners (Adams, 2003; Qi & Lapkin, 2001; Yang & Zhang, 2010). This form of feedback, called *reformulation*, gives the student's original text to a native-speaker to rewrite in such a way as "to preserve as many of the writer's ideas as possible, while expressing them in his/her own words so as to make the piece sound native-like" (A. Cohen, 1989, p. 4). With the rewriting of the student's work, reformulations can be seen as a method of providing learners with whole text feedback (Adams; Qi & Lapkin; Yang & Zhang). The following is an example of a reformulation.

Example 1.

> **Original**: One day, they have a dinner and meeting because their mother want to share her's wealth after dinner the old mother think they need share all of the family wealth.
>
> **Reformulation**: That day, the family had dinner together at the mother's home. After dinner, the mother wanted to have a family meeting to discuss how to divide her money among her three sons. (Qi & Lapkin, 2001, p. 302)

While Example 1 offers grammar amendments along with content changes, students using reformulations often complain they do not understand why the changes were made. In early research on reformulations in the language classroom, Allwright, Woodley, and Allwright (1988) explain that polishing a text involves changes at multiple levels, (e.g., syntactic, lexical, pragmatic), so they suggested learners in a class setting receive supplemental support through discussion of reformulated texts. Despite having learners discuss their reformulated texts, some researchers noted that learners still struggled to understand the feedback. For example, Yang and Zhang (2010) interviewed 10 EFL students who expressed their frustration with reformulations:

Example 2.

> **Wang**: I hope that the reformulator could tell us why he reformulated our writing this way.
>
> **Han**: It might be better if the reformulator could give us some comments under each reformulation. (both excerpts from Yang & Zhang, 2010, p.479)

These interview comments give the impression that reformulations can be more frustrating for students than they are useful. Allwright et al. (1988) posited that one down fall of reformulations keeping faithful to the learner's intention is the reformulated text may not represent "a good sample of native writing" (p. 254). Thus, the reformulation can be only as good as the reformulator can predict what the original intention of the writer was. This led some researchers to point out that when a reformulator misunderstands or misrepresents the original meaning in the learner text, an *appropriation* effect occurs (Tardy, 2006). The potential for learner confusion resulting from reformulations caused other researchers to propose that learners receive a *model* text that has been tailored to the task (e.g., Hanaoka, 2007). In doing so, they suggest models could eliminate appropriation and learner bewilderment.

Researchers claim that through models, learners can notice language as they analyze then emulate the model in their subsequent writing (Eschholz, 1980; Watson, 1982). However, some critics voice concern that models create an opportunity for mimicry. Eschholz stated, providing models can limit creativity and purpose because “a form is assigned before they know what they want to say’’ (1980, p. 36). To address this concern, Watson suggested that teachers ask learners to write a draft before they look at a model essay. This intentional postponement of feedback may push students to produce new language in their revisions as they consider solutions from hints in the model. Used in this way, models fit well with Second Language Acquisition (SLA) theories of language learning (e.g., noticing and output) as they become a part of the process of composition and a language resource for comparison. This means research on models can inform both practical and theoretical perspectives.

From a practical perspective, current research (e.g., Hanaoka, 2007; Hanaoka & Izumi, 2012) shows positive findings in students’ changes to their writing using models as feedback, which speaks to the pedagogical tenor of giving learners feedback. Another reason research using models as feedback might inform pedagogy is the efficiency of creating models. It will take far less time for an instructor to generate one or two models for students to use than it will take the same instructor to explicitly correct or reformulate each student’s paper (Ferris, 2011). Furthermore, models are less intrusive than direct correction or reformulation, which some researchers identify as discouraging for students (Hedgecock & Lefkowitz, 1994; Hyland, 1998; Truscott, 1999).

From a theoretical perspective, if language, structure, or content are incorporated from models there could be implications for SLA theory and for discussions on learning transfer. However, if learners who merely complete revisions without models yield similar language changes as those who used models as a resource (i.e., task repetition without feedback yields the same results as receiving feedback), the results of such research will not provide evidence for the role of models in promoting noticing and subsequent output in second language learning. This finding would also contribute to discussions on the role of noticing and output in SLA, if not to the claim that transfer in L2 writing classrooms is fragile.

2.2 Models and multistage writing tasks

One way to investigate the diverse, and sometimes divergent outcomes on feedback studies, is through alternative writing tasks and learning theories that might contribute a renewed understanding of how and why SLW research results vary. Within the past decade, a promising line of L2 writing investigations has become increasingly more popular, multistage writing tasks (see Table 2.1 for a description). Driven by the notion that writing is not a fixed or one-shot event, some researchers have examined L2 writing as a cyclical process (Raimes, 1985; Zamel, 1983, 1985). The perspective that writing occurs in cycles (sometimes called, stages) can be a useful way to investigate L2 writing because it approximates the writing process and sequenced writing formats commonly used in L2 writing classrooms (e.g., Leki, 1998; Zamel, 1985). Several L2 writing researchers (e.g., Adams, 2003; Hanaoka & Izumi, 2012; Qi & Lapkin, 2001; Yang

& Zhang, 2010) have chosen to operationalize the stages into three parts, (1) composition stage, (2) feedback stage, and (3) revision stage. It should be noted that the representation of the writing process in this linear manner is for sequential accuracy and not for descriptive precision. Indeed, multistage writing studies have shed some light on how learners use written feedback to solve a range of language issues, depending on the mode and purpose of the feedback (e.g., Adams; Hanaoka, 2007; Hanaoka & Izumi; Sachs & Polio, 2007; Tocalli-Beller & Swain, 2005; Yang & Zhang).

Table 2.1. *Description of Stages in Multistage Writing Tasks*

Stage	Writing event
1	Learners write on a given prompt. This has been done in dyads and individually, depending on the purposes of the stages that follow this stage.
2	Learners are given feedback. This has been done alone or with peer/tutor/instructor as a partner. Sometimes, feedback occurs through model essays, or by presenting learners with a corrected version of their original essay written in stage one.
3	Learners make revisions. These revisions are sometimes done individually or in dyads. Furthermore, some studies have asked learners to write individually in this stage, though they were in dyads in stage one and vice versa.

Note. Some studies use a fourth stage, which correlates with a delayed posttest.

Why models. To date, only a few empirical studies have explored models as a feedback in the L2 classroom. The primary studies have been in EFL environments, such as Japanese university students (Hanaoka, 2006a, 2006b, 2007; Hanaoka & Izumi, 2012), Chinese university students (Yang & Zhang, 2010), primary school learners in Spain (e.g., Martinez & de Larios, 2010), and with university-level ESL students learning the past hypothetical conditional[1] (Izumi, Bigelow, Fujiwara, & Fearnow, 1999; Izumi & Bigelow, 2000). These studies have compared how learners use models and reformulations (Hanaoka; Hanaoka & Izumi; Yang & Zhang). Together these studies have contributed insight into the effectiveness of models to promote writing change.

In an early series of investigations into the influence of models on noticing target language forms, two studies by Shinichi Izumi (Izumi, et al., 1999; Izumi & Bigelow, 2000) showed writing tasks that used model feedback were no more facilitative of noticing than reading comprehension tasks. The targeted specific grammatical feature, the past hypothetical conditional, was the focus in both studies. Both study designs are quite complex to describe, but relevant to the current dissertation is how the models were used and what their effects were. I will only review phase one of Izumi et al (1999) because it is most relevant (each study had two phases, they were the same but counterbalanced). For the treatment group, phase one began with a reading comprehension task that asked learners to highlight important information in the text. Highlighting was evidence of noticing the target form, and was an outcome measure. The reading text was the model, and it provided the forms needed to complete a text reconstruction task that followed the reading/highlighting stage. After completing the text reconstruction task, the learners read and highlighted another text with the target forms and reconstructed another text. This series of input, output, input, and output led to no significant difference in the noticing of the target forms when contrasted with

[1] An example of this structure is, “If I had gone to the movies, I would have eaten popcorn.”

the comparison group. The comparison group did not have a text reconstruction task, so they did not have an output stage. Instead, they answered comprehension questions about the model. For this group, they read the model and highlighted terms, answered comprehension questions, read and highlighted the model again, and answered comprehension questions. In this phase, the noticing of the past hypothetical conditional was the same whether learners answered comprehension questions after reading the model or wrote a structured essay after reading the model. These results are interesting from the perspective of the Output hypothesis. According to Izumi et al (1999) explicit outputting through writing was no more helpful in pushing learners to notice grammatical structures than the highly internal act of answering reading comprehension questions.

For Izumi & Bigelow (2000) the findings were largely the same as Izumi et al (1999). Both studies noted that writing output and reading comprehension did not have significant differences in pushing learners to notice target grammar forms. With that result, they also noted that the text reconstruction task seemed to focus learners more on the target forms than the learner generated writing. This meant that the differences between learner output and the model might have minimized opportunities for learners to identify target forms in the model (Izumi et al.; Izumi & Bigelow). Thus, their results did not provide support for the Output hypothesis, and failed to provided evidence supporting the use of models as a mode of grammar feedback.

To find out how learners might use models and reformulations differently, Hanaoka (2006b) conducted a case study of two Japanese EFL university students (one “more proficient,” the other “less proficient,” p. 167). Using a multistage writing task and a think aloud protocol, he investigated the use of two models (to avoid a mimicry effects, and to provide more coverage for language input), and a reformulation in the revision of learner texts. The stages of his study were: (a) compose, (b) review the two models and the reformulation and compare them to the original text, (c) revise the original writing. Hanaoka found that each learner used models and reformulations in their revisions but in different ways. The higher proficiency learner would actively search the model for unique information, such as alternative phrases and synonyms for their original text. In the Stage 3 writing, she incorporated nearly double the language (9 tokens) from the models than from the reformulation (5 tokens). The lower proficiency learner made equal amounts of changes with four tokens from each source. Interestingly, across the two learners, 14 tokens could have come from either of the models or the reformulation. This is an interesting finding because it identifies that learners can notice language they want to use from a model equally as well as they can in a reformulation. However, that should be viewed in light of Hanaoka’s finding that across all language noticed, 74% was lexical.

Hanaoka (2006b) also noted that learners used avoidance strategies in their initial writing, such as using their native language instead of attempting an English vocabulary term. In fact, about 30% of the problematic features in their original texts were covered by these strategies. This is an important notation because if an error does not exist, a reformulator cannot address it. Since Hanaoka used a think aloud protocol, he was able to identify these instances. Hanaoka labeled these problems “covert” as they had been veiled by avoidance tactics. Hanaoka compared the types of corrections and additions made from the two feedback forms. He found that both learners incorporated expressions and new content from the model passages but used reformulations to correct their explicit linguistic errors. This finding led Hanaoka to conclude that while both content information and language information may be gleaned from models, reformulations may help learners to noticing the gap in their grammar. This may shed some light on why Izumi et al. (1999, and Izumi & Bigelow, 2000) did not find positive results for their grammar-focused studies that employed models.

Adding to the findings of Hanaoka (2006b), Hanaoka (2007) used a delayed posttest design (i.e., a fourth stage) to identify what difference may exist for Japanese EFL learners noticing and output on immediate revisions (i.e., Stage 3) and on a two-month delayed revision. Hanaoka investigated 37 Japanese university students' writings and revisions to a two-panel picture prompt. In Stage 1 students wrote a response to the prompt, then they received two model responses for the same prompt. As in Hanaoka (2006b) two models were used to maximize possible solutions and to prevent learners from mimicking a model from memory. Distinguishing this study from the 2006b study, Hanaoka incorporated note taking as the mode of noticing rather than think aloud protocols. For the note taking stage, learners took notes on useful information from the models, and then they rewrote their essays. Stage 1, 2, and 3 occurred in a single class period, while the delayed posttest occurred after summer break, two months later.

Hanaoka (2007) asked learners to reflect on their Stage 1 essays while looking at the Stage 2 models. This allowed learners to directly compare the model to their own writing, and it allowed them to make specific notes on how something noticed in the model might help them in their revisions. The results indicated that learners had a strong preference for using the models for lexical features in their subsequent revisions. Over 90% of noticed language was lexical (out of 162 tokens), and content made up one third of those items. He also found that learner proficiency played a role in what learners noticed and incorporated; specifically in both the models and their original writings, the more proficient learners noticed significantly more features than less proficient learners did. This is an important finding in light of the delayed posttest revisions, which identified that learners retained the solutions longer if they noticed the error in the initial stages. This means more proficient learners may have an advantage in using a model for improving their writing over time, and research investigating models should consider proficiency as a confound for uptake.

One study contributed a perspective on using collaboration in a multistage writing task. Yang and Zhang (2010) used a three-stage writing task with the first two stages being collaborative, and the final stage being individual writing. Their study provided evidence for both reformulations and models as useful tools for learners who collaboratively construct the Stage 1 writing and view The Stage 2 model together. To show this, Yang and Zhang audio recorded 10 EFL students (put into five pairs), who worked collaboratively for one hour on composing a response to the classic multi-panel picture prompt “The tricky alarm-clock” (adopted from Lapkin, Swain, & Smith, 2002).

One week later, in Stage 2, each pair collaboratively looked at a reformulation of their text and a model essay. Each pair received one reformulation, which they read at the start of the class, and a model response to the prompt, which they read during the last half of class. During this time, the learners were asked "… to compare the two versions of the story and find all the differences they could, discussing together why they believed that changes had been made … [then] discuss the differences between their original text and the model." (p. 469). During these interactions, the pairs were audio recorded. The findings showed that students noticed the most differences between the reformulation and their original text, and noticed only a few between the model and their original text. Yang and Zhang also identified that 38% (p. 471) of the talk centered on content related issues (such as clarifying the content or generating ideas) and 62% focused on grammar. In the final revision stage, despite looking at feedback collaboratively, learners wrote individually. This change in writing styles from pretest to posttest makes conclusions difficult to generalize, but the primary findings highlighted the fact that students desired explanations for reformulations, and that most over half of the necessary changes (55%) could have been made from either feedback mode. Adding to this, they found no significant difference between the percent of correct solutions incorporated from reformulations or models. In this way, the study contributed support for the use of models in terms of possible solutions to errors, and provided evidence for using feedback in collaborative conditions.

As Yang and Zhang (2010; Hanaoka, 2006b) have reported, learners appeared to be able to identify their original errors more easily in reformulations than in expert models. They also remarked, though, that models were seen by learners as helpful in providing alternative vocabulary terms and expressions that were not present in their own writing. The notion of the overlapping errors found in the original source text and the content of reformulated and models texts was meticulously investigated by Hanaoka and Izumi (2012). Using a three-stage writing task with 38 freshman Japanese university students, they investigated how reformulations and models address learner errors. They focused their attention to how the information in the model or in the reformulation would address errors that were *overt* (an errors the learners notice in their own writing), and *covert*, which are errors avoided by the learner (Hanaoka & Izumi, 2012, p. 332). This study also had students compare their original writing to the two feedbacks given in Stage 2, a model passage and a reformulation. Hanaoka and Izumi found that of 91.8% (123 tokens) of *overt* errors were addressed by the reformulation, and only 52.2% (or, 70 instances) were solvable with the model (p. 332). This difference was statistically significant; however, they also found that models enabled learners to locate solutions to significantly more *covert* problems than possible from the reformulation. This finding, taken with Zhang and Yang's (2010) result that either models or reformulations generate equal amounts of acceptable change, it would seem that model might be an excellent way to provide feedback addressing both overt and covert errors. Only two studies have investigated primary school learners, Guirao et al. (2015), and Martinez and de Larios (2010). As both of these studies use learners who have near zero learner abilities, they will not be reviewed here. It is important to note, however, that in both the Guirao et al. study and the Martinez and de Larios studies, the authors stated that when presented with the model and their writing, learners immediately focused on surface errors. They began a “spot the differences activity” (Guirao et al., 2015, p.70). This finding supports other research that demonstrated a proclivity for learners to focus on lexical and other surface features when using models as feedback.

Despite differences in study designs, populations, feedback types, and methods, these studies provide a nice synthesis of information for researchers and teachers alike. While the combined outcomes and findings of the studies reviewed cast a positive light on both reformulations and models in the L2 writing classroom, they also offer several points to consider when deciding on which to use. The collective results identify that providing solutions to students through reformulated texts or models will be sufficient for them to notice and utilize selected parts of the feedback in subsequent writing opportunities. This is strong support for whole text feedback as viable method to encourage text revisions. However, the target of what is to be transferred from the feedback to the subsequent writing is important when selecting models or reformulations. While Hanaoka showed that models *could* provide information for grammar errors such as covert errors, models were primarily useful for lexis and content changes. He also found that learner proficiency played a role in what learners noticed and incorporated. This was echoed by several of the studies as well.

As a whole, the studies generally pointed to reformulations as a useful way of directing learners to their original surface errors, especially the overt errors, and that models may help learners with language constructs beyond grammar errors. While teachers may feel like these ideas lend themselves to practical application, these findings need to be considered with two ideas in mind. Firstly, the prompts used in all the studies—though different in content and structure—were picture prompts. This static image-based representation of a snapshot in time, by their very nature, pushes learners to identify *surface features* in the frame. For instance, if leaners were looking at a prompt with a girl and a boy playing with a dog in a park. The learners might make assumptions about the relationship of the children. For example, the learners might state that the two children are friends, neighbors, or siblings. While any of these relationship labels might be possible, learners are left to make assumptions about the picture prompt. Notably, too, the learners in all the studies had no background knowledge shared on the topic or focus of the picture description tasks. This may have been why tasks limited the generation of novel content language or ideas. Thus, it is not surprising to see that learners stuck to surface features in their noticing stages. Secondly, in most of the studies, the learners looked at their original essays while looking at the feedback. While this may seem like an intuitive pedagogical move, it leads learners to compare the original text to the feedback rather than focusing on what they would like to take from the feedback regardless of their original production. This is a point relevant to the discussion because some L2 researchers have noted that when faced with feedback or the challenge of editing texts, learners will often start with—and maintain focus on—the words (J. Williams, 2001, p. 338). This implies that changes in the grammar of a text are not as salient as lexical alterations. Here again, it may not be surprising that the conditions in which learners were audio recorded (Hanaoka, 2006b;

Guirao et al., 2015; Yang & Zhang, 2010) learner talk focused on surface features for both the feedback and their original compositions.

Taken together, these studies on written feedback highlighted the potential usefulness of investigating feedback through multistage writing tasks. They also suggest some uses of models in L2 writing classrooms, such as support for deep structure concepts such as covert errors and cohesion rather than surface features. The studies have provided some input on what to avoid when using models, such as collaborative tasks where "spotting the difference" becomes the task. This brings us to a more detailed discussion on collaborative learning research.

CHAPTER 3. COLLABORATIVE LEARNING

According to collaborative research in the L2 writing classroom, interactions may bring opportunities for idea sharing and question asking that foster language learning (Storch, 2013). In terms of the research investigating collaboration with feedback, Yang and Zhang (2010) bring up some important concerns for study design. Their findings showed most learners focused on lexis and other surface errors. This might have been connected to the directions given to the student. While learners were told to “consider why the changes were made,” interview data showed that learners often struggled with understand why changes were made in the reformulation (recall Example 2 above, from Yang & Zhang, 2010, p.479). This is a real weakness for their study as one purpose of the task was precisely that, to explain why the changes were made. This fact reduces the difficult task of “find and explain the differences” to the simple task of “find the differences.” Simple tasks put to collaboration have been repeatedly shown to generate process loss (Steiner, 1972), and to create superficial understandings of the materials (Phelps & Damon, 1989; Sears & Reagin, 2013).

Time and again, collaborative research has shown that easy tasks put to groups for collaboration often produce *social loafing* effects (Latané, Williams, & Harkins, 1979). As the name suggests, social loafing is when members reduce output because the task does not require their full measure—which is the embodiment of a motivational process loss. This discussion on process loss and task complexity are interlinked, so it is important comment upon them briefly before moving into the literature on collaborative learning. Some researchers (e.g., E. Cohen, 1994) push for open-ended, ill-structured tasks that have multiple possible solutions in collaborative activities, but it is important to note that the tasks should also be sufficiently complex (Phelps & Damon, 1989; Sears, 2006) and demonstrable (Laughlin et al., 2003, 2008).

One may wonder, for example, why is a collaborative task so relevant to individual outcomes, especially for tasks such as reading feedback and writing revisions? There is no quick answer, but, for now, the main idea is to consider how group performance might influence what a person does later as an individual, after participating in a collaboration. Group performance has been linked to and showed strong association with individual learning (e.g., Laughlin et al., 2008), what is important to note is that positive group-to-individual transfer is a common consequence of collaboration (e.g., Johnson & Johnson 1989; Laughlin et al.; Olivera & Straus, 2004). This means, the more successful group collaboration is, the deeper and more durable the learning trace becomes on the individual level. For writing instruction, this means working collaboratively with feedback might provide a more robust understanding and application of the feedback.

3.1 Collaboration in education

The use of collaboration in learning contexts is not a novel pursuit, and over half a century of research has shown impressive gains in collaborative learning outcomes when appropriate scaffolds are implemented (Johnson & Johnson, 1999; Slavin, 1996). This expectation for a positive learning experience for students prompted many educators to adopt collaborative learning activities in their classrooms, but their zeal tumbled when it came time to employ efficacious scaffolds (Antil, Jenkins, Wayne, & Vadasy, 1998). Antil et al. reported that while many teachers claimed they used collaborative activities in their classrooms, very few teachers engaged the learners with the appropriate scaffolds associated with successful learning in groups. This finding is problematic in light of Slavin's (1996) meta-analysis, which found collaborative learning groups showed no significant gains in learning outcomes when key scaffolds were not implemented. In the sections that follow, I will discuss task support structures as they are related to task complexity, the relationship of process gain and process loss with task complexity, and the truth-wins approach to understanding learning gains in collaborative conditions. These ideas build a foundation for Chapter 4 on Transfer.

3.2 Task support structures and task complexity

Collaborative learning can take many forms. According to Johnson and Johnson (2009; O'Donnell 2006), approaches to collaboration in the classroom vary in the size of groups. The magic number for group size is an enigma, but debates about the optimum group size are not random. Research has sought out whether groups as small as two people should be considered as a "group," noting that group phenomena (such as *ostracization*; K. D. Williams, 2001) are typically found in groups of three or more. However, negative social phenomena associated with groups of three or more has led other researchers (e.g., Rau & Heyl, 1990) to note how achieving meaningful group interaction and coordination becomes proportionally more difficult as group size increases. This idea calls into question the notion that more people are needed to be a group, at least if successful interaction is a trait of collaboration. Indeed, as group size increases members experience an inverse relationship in their opportunities to contribute to the group's efforts. For this reason, some researchers have pointed to the value of using paired (i.e., dyadic) interactions via activities that employ scripts, prompts, and roles to promote interactions associated with successful learning (Coleman, 1998; King, 1999; O'Donnell, 1996; Webb, 1989, 2009).

Structured group interaction with scripts helps learners focus on the content of the task, and can remove possible monopolization of group time with learners who are intensely committed to leading the group's interaction. The scripts do not remove learner autonomy; rather, they are used to specify the roles of each group member and the sequence of their task engagement. One benefit of scripted interaction is they have been shown to increase cognitive dispositions toward learning through, asking question, offering hypotheses, checking knowledge, and explaining information (Coleman 1998; O'Donnell et al., 1987; Palincsar & Brown, 1984; Webb, 1992). O'Donnell et al. provide an example of scripted interaction. In their study, scripted cooperation required learners to read and then summarize sections of a text to their partner. The role of the partner was to listen to the summary, and identify any errors or omissions. O'Donnell et al., found students in the scripted condition performed better than learners in the non-scripted condition; furthermore, the scripted learners showed a positive attitude towards their partners and the tasks (O'Donnell 1996). While this may suggest that carefully structured tasks might foster positive learner interaction and learning, it should not be assumed that employing a collaborative task may only need well-defined directions for learners. It may require that students receive training in how to successfully collaborative (Gilles, 2004; Storch, 2013). Storch suggests that in L2 classrooms, teachers should help learners recognize the importance of asking for and listening to their partner's input before suggesting a solution of their own to a language problem. Ashman and Gillies (1997) provide support for Storch's point from the primary school classroom. In their study, primary school children trained in interpersonal skills (e.g., turn taking) not only excelled on a social studies posttest exam, they maintained both the positive interpersonal skills and the knowledge over time. Although the scaffolds of interpersonal training, scripts and role structures may promote interaction and learning, structure alone does not make up a successful collaborative task—the task is also important.

At face value, this seems to be a no brainer: tasks are important. However, the nature of a task can oscillate on the continuum of difficulty from simple to complex. Relatively simple tasks are not usually touted as ideal tasks for collaboration (Steiner, 1972). For example, brainstorming has been identified as an activity that often fails to use a group's natural resources for generating ideas. In fact, Lamm and Trommsdorff

(1973) found combining the ideas of people working alone created more unique ideas than the number of unique ideas produced by people working in groups. They suggested some participants are too shy to speak, fear judgment of their peers, or suffer *production blocking* (e.g., a dominate speaker prevents others from sharing their ideas). These negative consequences inhibit positive and fruitful interactions, but this is not limited to the generation of ideas. It is also connected to the recall of ideas. For example, Andersson and Rönnberg (1995) found college students had difficulty recalling single-word vocabulary terms because of social interactions that impeded learners recall. While most research findings demonstrate simple tasks are poor venues for collaboration, some research studies have reexamined this topic with the incorporation of highly structured interventions.

Dennis and Valacich (1993) employed a computer interface that allowed learners working collaboratively to share ideas anonymously on a brainstorm task. Although brainstorming tasks have repeatedly proven they are poor tasks for encouraging successful collaborations, Dennis and Valacich showed collaborative groups can significantly surpass the number of unique ideas generated by their individual counterparts when production blocking is removed and anonymity is provided. Research like Dennis and Valacich provide evidence for the potential of structuring easy tasks so they facilitate contributions from the group, there is a different line of thinking when it comes to complex tasks.

When task complexity is high, some researchers have shown the benefits of collaboration come when the external structures are relatively minimal. For example, Barron (2000) investigated sixth grade students working in triads on a problem-solving task requiring them to generate solutions to a math problem. The learners had to create plans and offer quantitative solutions for a series of rate problems (rate = distance x time) for a riverboat traveler with limited resources (including things such as petrol, and daylight). Her research demonstrated that the triads outperformed learners who learned about rate problems in content mastery conditions. Her study shed light on the value of ill structured tasks that do not specify interaction processes through scaffolding such as scripts, roles, or rewards. However, Barron did provide some guidance to learners. Her study employed directions for students that reminded them to ask other group members to clarify and identify key data points before constructing their solution. This study helped show that complex tasks can lead to significant gains when interaction is encouraged by task demands.

One success of collaborative learning comes from research that shows learners tend to maintain their learning gains and transfer their learning into other contexts and domains. For example, Laughlin et al. (2003, 2006, 2008), set college students to the challenge of cracking a randomized code of letters and numbers. The letters-to-numbers decoding problem required a group to decode which of 10 letters (A – J) equaled which of 10 numbers (0 – 9). The point of this task was to have the learners (working in groups of three v/s working alone) solve the problem in as few attempts as possible (within the maximum of 10 attempts). To do this, participants generated a letter formula (e.g., D + H) and they would take that to the researcher. The researcher would respond with a letter combination that matches the appropriate number. In this case, if D = 4 and H = 6, the researcher would reply with the letters that corresponded to the numbers 1 and 0. Hypothetically, then, if 1 = C and 0 = J, the researcher would write "C J" to the students initial formula. What was impressive about this activity was the sheer difficulty of the task compelled learners to think about and create strategies to get as many numbers on one try as possible. This meant learners had to adapt their problem solving skills from mathematics into strategies as limited by solution attempts. For example, learners who figured out C = 1, might ask the researcher: "C + CC" (i.e., 1+11), which would equal 12. This strategy would give them one more letter's value, "2". Another strategy might capitalize on commutative trait of operands in addition problems. They could ask the researcher what "the sum of A through J" equals because 0+1+2+3+4+5+6+7+8+9 = 45, providing them two more letters that would have known values (i.e., 4 and 5).

One reason Laughlin et al., (2003) offered for the superior performance of the groups was the demonstrability inherent to the task. In prior research, Laughlin and Ellis (1986) listed four elements of demonstrability: (a) members of a group can come to a consensus for an agreed upon conceptualization, (b) members must have sufficient information to solve the task, (c) participants must be able to recognize a correct solutions when they see it (even if they could not, at first, generate the solution), and (d) members should be able to explain a solution to a partner. In the end, since the demonstrability conditions were all met in the letters-to-numbers task, Laughlin et al. hypothesized that demonstrability was the driving force behind the task's success.

One point to take away from the letters-to-numbers task is learners could take multiple approaches to finding a solution. This means that while there was only one numerical answer per letter, there were multiple ways to derive the answers. This facet allowed individuals in a group to learn from their partner's contributions, examining their personal approaches to solving a complex problem against the contributions of their partners. Markedly, Laughlin et al., (2003, 2006, 2008) noted that the members of a group (who out-performed individuals) performed better in later course work when working individually. This maintenance of learning strategies is a result that supported the notion that working collaboratively can lead to learning being transferred into individual contexts.

3.3 Task complexity and productivity: Process gain and process loss

Collaborative interaction does not always yield a positive outcome. Sometimes it can negatively affect performance. Naturally, positive gains are sought when collaboration is employed, but what constitutes positive group work? The positive effects of successful group interaction has been called *Process gain* (Steiner, 1972). Process gain occurs when the group interaction provides a benefit or outcome that was not possible without the shared activities and abilities of each group member. Process gain, then, is a unique characteristic of successful collaboration because it achieves something that could not be had without the interaction of the collaborators. Contrarily, process loss manifests when a group fails to perform to their maximum potential because of an individual who limits the effectiveness of the collaboration

(Steiner). While process loss can never wholly be eliminated from a collaborative task, tasks can be designed to maximize process gain (Steiner; Sears & Reagin, 2013).

The relationship of process loss and process gain is important to collaborative researchers for several reasons. First, as Steiner noted, understanding what a group is capable of doing and identifying what a group actually did lays at the heart of understanding how to determine if an activity was worth doing in a group. In fact, Steiner puts forth his "Law of Group Productivity" in a straightforward formula: Actual Productivity = Potential Productivity – Losses due to Faulty Process (1966, p. 274). This formula distinguishes the difference between a group's "potential productivity" (i.e., what a group is capable of doing) and what a group actually does (i.e., "Losses due to faulty processes"). Second, if, as this formula implies, there is a way to measure the potential productivity (i.e., the process gain) of a group there is also a way to identify the amount of loss due to faulty processes. Put another way, being able to measure the productivity of a group allows researchers to see how much was lost, or gained, through the interaction. This is a key idea because if an average (or taken to an extreme, the most capable) group member can perform a task better when they are alone than when they are in a group, the group will have shown process loss.

According to Steiner (1966, 1972), identifying the appropriateness of an activity is a first step to successful collaboration. To help identify what successful tasks will be, Steiner identifies where the sources of process loss spawn. He noted that a reduction in actual group productivity is triggered by two primary causes: a loss of group coordination (e.g., a group cannot coordinate their efforts for a lack of aim or synchronization), and a loss (or reduction) in personal motivation (e.g., a group has at least one member who fails to put out maximally towards the successful completion of the group's task). Of course, these sources of process loss are not theoretical, and examples of coordination losses exist in educational literature. For example, *social loafing* is a motivational loss investigated by Latané et al. (1979). Kerr and Bruun (1983) researched free-rider effects, another motivational imperfection that generated process loss. Earlier we saw Dennis and Valacich (1993) combat production blocking in their innovative brain storming activity.

As mentioned before, some researchers (e.g., Cohen, 1994) contend that successful collaborative tasks will encourage interaction and foster learning when they are designed to be maximally complex. In fact the tasks should be so complex that no single individual can complete the task alone (Cohen). This argument is made in light of the process loss that Steiner (1966, 1972) discusses because by definition process loss will occur if members of a group are not needed to complete a task (cf. unless the task was designed for task specification, like a jigsaw, which is not addressed here). By extension, then, complex tasks should yield a net process gain. From this perspective, complexity necessitates participation from the whole group, which minimizes opportunity for social loafing, or coordination loss. Seen this way, complex tasks provide learners with opportunities to coordinate their ideas towards task completion (Cohen & Lotan, 1997).

Cohen's (1994) claims echo similar points by Schwartz (1995) who suggested that learners, who coordinate their ideas (and actions) with another member of their group, tend to find and learn the deep structures of a task. In an investigation of gear motion, Schwartz constructed a task requiring participants to predict the motion of the final gear in a chain of gears. They were to determine the direction the final gear moved based upon knowing the direction of the initial gear in the sequence; this is known as the parity rule. Learners were not taught the parity rule, but were merely provided a series of problems with multi-gear chains. The problems were in word form, like, "Five meshing gears are arranged in a horizontal line much like a row of quarters on a table. If you turn the gear on the furthest left clockwise, what will the gear on the furthest right do?" (1995, p. 327). Learners heard eight of these problems, but with varying gear lengths (including 3, 4, 5, 6, 7, 8, 9, and 131). Schwartz counterbalanced the presentation of items, and carefully monitored learners attempting to solve the examples. Of the 12 dyads 58% found the parity rule, compared to 14% of the individuals, a statistically significant difference in the rate of discovery. This finding was attributed to the high frequency of abstract representations generated in the dyad condition. Schwartz contended that the dyadic condition fostered abstract representations more readily than in the individual conditions. This was attributed to dyads striving to create a common ground, which pushed dyads to generate a common representation that could coordinate the group's effort. For example, learners in dyads often used gestures to coordinate and demonstrate their ideas. Learners working alone did not need or use abstractions.

3.4 Truth wins and task complexity

A common theme in the discussion above is performances by learners in collaborative conditions surpass those by individuals. Several of the mentioned studies included tasks that showed group performances are related to (if not dependent on) task complexity. As complexity can come in many forms, logic problems (e.g., Laughlin et al, 2003), scientific concepts (e.g., Schwartz, 1995), low structured tasks (e.g., Barron, 2002), etc., putting learners within each study on the same scale might be one way of seeing how task complexity relates to successful collaboration. In order to do this there is a need to compare performances of groups and individuals on complex tasks. Doing so will help researchers see not only if a task might be a productive collaborative task, but also if individuals might perform more successfully on that same task.

Comparing performance gains can be surface comparisons (i.e., direct comparisons of outcome percentages), or they can be theoretical (i.e., comparisons of actual performances to theoretical models of outcomes). The prior is by far the most common method, but there are several ways to compare direct scores when groups are compared to individuals. For example, Figure 3.1 below shows comparison models for outcome scores.

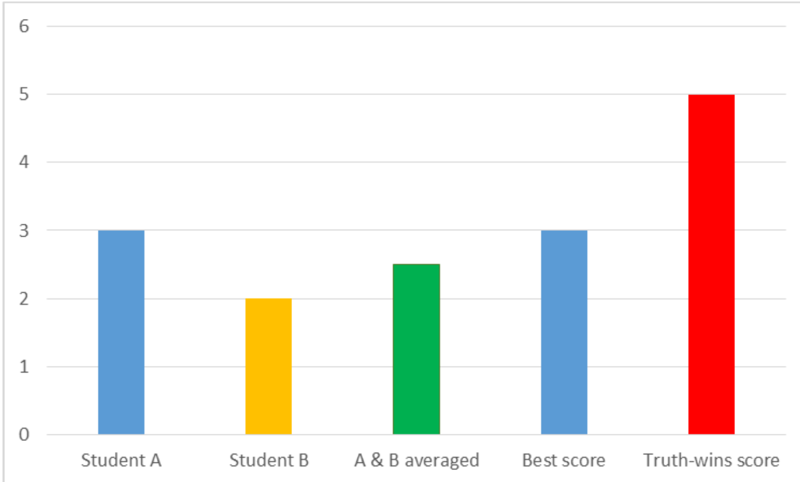

Figure 3.1. *Scores for a hypothetical test of 6-points, adapted from Sears and Reagin (2013)*

Figure 3.1 shows hypothetical scores for a pair of learners, Student A and Student B, on a six-point exam. The scores for the scoring cases are found outside the column ends. The first column (in blue) is Student A's score and the score for Student B is the second column (in yellow). With those scores in mind, consider the scenario where Student A gets items 1, 3, and 5 correct, and where Student B gets items 2 and 4 correct. Theoretically speaking, if Student A and B worked together there are several ways to consider their outcome. A researcher might take their averaged scores (i.e., 2.5, in green). Another way to determine a group's grade would be to give all members the best member's score (i.e., Student A's score: 3, also shown in blue), another perspective might take a punitive approach, giving the lowest score to all members (which is not in the Figure). However, consider a situation where Student A and Student B were able to perfectly share their knowledge. If they were able to pool their knowledge, they would yield a score of five (seen in last column, which is red). This theoretical pooling of knowledge is called the Truth-wins score.

Surpassing the truth-wins score would require Student A and Student B to put their collective knowledge together to co-construct the correct response to the final item #6. Doing this would allow them to exceed their theoretical best under the truth-wins condition. The truth-wins comparison of learning gains puts the theoretical best possible for a population of learners and pits it against the actual score of the collaborative condition. The truth-wins comparison constitutes one of the most rigorous tests of learning gains because it creates a theoretical scenario in which knowledge is shared perfectly between members of a nominal group who are "mathematically" paired. An example may help with understanding this complex idea. The Student A and B example has limitations, so let us consider the parity study by Schwartz (1995).

Schwartz's (1995) findings showed that 14% of the learners came up with the parity rule in the individual condition, while 58% came up with the rule in the dyad condition. To compute the best theoretical mean for individuals, we use the Model A formula from Lorge and Solomon (1955). This formula (commonly known as the "Truth-wins" formula) is [$1-(1-\text{observed mean})^2 = \text{Truth-wins value}$]. This truth-wins value represents the theoretical best average score for the individuals, if they mathematically pooled their knowledge. Thus for Schwartz's study, the individual condition's truth-wins score is [$1-(1-.14)^2 = .2604$]. This means—if individuals were to share their knowledge perfectly—they would have averaged 26% success discovering the parity rule. Since the observed score for the collaborative dyads was 58%, the real dyads beat the theoretical best of the individuals, which is to say the dyads beat the truth-wins comparison. This is a truly impressive finding, and extremely rare! Consider another example, one about a hypothetical vocabulary test between two classes.

Two classes take different approaches to learning a list of vocabulary terms. On the one hand, Class A uses a learning technique using individual conditions, such as learning with flashcards. Class B, on the other hand, uses a novel approach to learning vocabulary terms and has learners play a charades activity, using gestures to demonstrate what the vocabulary terms mean. After a two-week period of studying, the classes take a test on the same list of words. Each class takes the test individually. Class A averages 40%, and Class B averages 60%. Now, these scores are disparate and a statistical test will likely show that Class A and B are different. However, the truth-wins comparison allows us to compare what Class A might have scored if they were allowed to work together and share their knowledge perfectly. The truth-wins calculation would be: $[1– (1 – .40)^2 = .64]$. This means, by the truth-wins comparison, Class B (the collaborative class) did *not* surpass the theoretical best of Class A. This does not mean the statistical significance between the scores evaporates. What is does mean is Class B likely encountered some process loss, since easy tasks (such as memorizing vocabulary terms) are not likely to yield strong learning gains under collaborative conditions (Phelps & Damon, 1989). To help conceptualize learning gains and truth-wins comparisons, Sears and Reagin (2013) have offered the complex-demonstrability framework (see Figure 3.2, below).

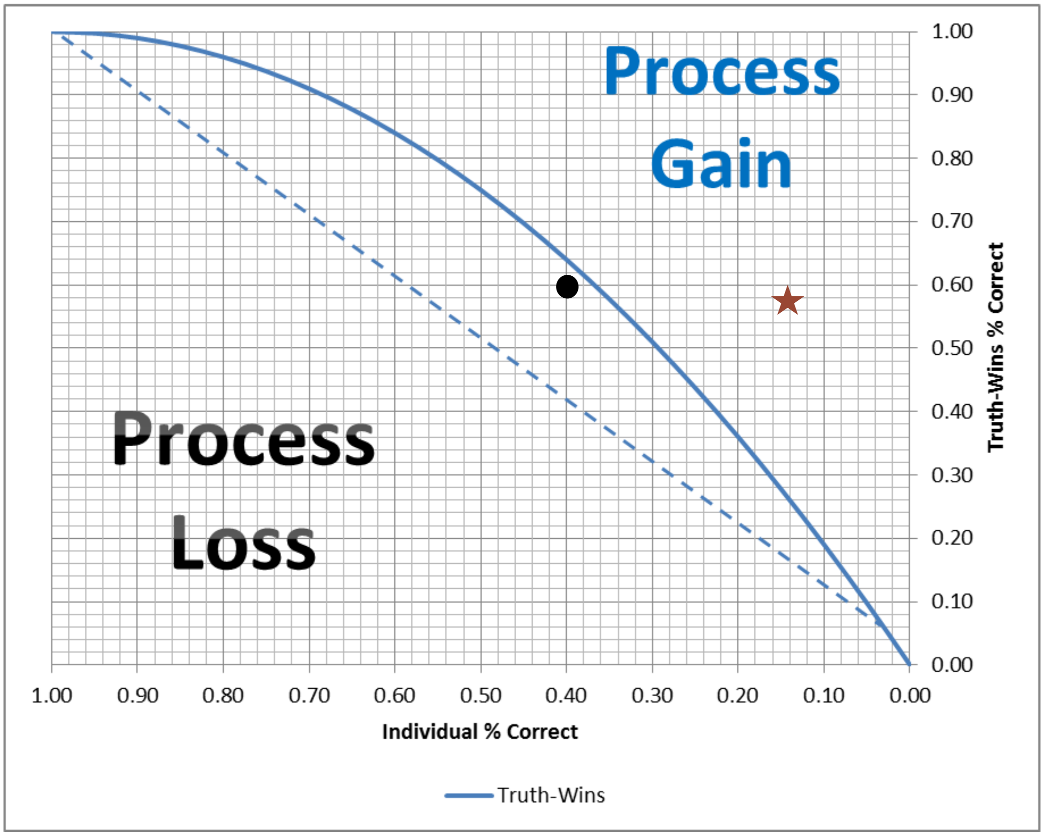

Figure 3.2. *Complex-demonstrability framework (adapted from Sears & Reagin, 2013). The black dot is the theoretical best for the vocabulary test scenario between Class A and Class B, and the blue star is the plot for the truth-wins comparison from Schwartz (1995).*

Figure 3.2 shows the theoretical best average score for individuals (i.e., their truth-wins score) put into theoretical dyads (also called *nominal dyads*) along the x-axis, in two point increments. The true dyad score is plotted along the y-axis. A conceptual understanding of this graph compels inspection of the x- and y-axis, and how they interact. The x-axis runs from left to right, but in descending order from 1.0 to 0.00. The y-axis is on the right, and runs from 0.00 at the bottom to 1.0 at the top; it is on the same scale as the x-axis. The interaction of the x- and y-axis creates a field on which actual dyad performance (on the y-axis) can be plotted against the theoretical performance of truth-wins scores of individuals (on the x-axis). The dotted line splitting the graph into two diagonal sections is the delineation between the process loss portion of the graph (noted in black text in the bottom left) and the process gain portion of the graph (noted in blue text in the upper right). A score comparison that exceeds the dotted diagonal mean the dyads beat the average individual's score; however, a score comparison below the dotted diagonal is an indication of strong process loss. The downward-turning blue-arc is the plotted truth-wins curve, which identifies the threshold of when a collaborative condition has surpassed the truth-wins comparison (which in this case is dyads, note that the curve must be recalculated and redrawn for groups larger than two). Surpassing the truth-wins curve will indicate a net process gain, and suggests there was something unique to the collaboration that was not seen in individual condition despite the perfect pooling of knowledge.

Plotting the fictitious vocabulary study between Class A (individuals) and Class B (collaborative) is shown with the black dot. Figure 3.2 graphically depicts what the truth-wins calculation has already shown us, but Figure 3.2 is useful for its ease of facilitating raw score comparisons. In the fabricated vocabulary test given above, we plot Class A's 40% along the x-axis, then we search for Class B's true score of 60% along the y-axis. The intersection of Class A's score and Class B's score is noted with the black dot. The location of the dot is above the diagonal, which indicates the collaborative Class B's 60% is better than the average score of Class A; however, since the dot is below the plotted truth-wins curve, their 60% does not surpass the theoretical pooling of knowledge for Class A.

Conversely, we can see Schwartz's (1995) study plotted in the process gain region, with a blue star. In his study, the individuals had a truth-wins value of 26% and the dyads had a 58%. By plotting the raw scores, we can see the intersection of the 58% for dyads and the 14% for individuals places Schwartz's study above the truth-wins curve and in the "process gain" field. Figure 3.2 provides a quick truth wins comparison for raw scores, without calculations. However, it also demonstrates how stringent the truth-wins comparison is. Consider the example vocabulary study. A 20% difference in raw scores between the two classes did *not* beat the truth-wins comparison. Surpassing the truth-wins is no small achievement.

CHAPTER 4. TRANSFER

The application of knowledge from one context to another is a form of generalizing knowledge constructs. This is called transfer, and it is at the heart of educational practice (Bransford & Schwartz, 1999). Despite the seemingly universal nature to aim for transfer in educational practice, some researchers (e.g., Detterman & Sternberg, 1993) claim the transfer of knowledge is rare if ever possible in education. Collaborative researchers, however, have consistently shown how the cognitive processes of explanation, elaboration, and clarification evident in student interaction play a role in supporting durable learning (Webb; Webb & Farivar, 1994). Learning in groups, then, is important to education because it prepares learners to apply that knowledge to other situations (i.e., transfer their knowledge).

In L2 writing research the picture of transfer from the language classroom to other coursework is blurry. Some researchers claim transfer, while limited, is possible (James, 2006a, 2006b, 2009, 2010; Larsen-Freeman, 2013). Other researchers take the position that transfer is not a reasonable expectation for an L2 writing curriculum, given the expansive nature of possible educational context in which writing can be applied (e.g., Leki & Carson, 1997; Spack, 1997). Dyson (1999) took a different perspective on transfer research, stating that researchers have not asked the right questions nor looked in the right places for evidence of transfer. One approach L2 writing specialist James has taken is to look at how writing course outcomes transfer to the other academic courses.

James (2010) interviewed and collected writing samples from 11 students enrolled in an English for Academic Purposes program at a large state university. The participants provided sample essays from across academic courses. The data showed that learning outcomes (e.g., "avoiding confusing sentences"; James, 2010, p. 183) transferred to their coursework outside of the language classroom, but this was not true for all learners across all disciplines. James also found some elements, such as "describing visual elements," and "using similes and metaphors" rarely transfer (p.183). While there were no explicit patterns to which outcomes would transfer into disciplines seamlessly, this study showed how transfer of L2 writing skills, as operationalized through course outcomes, transfers to some educational contexts outside the language classroom. This finding reflected similar findings of James (2009) who discussed the variation in transfer of learning outcomes in terms of a transfer task designed to reflect writing done in academic classrooms.

James (2009) asked 30 university students enrolled in an English for Academic Purposes (EAP) course to respond to a prompt that was chosen for its similarity with the tasks that learners do *outside* of the L2 writing classroom. His study was designed to test the notion that L2 writing students would transfer their learning experiences to other writing tasks. James found that very few outcomes transferred, but he questioned this depiction of transfer because outcomes may not all manifest in a single writing assignment. This idea was driven by findings in James (2006). James (2006) observed that across a series of interviews with five students, spanning an academic year, that the learners infrequently noted the transfer of skills they learnt in their ESL courses. The same learners pointed to similarities in the task that helped them transfer. At the same time, they also identified how differences in the writing task prevented from seeing how to (or when to) transfer their learning. As an exploration of the possibility of drawing learner attention to the opportunity for transfer, James (2010) designed his investigation.

For the 2010 study, James met with each student for one hour to collect data. Each student completed a reading of a science article that described the biotechnological process of turning animal waste into fuel. After reading the article, the learners answered two questions in writing. The first question probed their comprehension by asking them to explain the process described in the article. The second asked them to describe why it would be important to the science field to promote the development of this technology. James split the group into two halves, with half of the students writing on the science article task as it was given to them. He had the other half complete an extra step. He asked these students to identify at least four similarities between the writing task given to them (i.e., the animal waste article) and the tasks they had completed in their L2 writing course work. The purpose of this was to assess if learners could identify similarities in their course work, thereby promoting a focus on skills learnt in the classroom. After completing the tasks, he interviewed each participant. His findings showed learners transferred reading outcomes (e.g., finding the main idea, guessing vocabulary) and writing outcomes (e.g., using transitions for organization). However, these outcomes were only noted by 13 learners (43%). From this 13, five learners only noted reading outcomes and the others noted both reading and writing (n=4) or writing only (n=4). The results provided evidence that outcomes did transfer from the course to the task, but in very limited ways.

Another finding James (2010) sought identifying the effects of drawing learners' attention to the similarities between the science article task and their writing class outcomes. The results showed a non-significant result between learners who were asked to look for similarities and those who did not, X^2 (1, N = 30) = .621, p = .431 (p. 77). Taken together, James showed that outcomes do transfer, but not across all contexts or assignments. He also found that even if learners could identify similarities between a transfer task and the course outcomes it did not help them transfer. This research provides a nice systematic investigation of the transfer of outcomes. However, what this points to is a need to investigate transfer from a classroom perspective, such as a study that investigates how tasks in the language classroom can foster a deeper understanding of language. By having a deeper understanding, learners are more likely to be capable of transferring their skills to novel situations. This is a good transition to discuss how collaborative researchers have investigated transfer in the classroom.

4.1 Collaborative learning and transfer in the classroom.

Bransford and Schwartz (1999; Larsen-Freeman, 2013) have called for investigations of transfer across educational activities to consider learning from alternative perspectives. Specifically, Bransford and Schwartz call for a reconsideration of a “sequestered problem solving” (SPS) view of transfer, which emphasizes rote learning. Instead, they promote a vision of transfer that extends learning assessments and activities to a forward-looking definition learning, called, “preparation for future learning” (PFL; see also Schwartz et al., 2005). In the PFL framework, learners engage in *innovation tasks* that push learners to generate a solution to a difficult problem, work with a resource, and then tackle problems that are more complex. The innovation tasks compel the use of prior knowledge as learners attempt to construct, or at a minimum offer, solutions to problems that are unfamiliar or unknown to them (Sears, 2006). This approach is opposed to the *efficiency* model of learning tasks, which are normally associated with SPS assessments (Sears). Learning facts and being able to recall them on demand is not bad per se. However, efficiency tasks restrict the mental dexterity of learners, an essential trait for transfer (Schwartz, 1995; Schwartz et al., 2005). A primary objective for having students innovate their initial solution is to prepare them to perceive and appreciate a resource from which they can learn. The resource could be a person, an expert solution, or other source of useful information that enhances their knowledge base (Schwartz & Martin, 2004). This change in perspective from SPS to PFL has important implications for how language classrooms might design, employ, and measure effective learning tasks.

The PFL framework posits that the innovation of ideas and solutions through dyadic interaction can be a valuable catalyst for learning transfer of the deep structures of knowledge that apply across multiple contexts. As noted above, fostering (or identifying) transfer is not an easy task. One reason this is particularly difficult in learning new information is novice learners primarily focus on surface features (Chi, Feltovich, & Glaser, 1981). Attention to surface features may not seem like a confound for transfer, but it is necessary to consider how pliable superficial knowledge of external structures are across learning contexts.

For example, Chi et al. (1981) found that the categorization of physics problems was divergent for novice (undergraduate students with an introductory knowledge of physics) and experts (Ph.D. students in Physics), but perhaps not for the reasons one might think. For this experiment, Chi et al. asked novices and experts to separate physics problems written on 3 x 5 index cards into categories, without calculating the true answer. Not calculating the items forced the participants to estimate how the physics problems were related at a glance. The index cards listed terms (e.g., "weight") and had pictures of objects typical in physics classes (e.g., an incline plane, arrows for direction, spheres, pulleys, and the like). Surprisingly both the experts' and the novices' classification of the problems were not distinct. There was no statistical difference between the numbers of groupings created by the participants or the problems classified under the categories. The difference between the novices and the experts was the reasons for classifying the physics problem into a given category. The findings showed that experts focused on deep structures of the problems, identifying generalizable trends that can be applied to multiple situations (e.g., "Newton's second law," "conservation of energy"). The novices categorized through an analysis of the surface features presented on the index cards, without consideration of the interconnections of the objects to the physical theories they learnt in introductory physics courses. For example, they mentioned "a spring," or a vocabulary term found in the question stem listed on the card (e.g., "friction"). Their statements on why they classified the problem into categories were also superficial, "These deal with blocks on an incline plane" (Chi et al., p. 126).

The physics study provides some evidence to the nature of how novices struggle to identify deep structures (Chi et al, 1981). Similar to the James (2009) study, psychological research has shown how learners struggle to see when their prior knowledge applies. Gick and Holyoak (1983) identified that learners often do not spontaneously transfer their knowledge, and often they need prompting. The research by Chi et al, Gick and Holyoak, and James are indicative of investigations into individual learning conditions. Small-group and collaborative learning, however, has been shown to enable learners to interact with partners to discuss, explain, and negotiate a deeper understanding that is typically not available to those working alone (e.g., Schwartz, 1995). This interaction often stimulates knowledge in the participants, preparing them to transfer on later tasks. This interaction also provides a rich environment for giving and receiving explanations, which have been connected to learners achieving an understanding of the deeper constructs behind the materials (Chi et al. 1994; Webb, 1992). This process of bring multiple perspectives together develops a shared understanding that leads to a more abstract representations of the problem (Schwartz).

4.2 Applying PFL assessment of transfer to writing classrooms

Why is the literature from collaborative learning applicable to investigating feedback in language classrooms? Collaboratively learning is a natural phenomenon. According to Vygotsky (1978), cognitive development and knowledge construction are a result of social interactions. Fundamental to his theory is the premise that learning (i.e., the development of knowledge) is a social process of interactions by individuals. These individuals engage with the learning process by interacting to create and internalize knowledge (Vygotsky, p.128). For language learning, it is commonplace for group work to occur in the communicative language classroom.

Much of the group work on interaction in the L2 classroom is no doubt connected to the seminal work of Long and Porter (1985). They identified five arguments in support of group work activities in the language classroom. They explain, from a psycholinguistic perspective, that group work "increases language practice opportunity, improves the quality of student talk, helps individualize instruction, promotes a positive affective climate, … and … motivates learners" (pp. 208–212). Not unlike the collaborative research reviewed earlier, the support for group work in Long and Porter's article is founded on the notion that negotiating meanings is an essential and necessary means for language acquisition.

Of the many methods possible for instruction using collaborative learning, some are familiar to second language research, while others are less commonly used in L2 classrooms or research. For example, the *jigsaw* task is a common task employed in communicative research and classroom activities in foreign and second language environments. According to Pica, Kanagy, and Falodun (1993) there are other classroom tasks (e.g., information-gap, problem-solving, decision-making, opinion exchange) available for teachers looking for tools that push learners to interact with a partner; however, L2 research studies have typically used interaction tasks to investigate linguistic features of language use and acquisition.

The long line of research on Complexity, Accuracy, and Fluency (CAF) structures (see Norris & Ortega, 2009 for a review) and Language Related Episodes (LREs) are examples of how language research has considered components of language as indicators of uptake and transfer. On the surface, this makes sense. The notion that grammar applies to multiple contexts is a sound one. The idea, too, that a strong control of grammar and lexis will more likely help learners across multiple contexts is logically more viable than a lack of control doing so. With that noted, measures of language features such as CAF metrics do not have predictable growth patterns. For example, Koyama and Sun (2012) showed in their two-year longitudinal study of 90 EFL learners' L2 English writing placement tests, fluency (i.e., number of words per essay) was the only CAF metric (among the 14 features examined) to move in a predictable direction over time. Koyama and Sun noted that CAF metrics might not be sensitive to change over an academic year, especially in one-shot assessments of writing. The notion that snap shot assessments of learner writing may not provide a robust picture of learner processes in writing is one reason why Neomy Storch began her line of inquiry into collaborative writing (CW).

Collaborative writing started for Storch when she realized that learners in the classroom were actively engaging with each other during a classroom writing assignment (Storch, 2013). Watching learners interact made her question why researchers continued to investigate L2 writing through individual writing assignments. Her research was inspired by long time advocates of collaborative writing, Ede and Lunsford (1990). Ede and Lunsford defined collaborative writing as a social act of equally shared responsibilities. In their book length treatment of the topic, they put forth a bold call for institutions valuing individual writing to consider a paradigmatic shift in those power structures to accommodate and equally value collaborative writing. They called for an increased awareness of the value and importance of collaborative writing as a reality of their professional experiences; they simultaneously called on collaborative writers to engage in meaningful interactions throughout the composing process. This included an identification of shared responsibilities at every stage of writing, such as planning, creation, editing, and the like. Thus, from this viewpoint, it is not only the writing process but also the writing product that becomes collaborative (Koyama, 2014).

4.3 Lessons from the collaborative writing research

In one of the first L2 collaborative writing studies Storch (1998) examined Rutherford's (1987) *propositional cluster* task. This task presented students with content words that were to be used in the construction of a text. Thirty ESL learners were put into groups ranging from two to five members, and each group's interactions were audio recorded. Grammatical issues were the focus of the analysis as identified with Swain's (1994) language related episode (LRE) framework. Storch's findings indicated that students focused most of their collaborative attention to grammatical features, and did so through justified resolutions by recalling grammatical rules and conventions (p. 297).

In an effort to elaborate on how task effects might affect collaboration, Storch (1999) investigated a set of grammar-focused communicative tasks: a cloze exercise, a text reconstruction, and a composition task. Storch (1999) found, when writing in pairs and groups, learners generated more accurate essays, but they wrote more complex essays when writing individually. This brought attention to the possibility that collaborative task structure can affect the quality of product in terms of not affording opportunities for an individual's best performance. From the stance of the literature on productivity in collaboration, we know this outcome to be process loss.

In a follow up study that examined the patterns of pair interaction, Storch (2001b), investigated the nature of pair interaction of three pairs of adults in an ESL classroom. Storch (2001b) analyzed transcripts of the pair talk during collaborative writing sessions, and the results of the linguistic features showed that first person pronouns were used differently by groups but 11 of the 13 uses (or, 85% of the uses) were with force and to emphasize their opinions (Storch, 2001b, p.34). Based on the analysis of pronouns, as being one way to identify the degree to which a dyad functions in collaboration, Storch (2001b) identified two of the three dyads as being collaborative and the other dyad as being non-collaborative. This finding aided in the interpretation of the text construction behavior, which showed that the collaborative pairs discussed the text in terms of how it should look, and actively tried to incorporate suggestions from one another. Non-collaborative pairs had overtones of hostility and defensive strings of interactions, such as dominating the interaction, ignoring a suggestion, and arguing over the accuracy of revisions (Storch, 2001b, p.45). These are traits of production blocking, and coordination loss, which are symptoms of process loss. From the perspective that the writing process of collaboration through interaction facilitates language learning more than the writing product itself, the findings of Storch (2001b) are important to collaborative writing research as they point to the potential pitfalls of collaboration while writing without task supports such as scripts and roles.

Expanding the findings of patterns of interaction from Storch (2001b), Storch (2002a) conducted a longitudinal investigation collaborative interaction in another adult ESL classroom. This time she analyzed interaction data from 10 dyads over an academic semester. She drew from collaborative learning research by Teasley and Roschelle (1993), and identified *joint problem space* (JPS) creation through four distinctive patterns of interactions. A JPS is created by learners interacting to support problem-solving by integrating their goals, actions, and intentions (Teasley & Roschelle, p.229).

These interactions were put on continuums of *equality* and *mutuality* that intersect to make a grid (see Figure 4.1).

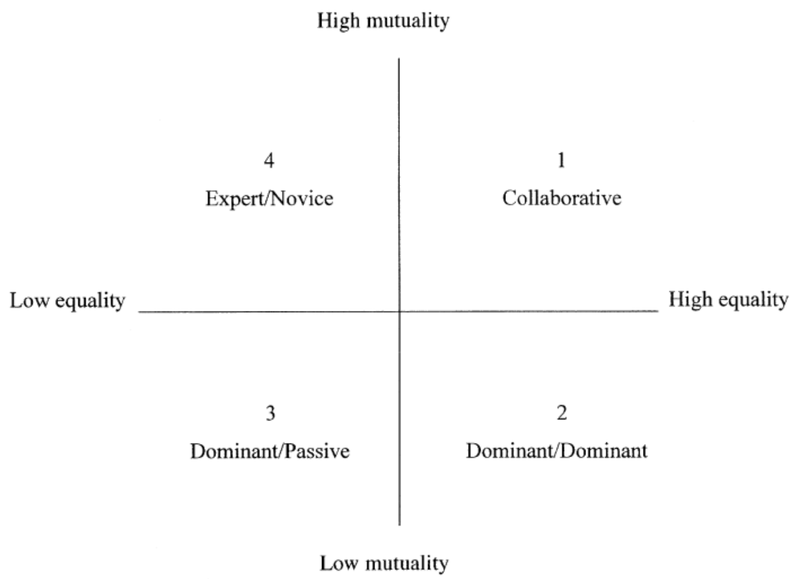

Figure 4.1. *A model of dyadic interaction from Storch (2002a, p.128).*

Figure 4.1 shows quadrants that represent the collaborative nature of the interactions of dyads on a writing assignment ...

> Quadrant 1 represents a pattern of dyadic interaction where there is moderate to high equality and moderate to high mutuality […] Quadrant 2 represents a pattern of interaction where there may be moderate to high equality, but a moderate to low level of mutuality. […] Quadrant 3 represents a pattern where the level of equality and mutuality are both moderate to low […, and] Quadrant 4 represents a pattern of dyadic interaction where there is moderate to low equality but moderate to high mutuality (Storch 2002a, p.127-129).

This framework for identifying which interactions are collaborative was offered as a useful tool for researchers considering the effects of task design on dyadic collaboration. It is also a nice bridge to the collaborative learning literature of educational studies. What Storch's series of studies has done is illuminated how task structure and student interactions intermingle. The identification that some prompts, such as text reconstruction task, encourage students to produce grammar rules to support their decisions (Storch, 1998), is also supported by Storch (1999) who provided evidence that task type and condition of interaction affected the accuracy of learners' production. These findings suggest that while close-ended grammar tasks completed collaboratively can encourage learners to share knowledge, not all exchanges were collaborative (Storch, 1998, 1999a, 1999b, 2002a, 2013).

This trend for digressions and non-collaborative exchanges was identified in additional research by Storch (2005), and Fernández-Dobao and Blum (2013). These researchers interviewed different groups of learners and identified that learners found the collaborative writing experience to be very positive, but students in both studies noted that sharing information was not always productive, and sometimes controversial. While this should be considered in light of the findings of previous research studies by Storch—that showed task effects for the kinds of interactions in a given collaborative activity—poor collaboration is not to be taken lightly. When the learning processes of individual learners are undermined, it also erodes successful collaborative interaction and can encourage process loss (Steiner, 1972).

As mentioned above, collaborative researchers suggest scaffolds to guide and promote the collaborative learning task and to abate opportunities for process loss. Scaffolds can be designed to highlight features of a task content (e.g., using graphic cues to promote learning transfer: Gick & Holyoak, 1983), to promote structured interaction (e.g., scripts and roles: O'Donnell et al., 1987), or to develop a network of reliance on other group members (e.g. interdependence: Johnson & Johnson, 1999; rewards: Slavin, 1996). Thus, the charge for L2 writing instructors is to create tasks that support collaborative dialogue.

4.4 Innovation and preparation for future learning

Research has shown (e.g., Barron, 2003; Sears, 2006) that collaboration while innovating can create interactive experiences that promote preparation for future learning (PFL) . For example, Barron showed how joint attention to a novel problem generated new information that was only possible through interaction. This points to the potential of collaborators, who bring their own backgrounds and experiences, to interact to solve problems in unique ways that problem solving alone cannot. Sears (2006) found that when college students collaborated to innovate and learn the chi-square formula, they outperformed students who were given the formula and time to practice examples (i.e., efficiency tasks). The efficiency dyads displayed a lower performance on transfer tasks than those who were put to task trying to invent the formula before seeing the expert version. Sears' finding underscores how invention in collaborative tasks pushes collaborators to share ideas towards creating abstractions that are generalizable to other situations, rather than focusing on finding the "right answer" (Schwartz, 1995; Schwartz, Chase, Oppezzo, & Chin, 2011; Schwartz & Martin, 2004).

Preparation for future learning assessments are not like traditional assessments of learning. Most commonly, exams use measures to see what learners already know, retrospective assessments (Sequestered Problem Solving); PFL assessments tap the constructs that transfer, knowledge that was not explicitly taught (Schwartz & Martin, 2004). PFL measures differ from SPS measures more specifically because PFL assessments allow resources for students to learn from during the test. The resources vary in presentation, and can include worked examples, explicit feedback, and even other people (Schwartz & Martin). However, as Gick and Holyoak (1983) noted, learners often do not spontaneously transfer their knowledge. In the case of PFL assessments, learners need to be prepared to learn from the resource materials so they can solve a novel problem when they encounter it (Schwartz & Martin; Sears, 2006) What is particularly useful about PFL measures is the information they provide that cannot be gleaned from SPS assessments (Schwartz et al., 2005). A classic example of how a PFL assessment can highlight learning is Schwartz and Bransford (1998).

In their study, Schwartz and Bransford (1998) investigated two groups of university students learning about psychological theories. One group of students read a summary table of data sets from classic psychology experiments, and the other group wrote analyses of example cases from a chapter that described the same experiments (Schwartz & Bransford). Immediately after completing their tasks, the researchers gave both groups a true and false test (a classic SPS assessment) about the contents of the studies. The results showed that the learners in the group that received a summary of the studies performed better than the group who wrote their own analyses of the psychological studies in the chapter. If the experiment were completed at that point, the true false test could have been an indicator that providing learners with a summary table will help them learn more than reading and analyzing example studies. This is where the PFL assessment helped.

After completing the true and false test, the learners were all exposed to a lecture that discussed the psychological experiments they had just read. The lecture covered the results and the broader implications the study had for human behavior. The lecture was the resource in this PFL assessment. Students from both conditions heard the same lecture, and they were asked to predict the outcome of a novel experiment. This novel experiment was very different from the studies they had just learned about, and it had very different surface features (Schwartz & Bransford). This time, the learners who analyzed contrasting cases from the chapter surpassed the learners who read summaries of the chapter contents. This was interpreted to mean that the analysis of the cases better prepared the students to learn from the lecture than just summarizing the chapter contents. This meant the PFL assessment of having learners read and analyze, prepared them to learn from a lecture, which further prepared them to transfer that knowledge to answer the transfer prediction problem. Recall that the group with the summary chart excelled on the true and false test that the read and analyze group did poorly on. This difference in assessments is, Schwartz and Bransford assert, how PFL assessments can be seen as a way of capturing transfer that cannot be detected with SPS assessments.

CHAPTER 5. THE STUDY

5.1 Study overview

This study investigates the effects of learner-learner interaction with an expert model in a multistage writing task. The current literature on multistage writing tasks and learner use of feedback has painted a picture that individual learners can autonomously glean useful language from a model essay, but to date learner interaction has mainly been incorporated in unstructured ways (e.g., collaboratively composing a writing sample). The collaborative writing literature has shown many instances of poor collaboration when learners intact in the invention stage (i.e., the writing stage), but also identifies that when learners share task time and listen to each other they collaborate well (Storch, 2013). This study contributes to the SLW literature in three primary ways. First, it investigates interaction as an important facet for uptake from models. Second, it will do this by taking into account the overt focus on lexis when learners compare their writing to feedback; thus, this study will offer learners the model without access to their original writing. Third, it will provide a fresh method of feedback using an isomorphic (or, parallel) model essay to obviate the potential of mimicry in subsequent writing samples. This has been a concern for L2 writing studies utilizing models (see literature review). These combined traits are hypothesized to push learners to focus on content, by interacting with a peer while reading an isomorphic model.

This study also adds to the collaborative learning literature by investigating the flexibility of PFL assessments. The twist on PFL assessments in this study was made based on the findings that learners have difficulty giving (or taking) constructive feedback on collaborative writing tasks. Instead of having learners collaborative write (i.e., collaboratively innovate a response) in Stage 1, learners will collaborative learn from a model in Stage 2. This was done to illuminate the process blocking that often occurs in collaborative writing task when learners cannot agree on language. Since the task used was deemed complex, this change was made. With that notation, the resource model provides an expert solution, and will be a safe topic for discussion as their writing and ideas are not at stake. This means, however, that unlike the PFL framework standard of innovating in dyads then working independently, this study flips the stages (innovate with individual writing then collaborate on the resource model), but maintains the opportunity of learning under collaborative conditions. Investigating this will contribute to evidence of the plasticity of PFL assessments, and the application of collaborative task with feedback in the L2 writing classroom. Based on research highlighting the importance of incorporating additional sources of information in writing as an essential skill for the academic success (Campbell, 1990; Leki & Carson, 1997), the task chosen for this study was a graph description task.

5.2 The rationale for graphs

Several elements of graph design can affect reader perceptions of the data displayed (Carswell et al., 1992). Therefore, the graphs were carefully crafted for this study. Column graphs were chosen because they can display trends in multivariate data in ways that are clearer to readers when compared to readers of line graphs or pie charts (Carswell et al.). This is especially relevant for this study because data comparisons were part of the overall messages in the Graphs[2] used (e.g., for Graph 1: which bookstore sells more books, and which book condition is most popular, see Appendix B). While line graphs and pie charts can be used for multivariate data, there are normally several charts required and the interpretations are often confounded as readers struggle to merge and interpret data across separate graphics (Shah & Hoeffner, 2002). Since the data in Graphs 1 and 2 (See Appendix F) were not trends over time, a line graph was deemed inappropriate. Additionally, research on reading graphs shows that pie charts are not suitable for showing interactions in data, but are better aligned with displaying information about relative proportions of the data through percentages (Shah & Hoeffner). With those ideas in mind, grouped-column graphs were chosen and were designed to facilitate comprehension.

Since the ability to identify trends and making data integrative comments (such as summarizing data points or comparing data points) can be interpreted as understanding a graph (Carswell et al., 1992), it was important that surface features of

[2] Graphs with a capital "G" will be used to reference the graphs used in the study.

the graph did not create undue stress on the readers that might distract them from comprehension. To assist the readers, colors were used to clearly identify each of the superordinate groups (e.g., bookstore location), and data labels were displayed over each column because colors cannot easily (or clearly) represent specific numerical information (Cleveland & McGill, 1985). Furthermore, the surface features of the Graphs were clearly labeled (i.e., x- and y-axis, superordinate and subordinate column names, and all data values) instead of using a legend or key, which have been found to tax working memory (Kosslyn, 1994). Additionally, the topic of the Graphs (for Graph 1: bookstore purchases by location and condition, and for Graph 2: living conditions and locations for freshman) were familiar topics to students as they all have purchased a book at one of the locations, and they all live with or without a roommate on or off campus. This is important because Shah (1995, 2002) showed that when students have background knowledge of a topic—and the data follow expectations—their interpretations are more accurate, especially when the data span two or more data sets. Thus, I elected to include data that made sense to students by topic familiarity, as learners would have to reflect on the graphs alone or in dyads.

The Graphs were also designed to have global and local data integrative points (Carswell et al., 1992). Global integrations are statements that integrate all of the data points in the graph, such as identifying an overall trend. For example from Graph 2, the statement, “a majority of students live with a roommate,” requires the reader to also note that there are fewer students that live alone regardless of the living location; thus all the data points are considered. A local data integration limits interpretations to a subset of the whole data. For example, from Graph 1, "60 students purchased a used book at the campus bookstore," does not require the reader to consider the new book purchases at the campus store, nor any of the purchases from the online store. Local data integrations can be summative, but they will not address all of the data in the graph. The scope of the data integrations is a key feature that distinguishes local and global data integrations.

As local data integrations are easily seen and produced by readers of column graphs (Carswell et al., 1992; Shah, 2002), global integrations represent a challenging aspect of this graph description task. In this way, global integrations were used to distinguishing Graph 1 from Graph 2. Graph 2 was the resource graph used in the treatment stage and was accompanied with a model description. It was designed to have a straightforward interpretation of a dataset without an interaction effect. Specifically, living with a roommate was the most popular living situation overall, and for students who lived on campus or in private homes. This means the local data integrations mirrored the overall global integration. In other words, so long as students saw one trend, the others were equally transparent (Carswell et al.; Cleveland & McGill, 1985; Kosslyn, 1994; Shah). This was not the case for Graph 1, which was designed to have an interaction in the data. This interaction was the target transfer element. For Graph 1, the overall trend must be synthesized from the data points and is not a foregone conclusion as it is in Graph 2. Since the resource model did not present students with a solution for identifying the interaction, Graph 1 presents students with the task to identify a salient interaction trend in a bar graph. This interaction is not easily detected in bar graphs, and can be seen as a complex task for learners (Carswell et al.; Shah).

To this end, the following research questions were developed:

Given that: the Control condition consists of learners who repeat the writing task without the model resource task, the Individual condition consists of learners who read an expert model without collaborating with others, and the Dyad condition consists of learners who work in collaborative pairs with an expert model ….

Research question 1: What is the difference in the amount of change in content overlap with an expert model between the Control, Individual, and Dyad conditions?

Research question 2: In terms of local and global data integrations, what are the differences for producing the expected number of data integrations represented in an expert model between the Control, Individual, and Dyad conditions?

Research question 3: Will learners in the Control, Individual, and Dyad conditions identify the complex global integration found in the interaction of Graph 1?

5.3 Methods

Participants. The students in the study were all (N = 54) matriculating undergraduate students registered for a required English Composition course at a large public university. Four learners did not complete at least one of the treatment days; their data was excluded from the study. The students came from four intact classes, each course nearing the final paper for the term: “the argumentative essay.” This put each course at a point in the curriculum where a lesson on using and describing data fit their studies.

Institutional constraints prevented a stratified-random sample for condition assignment, so whole classes were randomly assigned to one of three conditions: Individual, Dyad, and Control. Notably, as dyadic interaction was a variable in this study, and research has shown that rapport can be a moderating variable for successful collaborations in the writing classroom (Storch, 2013), it was decided keeping classes would not hinder interpretations of the data. In fact, an added benefit to this design is groupings can be seen as a depiction of what might naturally happen in a writing classroom (i.e., classmates working with other classmates, rather than with students from another class).

Fifty of the total 54 original participants had a TOEFL iBT sub-section score for writing, and, as it was impractical to collect a separate proficiency measure, the TOEFL iBT score was deemed a reasonably acceptable measure to ascertain the students' English writing ability. To test for potential bias of the placements into treatment conditions, the students' TOEFL iBT writing sub-section scores were submitted to an ANOVA to test for significant variations between group scores. The presence of a significant difference would mean learners in one or more conditions do *not* share the same language ability as determined by the TOEFL iBT writing score. The writing scores served as the dependent variable and the assigned condition was the independent variable. The descriptive statistics for the TOEFL iBT scores can be seen in Table 5.1.

In Table 5.1, column five shows that the means of each condition are nearly identical, ranging from 22.08 to 22.23. Further inspection will show that the mean of each condition is contained within the other groups' upper and lower 95% Confidence Interval (95% CI), and the 95% CIs of the groups all overlap. These are characteristics that indicate that there is no statistically significant difference between the mean scores of these groups. However, an ANOVA was employed to verify the condition assignments.

Table 5.1. *Descriptive Statistics for TOEFL Writing Sub-section Scores by Condition*

Condition	N	Min	Max	Mean	SD	Std. Error	95% CI (Lower)	95% CI (Upper)
Individual	12	17	28	22.08	2.84	0.82	20.28	23.89
Dyads	25	18	26	22.20	1.96	0.39	21.39	23.01
Control	13	19	28	22.23	2.59	0.71	21.67	23.79
Total	50	17	28	22.18	2.34	0.33	21.52	22.84

Note. SD = Standard deviation, Std. Error = Standard Error

Before an ANOVA can be conducted a test of homogeneity must be met. Table 5.2 shows the results of the Levine's test of homogeneity as outputted by SPSS[3] version 19. The Levene's statistic was not significant at the $\alpha = .05$ level, which means the assumption of homogeneity was met, and an ANOVA test for assignment condition was warranted.

Table 5.2. *Levine's Test of Homogeneity for TOEFL Writing Sub-section Scores*

Levene statistic	df1	df2	Sig.
0.347	2	47	0.708

[3] All statistical analyses were conducted in SPSS version 19 for PC, unless otherwise noted.

Table 5.3 shows the results of a one-way ANOVA with the TOEFL iBT writing sub-section score as the dependent variable, and the condition assignment as the independent variable. The ANOVA test revealed a non-significant F-value, *F*(2, 48) = 0.14, *p*= 0.986, which indicates no significant difference in the mean scores of the TOEFL iBT writing sub-section scores for students in any condition. By this standard, the students in each condition can be considered as having the same level of writing ability as determined by their TOEFL iBT writing score. Therefore, random assignment to conditions was considered appropriate.

Table 5.3. *ANOVA Test for Condition Assignment*

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	.156	2	.078	.014	0.986
Within Groups	261.224	48	5.558		
Total	261.380	50			

Materials. The materials used in this study can be found in the Appendices sections, but will be briefly outlined here and their roles made clearer in a Table 5.4 for the study design. The directions for Day 1 can be found in Appendix A. On day 1, Graph 1 (Appendix B) was the prompt used to stimulate the participants' initial response in both the first (i.e., pretest) and third (i.e., posttest) stages of the study. Graph 2 (Appendix F) was used only in Stage 2, which represented the treatment stage. Graph 2 was designed as a resource for learning, as it was also accompanied by an expert model description. The expert model was developed with consultation from three experts who use graphics often: one engineering professor, one statistics instructor, and one professor of educational psychology. The researcher met with each expert individually to discuss what information should be included in a written description of Graph 2. Notes were taken and the commonalities were included in the model graph description. Once this was complete, the description was written and given to the experts to check if the model was a representation of a complete graph description. No expert disagreed with the description, although the two of the experts noted they would like to have seen a statistical test of the means to accompany the graph. Since the interpretation of statistical test is not part of the current research project this was not deemed appropriate for the research questions asked herein.

Both Graphs are grouped-column graphs, but there is an important difference between them. Graph 1 was designed with a target transfer feature embedded in it. The target transfer feature was the interaction between the location students buy their textbooks and the kind of textbook purchased. Graph 2 was not designed with such an interaction, and the trends for students who live with a roommate are the same for each sub-data set as well as the overall trend. This means, for Graph 2, the local trends in the subsets of data reflect the same overall trend for the whole dataset. The overall trend in Graph 1, however, requires learners to identify an interaction effect in the dataset though it is not described in the model description of Graph 2.

All learners wrote their responses in Stage 1 and 3 on the “writing sheet” (Appendix C). In addition to writing a description of Graph 1, learners in the treatment conditions took notes in Stage 1 on the “notes sheet” (Appendix D), and made notes on the model in Stage 2. Stage 2 and Stage 3 directions can be found in Appendix E. The directions and materials for the Control condition can be found in Appendix I.

Study Design. Table 5.4 shows the progression of the study by Day (located in the first column), and the treatments conditions across the last three column headers.

Table 5.4. *The Flow of events by Day and Treatment Conditions for Study Design*

		Treatment condition		
Day	Event	Control (n=13)	Individual (n=12[a])	Dyads (n=25, 12 pairs[b])
Day 1	**IRB forms**	read and sign IRB forms	read and sign IRB forms	read and sign IRB forms
	Stage 1 part 1 (Pretest: Invention)	write on Graph 1 (individually)	write on Graph 1 (individually)	write on Graph 1 (individually)
	Stage 1 part 2 Note taking	--	take notes on your essay (individually)	take notes on your essay (individually)
Day 2	**Stage 2 part 1**	worksheet with whole class discussion (see Appendix I)	look at Graph 2 with model response and take notes on worksheet (individually)	look at Graph 2 with model response and take notes on worksheet (individually)
	Stage 2 part 2 Collaboration	--	--	discuss notes and the model in dyads
	Stage 3 (Posttest: rewrite Graph 1)	rewrite a response to Graph 1 (individually)	rewrite a response to Graph 1 (individually)	rewrite a response to Graph 1 (individually)
	Effect of	**task repetition**	**model**	**collaboration**

Note. The “n =” values represent the participants retained in the current study, a = one learner did not complete the study, b = three learners did not complete the study.

This study has several important components to its design. One is that students wrote alone in all stages of the tasks, and they wrote two passages: one passage before they had an opportunity to see an expert model response (which served as a resource for learners as they worked alone or with a partner), and another passage after they discussed the model with a partner or thought about the model alone. However, while the expert model was a resource, it did not contain all the elements needed to identify the interaction effect present in Graph 1. This Stage 3 writing sample, then, acted as a test to estimate how well Stage 2 would prepare students to identify a novel complex global data integration and not simply mimic a model.

Given the description that Laughlin et al. (2003; Sears & Reagin, 2013) proposed on the role demonstrability plays in allowing learners to share ideas on a complex task, it is relevant to describe how the graph description (and the task design) used in this study has demonstrable characteristics. Table 5.5 shows a breakdown of the four elements outlined by Laughlin and Ellis (1986).

Table 5.5. *Rationale for Demonstrability of the Model*

Demonstrability condition: Group members must …	Rationale for the model as demonstrable
1. agree on the underlying conceptual system	The leaners know and can agree on the meaning of the graph and the purpose for the content of the model
2. have sufficient information to form and identify correct solutions	The learners are given a “correct” solution to the graph description task in the presentation of the model; they are also told the approximate number of sentences and words needed to describe the graph

Table 5.5 continued

3. be able to differentiate correct versus incorrect answers	Again, the learners are given a "correct" solution to the graph description task in the presentation of the model. If students did not know the meaning of the model graph, they were told to ask the instructor. No learner did this.
4. have sufficient ability, motivation, and time to demonstrate or explain a correct solution to their partners	All learners had 12 minutes to read over the model, then 15 minutes to discuss their answers with their partner. All learners completed the tasks, and no learner seemed out of time for explanations or sharing ideas.

Note. Adapted from Sears & Reagin (2013)

As the materials were the same for the individual condition and the dyad condition, both conditions can be said to have been in demonstrable conditions. Importantly, during the writing stages, the expert resources were removed and learners wrote their responses independently. Furthermore, as Table 5.6 and Table 5.7 show, the time on tasks were held constant across the treatment and the control conditions. This study was conducted under normal classroom conditions, so no audio or video recordings were taken; however, the researcher took notes on how the lessons went. A time line for the treatment condition for the tasks can be seen in Table 5.6, and a time line for the Control conditions is shown in Table 5.7.

Table 5.6. *Time Line for Individual and Dyad Conditions*

Day 1 (Stage 1)	Day 2 (Stage 2 & 3)
5 minutes for IRB	5 minutes set up
5 minutes setting up the activity for the day	12 minutes for reading model, and completing the worksheet.
8 minutes for note training	15 minutes to work with partner, or to work alone
12 minutes for pretest essay	12 minutes posttest essay
8 minutes for note taking	

Table 5.6 continued

Total class time	38 minutes for the task activities. 17 minutes for administrative chores, such as taking attendance. (55 total minutes)	44 minutes for the task activities. 11 minutes for administrative chores. (55 total minutes)

Table 5.7. *Time Line for Control Condition*

	Day 1 (Stage 1)	Day 2 (Stage 2 & 3)
	5 minutes for IRB	5 minutes set up
	5 minutes setting up the activity for the day	12 minutes for answering worksheet
	12 minutes for essay	15 whole class discussion
	5 minutes for closing comments and preparation comments for the following day.	12 minutes post-test
Total class time	27 minutes for the task activities. 10 minutes for administrative chores. Students were let out early as per the course instructor's decision. (37 total minutes)	44 minutes for the task activities. 11 minutes for administrative chores. (55 total minutes)

5.4 Procedures

Participants worked in their assigned classrooms. The researcher taught each class to ensure equal administration across groups, and that participant questions were answered uniformly. After a brief introduction of why data is important for supporting arguments, all participants in the treatment conditions received instructions on how to take notes on their essay. This was done following Hanaoka's (2007) design, but was employed here as a means to prepare learners for the metacognitive activity of re-reading one's own writing while looking for areas to improve their writing (i.e., noticing the gaps; Hanaoka). The notations made by the learners were not a variable in this study, but were employed here as a pedagogical ends (i.e., pushing leaners to identify where they could improve). That said, the instructions for note taking occurred before learners wrote their initial response so they could take notes while writing their initial responses. If this followed the writing of their initial response, learners might have forgotten what they struggled with. This note-taking task was also employed to help them see Day 2's events as a useful extension of Day 1. Learners in the Control condition did not receive this training, as they did not take notes. Instead, learners in the Control condition received a small talk on the value of data in argumentative essays, wrote their initial responses, and received instructions for Day 2. Subsequent to these events, they were let out of class early as their normal classroom instructor asked the researcher to do.

On Day 2, the treatment groups were reminded of the importance of interpreting and describing graphs for argumentative essays. Then they were given instructions on the importance of understanding Graph 2 and the expert solution they were about to be provided. They were specifically told to try and understand the expert model well enough so they would feel comfortable writing another graph description at the end of the lesson. This gave them cause to try to learn the graph description well, and to not take the task as just an activity for the present. This also let them know that the knowledge they would get from the activity would be useful to them in the task that followed.

Graph 2 with an expert solution was provided to the treatment conditions, and a worksheet was provided to them so they could focus on comprehending the expert model. They were told the worksheets would be collected; thus, there was impetus for them to try as hard as they could on the task, as they did not know if the worksheet would be scored. Both the individuals and the dyads were encouraged to understand as much as they could and were told to ask the researcher if they had any questions about Graph 2 or its description. After 12 minutes passed, individuals were told to read silently to themselves each line of the description, and to think carefully about the graph, considering how they would apply their ideas about what they might add or remove from the model. In the Dyad condition, and after 12 minutes, the dyads were told the same directions as the Individual condition, except they were given directions to pair up with anyone in the class they wanted to work with. Instead of reading silently, though, they were told to read aloud each line of the model, taking turns with their partner who would check the content of their reading aloud. Then they were told to discuss the final two questions of the worksheet, which summarize to: (a) do you believe the data in the graph, and (b) what would you add or remove from the model. For this part of Stage 2, the individuals and the dyads each had 15 minutes. After 15 minutes passed, the dyads were asked to stop talking and to return to their seats. The individuals were warned that the final activity was about to occur. The experimenter collected all the Stage 2 materials, and distributed a clean Graph 1 (i.e., not the Graph 1 used on Day 1 as it had writing on it from the note taking task). A new "writing sheet" was also given to each person. They had 12 minutes to produce another description of Graph 1. No students had access to materials from Day 1 or Stage 2.

For the Control condition, on Day 2, they began with a reminder about the topic of Graph 1 (i.e., people buy books online and in physical locations like the campus bookstore). Then they were asked about their experiences using technology in class, and references were made to Graph 1 (e.g., people have begun to rely on the Internet more to buy their textbooks). Then they were given a worksheet (Appendix I) for completion; this took 12 minutes. After that, the researcher led a 15-minute class discussion on their responses to the worksheet. This was done by randomly calling on students for their responses. To each reply given by a student, the researcher offered his perspective. For example, when a student mentioned using BlackBoard, the researcher noted how he used BlackBoard in the classes he taught and how he thought BlackBoard was useful. The researcher engaged learners individually, but did not allow the learners to engage with each other. Most students sat still, waiting for their name to be called next. After 15 minutes passed, the researcher collected the materials and distributed Graph 1 with a "writing sheet." The students were given 12 minutes to complete another description of Graph 1.

For all conditions, the learners fully engaged with all tasks. Even on the final task, for which they would write *another* description of Graph 1, the learners seemed eager to have a second chance to respond. This was especially important for the Control condition because they represented task repetition. Since they had a meaningful lesson on the uses of technology for college students, they were not under the impression that they were repeating the task for the sake of repetition. Notably, a key difference between treatment conditions was that participants in the Individual condition invented a solution to Graph 1, and then received an expert solution for an isomorphic graph without any collaborative element. In other words, the learners in the

Individual condition had only the influence of the expert model, whereas, the learners in the collaborative condition had the support of their peer while considering the expert solution.

By attempting an initial solution to Graph 1, before seeing the expert solution (i.e., the model), the participants were expected to have a challenging time describing Graph 1. In Stage 2, they were able to contrast the expert example with their experience of their original attempt, but without looking at their original essay. Stage 2 also provided them an opportunity to compare the two subsets of data while looking at how the expert model described the data trends. By working without their original essays, learners in both treatment conditions were put to the task to resolve their own issues through reflection by considering the content of the expert model as containing solutions for them to find. It was hypothesized that learners in the Collaborative condition would be more prepared to transfer what they noticed and understood from the expert model, as they had to discuss the expert solution considering another person's perspective, making their interpretations of the model more flexible (Schwartz, 1995).

The Individual condition participants were put to the task of silently reading the model, and then mentally considering interpretations of the solution given to them. They did not engage with others, and they were not pushed to make their understanding of the expert solution more generalizable to others, like the dyads had done. It was hypothesized that the Individual condition would not likely form a broad understanding of the expert solution, but would likely focus on surface features of the canonical solution (such as the local data integrations, which is the most easily interpreted facet of a column graph; Carswell et al., 1992; Cleveland & McGill, 1985; Shah, 2002).

5.5 Measures

Two approaches to coding the data were taken. The first approach employed a coding scheme that provided a percentage of content overlap with the expert model—an indicator of factual recall and transfer. This was done by a line-by-line coding scheme (Glaser, 1978). The content overlap was calculated on a percentage of 37 specific points available in the original model (coding is discussed below). For this scheme, items of the 37-point scheme were only counted once, so the content overlap is exclusive of redundancy and representative of the percentage of recall. This percentage was then used to calculate a change score, and was subjected to an ANOVA to test the change in content overlap from the pretest to the posttest for a between groups comparison. The second approach considered how well the essays “fit” the expected data integrations provided in the model—an indicator of transfer of deep writing structure based on the relative balance of global versus local integrations. This approach used the model as a template for the expected number and kind of data integrations needed in a complete graph description. The coding schemes for each of these are described below.

5.6 Coding and reliability

A 37-point coding scheme was developed to demonstrate the amount of content overlap with the model. Since this study used an isomorphic model as the expert solution, equivalence must be made between what was given to the learners in Stage 2, and what the researcher expected as a possible response in Stage 3. Table 5.8 shows sentences numbers in column one that correspond to the sentences from the expert model for Graph 2 (found in column 2) and its converted text (found in the final column) for Graph 1.

Table 5.8. *Comparison of Isomorphic Model Sentences from Stage 2 Converted to Target Sentences for Stage 3*

#	Sentence from expert model for Graph 2	converted sentence to match Graph 1
1	Figure 1 is a grouped-column graph that shows the survey results for the living situation of 2,350 freshmen who live alone or with a roommate by student living location.	Figure 1 is a grouped-column graph that shows the survey results of 195 students from one physics class who bought a new or used physics textbook by bookstore location.
2	The number of freshmen is on the vertical axis and both the living location and the living situation are on the horizontal axis.	The number of freshmen is on the vertical axis and both the bookstore location and the book condition are on the horizontal axis.
3	The purple columns represent the freshmen who live in campus apartments, and the green columns represent the freshmen who live in private homes.	The blue columns represent the students who bought books in the campus bookstore, and the red columns represent the students who bought books from online bookstores.
4	For campus apartments, 1000 freshmen live with a roommate and 650 freshmen live alone.	For campus bookstores, 60 students bought used books and 40 students bought new books.
5	This means a majority of freshmen in campus apartments live with a roommate.	This means the majority of students bought used books from the campus bookstore.
6	For private homes, this pattern is the same with 500 freshmen living with a roommate and only 200 living alone.	For online bookstores, this pattern is reversed with 80 students buying new books and 15 students buying used books.

Table 5.8 continued

7	If the two living locations are compared, it can be seen that living in campus apartments is more common than living in a private home.	If the two bookstore locations are compared, it can be seen that buying books from the campus bookstore is more popular than buying books online.
8	Regardless of the living location, however, living with a roommate is more popular than living alone.	Overall, buying new books is more popular than buying used books.

Using the corresponding sentence from Table 5.8 above, Table 5.9 below identifies the scoring system and rationale for each sentence. The converted sentences are in the second column, and the targeted information has been bolded. Each bolded element has a subscript that represents the points scored for each element. In the final column of Table 5.9, a brief comment is offered on what students must do to earn points for the targeted elements. In this scoring system, it is important to make note that only the tokens of interest found in the model were coded in this 37-point scheme. For example, if a student summed the values for the online bookstore and wrote, “95 students bought books from the online bookstore” it was not scored. This is because the subtotals of the data subsets were not included in the expert model, and the 37-point was designed as a measure of recall of the model and transfer of the surface features of graph description. The second coding scheme accounts for additional information included by learners, but this is discussed in later sections.

Table 5.9. *Content Overlap Scoring System with Points and Rationale for each Sentence*

Sentence	Converted sentences and targeted information	points possible	comments
1	**Figure 1** $_{1pt}$ is a **grouped** $_{1pt}$ **-column graph** $_{1pt}$ that shows the survey results of **195 students** $_{1pt}$ from **one physics class** $_{1pt}$ who bought a **new or used** $_{1pt}$ physics textbook **by bookstore location** $_{1pt}$	7 points	must identify the type of graph, sum the total students, and use the information in the figure name. Points were awarded for “old” in place of “used”
2	**The number of freshmen** $_{1pt}$ is on **the vertical axis** $_{1pt}$ and both **the bookstore location** $_{1pt}$ **and the book condition** $_{1pt}$ are on **the horizontal axis** $_{1pt}$	5 points	must fully identify the x- and y-axis. The x-axis needed both the bookstore location and the book type, but points were assigned for each. Points were awarded for synonyms like, “x-y-axis, text book, text book style, and type of book”
3	The **blue columns** $_{1pt}$ represent the students **who bought books** $_{1pt}$ **in the campus bookstore** $_{1pt}$ and the **red columns** $_{1pt}$ represent the students **who bought books** $_{1pt}$ **from online bookstores** $_{1pt}$	6 points	must identify the two colors, that students are buying books, and the two bookstore locations. Points were awarded for “took, chose, selected, and preferred” in place of “bought.” Also, “Internet,” “offline,” and “Web” were accepted for “online”
4	For campus bookstores, **60 students** $_{1pt}$ **bought used books** $_{1pt}$ and **40 students** $_{1pt}$ **bought new books** $_{1pt}$	4 points	must identify the specific number with the book type. Points were awarded for “took, chose, selected” in place of “bought”
5	This means the **majority of students** $_{1pt}$ **bought used books** $_{1pt}$ from **the campus** $_{1pt}$ **bookstore** $_{1pt}$	4 points	must identify the trend (e.g., most students, the majority), the book condition, and the location because there are two bookstores.
6	**For online bookstores this pattern is reversed** $_{1pt}$ with **80 students** $_{1pt}$ **buying new books** $_{1pt}$ and **15 students** $_{1pt}$ **buying used books** $_{1pt}$	5 points	must identify the trend (“the pattern is …” “not the same”, “…different,” “…opposite” were accepted), the number of books sold by condition, and the selling location.

Table 5.9 continued

7	**If the two bookstore locations are compared** $_{1pt}$ it can be seen **that buying books from the campus bookstore** is **more popular** $_{1pt}$ **than buying books online** $_{1pt}$	3 points	must address the overall trend for amount of sales by bookstore location. Credit was given for, "thus, therefore, it is obvious," etc. in place of the conditional "If", and credit was given for "it is more common to" and "students prefer to" in place of "more popular"
8	**Overall** $_{1pt}$ **buying new books** $_{1pt}$. **is more popular than buying used books** $_{1pt}$	3 points	must identify the overall trend for the condition of book sold. Credit was given for, "however, in conclusion, finally, in the end, in sum" etc.

Note. Bolded portions were targeted ideas and phrases, subscripted values identify point values.

Inter-rater reliability for the coding-scheme in Table 5.9 was conducted. An EFL instructor with 15 years of language teaching experience and a master's degree in Applied Linguistics volunteered to read and rate a subset of the essays. After a one hour training session in which 5 randomly selected essays served for training, a random sample of 15 pretests and 15 posttests (or, 30% of the total essays collected), with five essays per condition, was selected for inter-rater reliability calculations.

Across the eight sentences for all 30 essays and the two raters (30 essays x 2 raters = 240 rated items), the overall percentage of exact agreement was 92.90%. For the sentence level analysis, chance of agreement between the raters was considered a potential confound, so the Kappa statistic was employed (Landis & Koch, 1977). The Kappa statistic is also useful in this case because not all items had the same number of possible ratings (Landis & Koch); however, to assist conceptual comparisons between raters the exact agreement for ratings across the 240 items were also calculated (see the final column in Table 5.10). Chance of agreement can be simply understood as the chance that the two raters would agree on a rating because there are a limited number of ratings possible for a given item. Landis and Koch suggested benchmarks for "strength of agreement" using the Kappa statistic, stating that values between 0.61 – 0.80 can be considered "substantial" levels of agreement, and values between 0.81 – 1.0 can be considered "almost perfect" (1977, p. 165). Table 5.10 shows the exact Kappa coefficients and the exact percent of rater agreement for each of the eight sentences.

Table 5.10. *Kappa Agreement Coefficients and Percentage of Exact Agreement for Content Overlap Coding*

Sentence	Kappa coefficient	Percentage of exact agreement
1	.831	86.7 %
2	.933	96.7 %
3	.872	93.3 %
4	.852	90.0 %
5	.849	93.3 %
6	.858	90.0 %
7	.876	93.3 %
8	1.00	100.0 %
Average	.8838	92.91%

Across all eight sentences, the average percent agreement was 92.91% (average Kappa = .8838; with a minimum exact agreement of 86.7% on any given sentence). Thus, the coding scheme can be considered reliable across raters, as the ratings surpass Landis and Koch (1977) suggested benchmark of 0.61 – 0.80 for "substantial" agreement between raters, and all items are contained within the "almost perfect" range of kappa coefficients 0.81 – 1.0 (p. 165). Table 5.11 provides examples of learner texts and the ratings given to them.

The items in Table 5.11 are representative of items that were agreed on by both raters. In the first column, the corresponding sentence from the expert model text is referenced by sentence number (see Table 5.8). Thus, column one may have more than one number, as more features from across the expert model were represented in the sample sentence (found in column two). It is important to note that scores were distributed by the presence of a feature, not the order by which the learners produced them. This means that learners received credit for incorporating parts of the model text regardless of where they incorporated it, just as long as the feature was present. Another important point, however, is only one point was given per feature. This means that if learners repeated a phrase, they received a score for only one instance (more on this later). Furthermore, neither grammatical accuracy nor spelling played a role in the scoring system. Looking at Table 5.11, the points awarded are in column three (for the sentence in column two), and column four offers some comments on the scoring.

Table 5.11. *Sample Scoring for Content Overlap with the Model*

Model sentence	Sample learner sentences	points scored	comments about scoring overlap
1	"Considering both the graphs together, the survey was conducted with **a total of 195 students.** $_{\textbf{1pt}}$"	1	1 point for communicating population elements from model sentence 1.

Table 5.11 continued

1,7	"**Students from physics classes** $_{1pt}$ are **more likely to buy books in campus bookstores** $_{1pt}$ **than online bookstores**. $_{1pt}$"	3	1 point for information about student population from model sentence 1, 1 point for identifying the trend, and 1 point for information about book condition and store location—each from model sentence 7.
2	"The **number of students** $_{1pt}$ **lie on the Y axis** $_{1pt}$"	2	2 points for including essential vertical axis information from model sentence 2.
2	"The **number of students** $_{1pt}$ is on the **vertical axis** $_{1pt}$, and **the location of bookstores** $_{1pt}$ **and status of books** $_{1pt}$ are on the **horizontal axis** $_{1pt}$."	5	5 points for including all information describing the layout of the graph, similar to model sentence 2.
3	"From the **red graph** $_{1pt}$, **we could see students** $_{1pt}$ will **buy books online.** $_{1pt}$"	3	3 points for including partial information similar to model sentence 3.
3	"The **blue** $_{1pt}$ is **students who buy books** $_{1pt}$ in **campus bookstore** $_{1pt}$ and the **red column** $_{1pt}$ is **students who buy books** $_{1pt}$ in **online bookstores.** $_{1pt}$"	6	6 points for including all information related to one subset of data, analogous to model sentence 3
3,4	From **blue graph** $_{1pt}$ we can see **the students choose to buy books in campus bookstores** $_{1pt}$ and **40 of them** $_{1pt}$ prefer to **buy new books** $_{1pt}$ and **60 of them** $_{1pt}$ would like to **buy used ones** $_{1pt}$."	6	2 points for information describing the data subset, comparable to model sentence 3, and 4 points for information describing the data trend, similar to model sentence 4.
4,5	"For **campus stores** $_{1pt}$, **60 people** $_{1pt}$ **bought used books** $_{1pt}$ and **40** $_{1pt}$ **bought new books** $_{1pt}$, so **most students** $_{1pt}$ **prefer used books** $_{1pt}$ at this store."	7	4 points for information describing a subset of data, similar to model sentence 4, and 3 points for highlighting the trend in the data subset, akin model sentence 5.
6	"**The preference of students who buy books in online bookstore just is opposite**. $_{1pt}$"	1	1 point for identifying the trend from model sentence 6.
6	"**80 of them** $_{1pt}$ **decide to buy new books online** $_{1pt}$, and only **15 students** $_{1pt}$ **choose to buy used books online**. $_{1pt}$"	4	4 points for specific selling information from model sentence 6.
8	**"All in all** $_{1pt}$ **new books** $_{1pt}$ **are the most popular by the rate 120/195.** $_{1pt}$**"**	3	3 points for all information from model sentence 8.

While points were scored for the presence of information across sentences, several sentences received no rating, or at best limited points. A sample set of these sentences in Table 5.12 shows how some sentences did not produce information from the expert model, and were therefore not scored. The first column offers the sample sentence (with portions receiving points bolded and subscripted with point values), and the final column in Table 5.12 offers a comment on why points were or were not awarded.

Table 5.12. *Sample Learner Sentences Receiving Some or No Points*

Sample learner sentences	Comment
"**Students who bought online 80** $_{1pt}$ **bought new books** $_{1pt}$ **and the rest** $_{1pt}$ **is bought used books** $_{1pt}$ which is surprised to me because I never buy used book online."	This sentence received points for the data description, but not for the supporting narrative parts.
"**60 students** $_{1pt}$ **buy used book** $_{1pt}$ **at campus store** $_{1pt}$**, and 40 students** $_{1pt}$ **buy new book**$_{1pt}$ there. [….] **80 students** $_{1pt}$ **buy new book** $_{1pt}$ **online stores** $_{1pt}$**, and 15 students** $_{1pt}$ **buy used book**$_{1pt}$. [….] At campus stores, 40 students buy new book but 80 students buy new book online because new books online are cheaper.	Points were awarded for the data description, but not for the summary statement that followed later in the passage.
"This is reasonable because online is cheaper and some professor use BlackBoard for handout with TA."	No points awarded for personal reflections and experiences.
"I much prefer the new book purchase online because I can stay home and because used books are not reliable online."	No points awarded for opinions

Note. [….] = deletions made for brevity.

The example below shows a whole text, so the points awarded can be seen in context. The example is a pretest that was rated nine out of a possible 37-points (or, 24%). The bolded sections were scored one point each, which is also indicated by a subscripted point value.

> According to a survey conducted on **the preference of buying textbook of physics students** $_{1pt}$, students showed different preference on buying textbook in different location. **40 students** $_{1pt}$ choose to buy **new textbooks** $_{1pt}$ in **campus bookstore** $_{1pt}$ while **other 60** $_{1pt}$ choose the **used textbooks** $_{1pt}$. However, **80**

> **students** $_{1pt}$ choose **new textbooks** $_{1pt}$ in the **online bookstore** $_{1pt}$. This is also my way of buying books. [pretest for ID:51020; 9 points of 37 possible points = 24%]

From a language-teaching stance, some points awarded in the content overlap coding scheme described above may seem prescriptive, but it is important to understand that this point system was developed to provide the maximum amount of points possible to students. Recall that points were awarded regardless of where the element occurred in the essay, just so long as the element was present. Repetition of meaningful elements were not considered in this analysis because the second method of analysis provided a method for counting repetition.

The initial 37-point scoring system checked essential elements that characterized a thorough description of a grouped-column graph, as presented in an expert solution. The model description included introducing the kind of graph, and identifying very specific information, such as noting that “15 students bought used books” not just stating “there are 15 used books,” which does not reflect what the graph represents. While the first coding scheme checked for content overlap, the second coding scheme checked for overlap with the underlying structures by coding for the global and local data integrations. As mentioned previously, column graph tasks require students to identify deep structure trends through global integrations (i.e., data interpretations that include all data points), and surface structure trends through local data integrations (i.e., comments on data that do not incorporate all data points, or are limited in their interpretations). That said, not all graphs have the same amount of global and local integrations. It is common, for example, for column graphs to have more local integrations than global integrations, especially when the data are multivariate (Carswell et al., 1992). Furthermore, in column graphs, global structures are less obvious than local structures to most readers (Shah & Hoeffner, 2002).

The following examples in Table 5.13 offer the coding scheme for the local and global data integration points. The purpose of the global and local distinctions in data integrations can be seen as a different approach to the discrete 37-point coding system for explicit recall of content. Coding of data integration types provides a deep structure view of the passages, and allows the Control condition to not be disadvantaged by not having exposure to the target model. Thus, each sentence from the expert model was converted to target information that could be found in Graph 1, which is offered in column two of Table 5.13. Italics and super-scripts indicate if the token was a local integration (i.e., "^{L}"), and bolded texts with sub-scripts indicate a global integration (i.e., "$_{\mathbf{G}}$"). An explanation of the target information is given in column three, and the number of tokens is counted by category of integration in the remainder of the Table.

Table 5.13. *Description of Target Sentences Used for Coding by Data Integration Type*

			Integration count	
Sentence	Target information	Explanation for classification	Local	Global
1	Figure 1 is a grouped-column graph that shows the survey results of **195 students** $_{\mathbf{G1}}$ from one physics class who bought a new or used physics textbook by bookstore location.	G1 incorporates all of the data points in the graph (Carswell et al., 1992)	0	1

Table 5.13 continued

2	*The number of freshmen is on the vertical axis* [L1] and **both the bookstore location and the book condition are on the horizontal axis** [G2]	L1 identifies a scale for the data and is not part of any trends, so it cannot be interpreted without other information (Carswell et al.). G2 identifies all possible categories in the multivariate data (Shah & Hoeffner, 2002)	1	1
3	The *blue columns represent the students who bought books in the campus bookstore* [L2] and the *red columns represent the students who bought books from online bookstores* [L3]	L2 and L3 individually identify an abridged portion of the data available; neither carry information about trends across all categories which make them local data points (Carswell et al.; Shah & Hoeffner). Despite identifying all the colors available, colors do not carry numeric information clearly, so they are not considered global descriptors of the data (Cleveland & McGill, 1985).	2	0
4	For campus bookstores, *60 students bought used books* [L4] and *40 students bought new books* [L5]	L4 and L5 describe a restricted portion of the data available, and do not carry information about trends across the data for the online stores (Carswell et al.; Shah & Hoeffner). See # in the ***Note*** below.	2	0
5	This means *the majority of students bought used books from the campus bookstore* [L6]	L6 identifies a trend in a subset of the whole data presented (Carswell & Wickens, 1987; Carswell et al.).	1	0
6	*For online bookstores this pattern is reversed* [L7] *with 80 students buying new books* [L8] and *15 students buying used books* [L9]	L7 describes a local trend related only to the online stores, without comparing specifics of the data (e.g., the pattern might be reversed, but there may be more books of each sold); thus the trend is only locally interpretable (Carswell et al.; Shah & Hoeffner). L8 and L9 describe a restricted portion of the data available (Carswell et al.; Shah & Hoeffner)	3	0
7	If the two bookstore locations are compared it can be seen that **buying books from the campus bookstore is more popular than buying books online** [G3]	G3 identifies a trend in the data that incorporates all of the data points given in the graph (Carswell et al.).	0	1

Table 5.13 continued

8	Overall **buying new books is more popular than buying used books** $_{G4}$	G4 identifies a trend in the data that incorporates all of the data points given in the graph (Carswell et al.).	0	1
Total			9	4

Note. Local data integrations are *italicized* and superscripted for reference with "$^{L\#}$" and global integrations are **bolded** and subscripted for reference with "$_{G\#}$". # If students included a summary data point, such as "For the **100 students** who purchased books at the book store, 40 of them…" a local integration was not added and the summary data description, and these were left out of the analysis. This was done because it was not available in the isomorphic expert model.

A summary of Table 5.13 above, shows there are nine local data integrations and four global integrations. Taken together, this means there were 13 possible data integration points available from the solutions provided in the expert model. Coding for these data points was conducted using auto-coding text queries available in QSR International's NVivo 10, qualitative data analysis software for Windows. This was done using key words and numbers (See Table 5.14). For example, the key word "blue" brought up sentences related to the local integration point for describing the data related to the campus bookstore purchases. The key word, "195" brought back terms related to the global reference for the population of the students surveyed. However, there were some exceptions. Looking at Table 5.15, for example, the last entry uses a ratio "120/195." In this case the "195" of the "120/195" was not coded as a global data integration point for the overall population because it was not being used as such in context. Rather, it was being used as a part of another global trend—the overall trend for condition of book purchased. Fortunately, too, NVivo has a synonym search function, which greatly increased the efficiency of the search. This search function looks for synonyms; for example, searching for “college” will also bring up “university,” and “school.” Once key terms were searched, each token was investigated to see if it fell into the global or local categories (see Table 5.13 for categories). Another useful feature of using NVivo’s auto-coding is NVivo does not double code data with the same code (Bazeley & Jackson, 2013), which further limits concerns for over coding an essay through auto-coding. Table 5.14 shows the 42 search terms used to query the data sets in NVivo.

Table 5.14. *Search Terms Used with Synonym and Wildcard Functions*

Search terms
15, 40, 60, 80, 195, 100, 120, all, axis, blue, book, campus, color, conclusion, down, every, few, finally, fresh, horizontal, Internet, majority, many, most, new, normal, offline, old, online, overall, perpendicular, red, regular, several, sum, text, textbook, total, up, used, vertical, web

Sample sentences from learner texts, and the coded elements are offered in Table 5.15. The sample sentences have bolded portions, which represent the coded portion. These portions of text are also labeled with subscripts to identify their nature as global (G) or local (L). The final column offers a description of the coding. Distinct from the 37-point coding scheme used above, if a learner had multiple data integrations from repetition, each instance was coded. Doing this provides a holistic perspective of what was included in the learner texts, which may not have been evident in the 37-point coding system for content overlap. To restate, the simple distinction of these two measures is content knowledge (number of points obtained out of 37) versus writing structure (relative use of global vs. local integrations).

Table 5.15. *Sample Coding for Local and Global Integrations*

Sample learner sentences	Coded as
"Considering both the graphs together, the survey was conducted with **a total of 195 students.** $_{G1}$"	1 global element for communicating the total population of students
"Students from physics classes **are more likely to buy books in campus bookstores than online bookstores.** $_{G1}$"	1 global element for communicating the overall trend total population of students
"The **number of students lie on the Y axis** $_{L1}$"	1 local element for delineating a restricted variable
"The **number of students is on the vertical axis** $_{L1}$, and **the location of bookstores and status of books are on the horizontal axis** $_{G1}$."	1 local element for delineating a restricted variable of the y-axis, and 1 global element for communicating the total possible data sources for multivariate data
"From the **red graph we could see students who buy books online.** $_{L1}$"	1 local element for defining the red graph is for students who buy textbooks online
"The **blue is students who buy books in campus bookstore** $_{L1}$ and **the red column is students who buy books in online bookstores.** $_{L2}$"	2 local elements. 1 each for delineating the meaning of the red and blue colors.
"For campus stores, **60 people bought used books** $_{L1}$ and **40 bought new books** $_{L2}$, so **most students prefer used books at this store** $_{L3}$."	3 local elements. 1 for describing the total trend for a subset of the data (i.e., L3), 1 each for specifying the new and used trend of buying in campus bookstores
"**The preference of students who buy books in online bookstore just is opposite**. $_{L1}$"	1 local element for identifying a local trend in the online bookstore graph
"**80 of them decide to buy new books online** $_{L1}$ and only **15 students choose to buy used books online** $_{L2}$."	2 local elements. 1 for describing the data for new books, and 1 for specifying the data for used books
"All in all new books are the most popular by the rate 120/195. $_{G1}$**"**	1 global element for identifying the overall trend including all data points available

The following are excerpts from one student's pretest and posttest, and key phrases and words have been italicized and subscripted to indicate the local data integrations and the bolded portions are global tokens.

> This graph shows the data of the same physics class who purchased a new or used physics textbook from campus bookstore and online bookstores. As you can see, *around 50% of students choose to buy books in campus bookstores*$_{L1}$, and *the rest of them got their textbook from online bookstores*$_{L2}$. Someone who purchased textbooks *in campus bookstore prefer to buy used textbook*$_{L3}$, and *the rate is about 60%*$_{L4}$. However, *more than 80% of students choose to purchase in online bookstores bought new textbook*$_{L5}$. [pretest,ID:20918, 5 local, 0 global]

For the tokens noted in sample 20918A, there are only five local data integrations that describe only a subset of the whole dataset. There are no global data integrations in this pretest sample. The following posttest example is from the same student and shows both global and local data integrations. Data integrations that refer to the same feature, but stretch across a sentence or more, are marked with the same subscript number and are underlined for emphasized, such as L4 "The blue columns represent$_{L4}$ … campus bookstores$_{L4}$."

> Figure 1 is a group column graph that shows ***195 students***$_{\mathbf{G1}}$ choices for buying a textbook from campus bookstores or online bookstores, and the status of the books. *The number of students is on the vertical axis*$_{L1}$, and ***the location of bookstores and status of books are on the horizontal axis***$_{\mathbf{G2}}$. *The blue columns represent*$_{L4}$ *40 students who buy new textbooks*$_{L2}$, and *60 students who buy used textbooks*$_{L3}$ from *campus bookstores*$_{L4}$. *The red columns represent*$_{L7}$ *80 students who buy new textbooks*$_{L5}$ and *15 students who buy used textbooks*$_{L6}$ from *online bookstores*$_{L7}$. Based on comparing columns, *about 100 students choose to buy textbooks from campus bookstores*$_{L8}$, and *95 students choose to buy textbooks from online bookstores*$_{L9}$, *so* ***more student use campus bookstore***$_{\mathbf{G3}}$. In addition, *more people prefer to buy used textbooks from bookstores*$_{L10}$, however, *a majority of people would purchase new textbooks from the online bookstores*$_{L11}$. ***All in all new books are the most popular book type purchased***$_{\mathbf{G4}}$. [posttest, ID: 20918, 10 local, 4 global]

What can be seen across these two examples are two important things, which brings us back to the two measures used. First, a lot more data points being covered, more facts are being identified. Second, and that is the purpose of this second analysis, a lot more global and local data points are being covered. In other words, the posttests more closely resemble the expert model. Using only a content overlap scheme does not present a whole picture of the differences in pretest and posttest essays. For example, students could increase their content overlap score by simply including information from the local data integrations. However, as repetition was not counted in the content overlap scheme, it was important to include an analysis of the types of data integration. After all coding was complete, the tallies were organized into a coding matrix using NVivo's node query function, and pre- and posttest groups for each of the three conditions were organized (available in the Results section).

5.7 Results

There are two primary sets of results. One set presents the "content overlap results" of the 37-point system (representing the amount of content overlap with the solutions from model). The other set considers the goodness of fit for the local and global data-integration (i.e., the "data-integration results" representing how well the learners incorporated data points in terms of matching those found in the model). Each of these results provide a different perspective of how learners used the model. The results will begin with the content overlap using the 37-point coding scheme, which is followed by the data-integration results.

Before the results are presented, however, it is important to make clear that the learners understood the model. This is important because both the 37-point coding scheme and the data-integration coding scheme are based on the assumption that learners understood the content of the expert solution given to them. There are two pieces of evidence for this. The first, no learners seemed to have struggled with the model during the lessons, and no learners asked for help understanding the expert solutions offered in the model. The second is data collected from Stage 2. In Stage 2, learners were asked to write down their understanding of how each sentence in the model was needed, or how it was connected to other sentences in the model. Their responses were scored for their acceptability by the researcher. Inter-rater coding produced an exact agreement ratio = 98%, with only 2 disagreements across 10 worksheets (i.e., 100 items across five worksheets from each dyad and individual conditions). For each treatment group (the control group did not have this worksheet), no learner provided an incorrect interpretation of the model. The disagreements in the inter-rater agreement came from the use of speech-marks to represent “same as above.” Some examples of the acceptable responses are: “this sentence describes bookstore data,” “it tells the graph kind,” “this shows the trend for data.” The only responses that were marked inappropriate were sentences left unaddressed by a blank response. The percent correct are shown in Table 5.16. This Table shows that learners understood what each sentence in the model meant, and how the sentences were related to Graph 2. This means the expert model was a demonstrable source of input for the learners, and misunderstanding the content of the model did not seem like a confound in the results that follow.

Table 5.16. *Percentage for Acceptable Responses to the Stage 2 Worksheet for Individuals and Dyads*

Worksheet Sentence	Individual (n=12)	Dyad (n=25)
1	1.00	1.00
2	1.00	1.00
3	0.92	0.96
4	1.00	1.00
5	0.92	0.96
6	1.00	0.96
7	1.00	1.00
8	0.92	0.96
Open-ended Questions		
9	1.00	1.00
10	0.92	1.00

For the results that follow, they will be presented with the research question they are related with.

5.8 Content overlap results

The changes in scores received by learners on the 37-point scoring system were subjected to an analysis of variance. However, in order to avoid inflating the sums of squares for the variance of the dyad condition (i.e., there are more learners in this condition so they have more variance) each dyad was merged into one entity, a common procedure in psychological studies (McNemar, 1955, pp. 148-149). The aggregated data point also allows for a more conservative estimate of change, as the average change for the dyads is calculated. To aggregate the data points, the aggregate option under the data tab in SPSS was used. Notably, the result of the aggregation changed the population of the dyads from 25 persons to 13 persons; however, the Individual and Control conditions were not aggregated, so their populations did not change.

The process of aggregation generates a change score by aggregating the pretest and the posttest scores of the dyads. The difference between the posttest and the pretest is the averaged-change in content scores for the dyad. Thus, if one student changes ten points higher on the posttest, and their partner changes 18 points higher, SPSS aggregates their change score into the average of the sum of their gain scores, which in this case would be: (10 + 18) / 2 = 14. The descriptive statistics for this data are given in Table 5.17.

Table 5.17. *Descriptive Statistics for Percent Overlap with the Model by Test and Change Scores*

Condition	*n*	Pretest mean ± *SD*	Posttest mean ± *SD*	Change mean ± *SD*	Cohen's d	95% CI for mean change	
						lower	upper
Individual	12	.09 ± .05	.46 ± .11	.37 ± .17	4.33	.30	.45
Dyad	13	.10 ± .03	.75 ± .13	.65 ± .15	6.89	.55	.73
Control	13	.14 ± .06	.36 ± .11	.22 ± .10	2.48	.16	.28

Note SD=Standard deviation

Since learners did not all receive a zero percent on their pretests, the starting point for content overlap with the model was different for each condition. This made it important to calculate the influence of the amount for content overlap in terms of the change in content overlap. This could be considered the percent of content transferred from the model to the posttest because the amount of overlap with the model from the pretest has been removed. To evaluate if the amount of content overlap changed significantly between the groups, a univariate analysis of variance was conducted with the dependent variable being the change in content score, and the independent variable being the condition. Before the results of the ANOVA can be examined, a Levene's test was conducted to check if the error variance of the dependent variable was equal across groups. Table 5.18 shows the Levene's test of homogeneity was not significant, and a group-wise analysis of variance was conducted to determine if being assigned to a condition influenced the amount of overlap with the model. This result can be seen in Table 5.18.

Table 5.18. *Levene's Test of Homogeneity for Change Scores*

F	*df1*	*df2*	Sig.
0.564	2	35	0.574

Table 5.19 below shows a large F-value, which indicates a significant difference in the group-wise comparisons for overlap with the model. This means that being assigned to one condition or another had a significant influence on the amount of change that manifested in the posttest writings.

Table 5.19. *Results for Univariate ANOVA of Group-wise Comparisons*

	Sum of Squares	*df*	Mean Square	*F*	Sig.	η^2
Contrast	1.194	2	0.597	39.19	.000	0.691
Error	0.533	35	0.015			

α = .05

The partial eta (η^2) value in the last column of Table 5.19 shows that .691 (which is equal to 69%) of the variance in the data can be attributed to the condition in which learners completed the tasks. With well over half of the variation in the change of content scores being attributed to the condition in which the students completed their tasks, a post hoc pairwise comparison was warranted for a parsimonious analysis of which groups may be contributing to this significant variation. As there are three conditions (i.e., Control, Individual, and Dyad), pairwise comparisons involves multiple independent significance tests, which increases the likelihood of a Type I error (i.e., finding a significant difference between groups when in fact there is none). To combat this, the experiment-wise alpha level was adjusted using the Bonferroni adjustment (Bland & Altman, 1955). Table 5.20 shows the results of this post hoc analysis.

Table 5.20. *Pairwise Comparisons of Change Scores by Condition*

					95% CI for Difference	
(I) Group	(J) Group	Mean Difference (I-J)	Std. Error	Sig.[b]	Lower Bound	Upper Bound
IND	DYD	-.271*	0.049	0.000	-0.395	-0.147
IND	CON	.152*	0.049	0.012	0.028	0.276
DYD	CON	.423*	0.048	0.000	0.302	0.545

Note. IND=Individual, DYD=Dyads, CON=Control, * = significant at α = .05, [b] = Bonferroni adjustments made for multiple comparisons.

Using the information in Table 5.20, we can answer Research question 1: What is the difference in the amount of change in content overlap with an expert model between the Control, Individual, and Dyad conditions? Referring back to Table 5.17, we can see that the mean amount of change (i.e., the difference between the pre- and posttests) for the Individual group was .37 (or, 37% gain), for the Dyads it was .65 (or, 65% gain), and for the Control it was .22 (or, 22% gain). The first row in Table 5.20 shows the 27% (taking the absolute value) difference between the change scores of the Dyads was significantly different from the Individuals. The second row shows the 15% difference in scores for the Individuals and the Control condition was significant. The final row shows that the 42% difference between the Dyads and the Control conditions was significant. These differences denote significant changes between all groups in the amount of content overlap with the model essay. Thus the treatment conditions improved their overlap with the model more that the task repetition condition, but the dyad condition improved more than the Individual condition. The data in Table 5.17 confirm this result with inspection of the 95% CI in the final two columns. The lack of an overlap between the 95% CI ranges for each condition indicates there is a significant difference between the amounts of change across conditions. As each treatment condition shows significant gains, it will be useful to make a distinction between these values. One way to partition a distinction between these significant gains is to calculate the effect size for the change.

Cohen's d (Cohen, 1977) is a measure of the magnitude of the "effect" of the significance for a given treatment. In his seminal work, Cohen delineated effect size guidelines as d = 0.20 is a "small" effect, d = 0.50 is a "medium" effect, and d = 0.80 is a "large" effect. Essentially these benchmarks were offered as a way to lexically express the size of the potential effects that might occur if an experimental treatment was adopted into clinical practice. The calculation for Cohen's d is the difference in the means of two groups divided by their pooled standard deviations. The result can be understood as the number of standard deviation units that distinguishes the two groups, which is the effect size of the difference in their treatment conditions. Without showing all the computations here (calculations for this were done in Microsoft Excel 2010), Table 5.21 provides the necessary information we need to determine the effect size of the treatments. To calculate Cohen's d, divide the number in the Mean difference column by the number in the Pooled SD column.

Table 5.21. *Cohen's D Effect Sizes for Change in Content Overlap Scores*

Comparison	Mean difference	Pooled SD	Cohen's d
Individual v/s Dyad	0.2712	0.1328	2.04
Individual v/s Control	0.1522	0.1094	1.39
Dyad v/s Control	0.4234	0.1266	3.34

Table 5.21 shows the size of the effect between groups. The first row shows that the difference in the gain of content overlap between the learners in the Individual condition and the learners in the Dyad condition yielded a Cohen's d = 2.04. This means the effect of adding collaboration to an expert model would roughly reap a two standard deviation difference in learner scores (i.e., moving from the 50th percentile to the 90th percentile). We can see the effect of having access to the model alone in the

Cohen's D for the difference between the Individual and Control conditions, found in the second row. This shows that the effect of access to a model without collaboration, but when compared to task repetition, yields a Cohen's d = 1.39 (i.e., moving a student from the 50th percentile to the 80th percentile). The difference between task repetition and collaboration with an expert model brings an effect that would move a student 3.34 standard deviations, which is equivalent to moving from the 50th percentile to the 99th percentile. All said, the effects of the treatments brought about "large" effect sizes (i.e., Cohen's d values greater than 0.80). With gains across all conditions found to be significant, and the effect sizes of all changes in content overlap found to be "large." It is now relevant to consider was it worth putting learners into dyads to begin with. If, as the data shows, the Individuals can show a Cohen's d of 1.39 over task repetition, and Individuals can show a significant gain in their content overlap score, why should teachers put them into dyads? This interrogation of making dyads is not a trivial matter, but with the data we have, there is a way to test if dyads outperformed the other conditions.

As mentioned in the Introduction, one of the most rigorous comparisons of learning gains for collaborative conditions is the "truth-wins" comparison. To distinguish the true dyads from this theoretical pairing of individuals, I will refer to a theoretical dyad as "*nominal dyad*" and the theoretical collection of *nominal dyads* as "*group*" using italics.

To compute the best theoretical mean for a *group*, the mean of all possible pairings for each individual is calculated. Since the Control and the Independent conditions had different treatments, their *nominal dyads* must be derived separately. Using Lorge and Solomon's (1955) formula [1– (1 – observed mean)^2 = Truth-wins value] we can quickly calculate the truth-wins score for Individuals and the Control. Recall, this truth-wins value represents the theoretical best average score for a *group*. The internal portion of the equation (1 – observed mean)^2 is the chance that <u>no one</u> in a *nominal dyad* (if we used groups of three, we should cube the difference) will know the correct answer. Thus, "1" minus this is the chance that <u>at least one</u> (or both) person(s) in a dyad will know the answer. This is how the formula determines the truth-wins value for perfectly shared knowledge.

According to Lorge and Solomon's (1955) formula, given that the Individuals scored a 46% on the posttest (see posttest mean scores in Table 5.17), the truth-wins calculation for the Individual condition's best mean posttest score for the *nominal dyads* is [1– (1 – 0.46)^2 = .7127] which is equal to 71%. For the Control condition, given that they scored a .36 (or, 36%) on their posttest, their truth-wins best possible mean posttest score for *nominal dyads* is: [1– (1 – 0.36)^2 = .5904] which is equal to 59%. The actual performance (i.e., the real observed posttest average) of the Dyad condition was 75%. Since 75% is higher than the truth-wins *group* average for the Individuals (71%) and higher than the truth-wins *group* average for the Control (59%), it can be said that the dyads beat the truth-wins *nominal dyad* condition for both the Individual and Control conditions. This fact reveals a net process gain for the Dyad condition that was not seen in the best possible situation of perfectly shared knowledge for the Control and the Individual conditions. This means there was something to the collaborative condition that pushed dyads to a level that exceeded the theoretical best of the other conditions. This points to something that cannot be seen through the analysis of change in content score. Rather, this points to a need to consider the whole graph description as produced by the learners. This brings us to the analysis of the local and global data integrations.

5.9 Data-integration results

Up to this point, the results have shown an analysis of variance on the difference of the amount of content overlap between the pretest was significant across conditions, and had large effect sizes. The truth-wins comparison showed that Dyads were able to beat the theoretical best average score for the Individual and Control conditions, but this was determined on the amount of content overlap (not including redundant features) found on the posttest averages. Neither the change score nor the truth-wins calculation can tell us if learners produced essays that match the ratio of data integrations found in the expert model. In other words, it is possible that the change in posttest scores came from learners identifying only local data integrations as they are the most common in column graphs. To shed light on this, a Chi-square goodness of fit for the local and global data integrations was conducted. Conducting this analysis requires all of the local and global data integrations be accounted for. Table 5.22 displays the local integrations for the pretest and the posttest for each condition. The specific local integration can be found in the first column, and the subtotals for each group are located across the final row with the grand mean for each test in brackets of the same row. The numbers of tokens are listed under corresponding conditions, and the percentage of learners using that integration can be found in parenthesis next to each count.

Table 5.22. *Local Integrations by Test and Condition*

	Pretest			Posttest		
	IND	DYD	CON	IND	DYD	CON
Local Integrations	n=12	n=25	n=13	n=12	n=25	n=13
15 used texts online	7 (.58)	14 (.56)	10 (.77)	8 (.67)	24 (.96)	10 (.77)
80 new texts online	7 (.58)	14 (.56)	11 (.85)	8 (.67)	23 (.92)	13 (1.00)
most new texts online	10 (.83)	22 (.88)	10 (.77)	13 (.26)	26 (1.04)	10 (.77)
40 new texts on campus	5 (.42)	15 (.60)	10 (.77)	8 (.24)	24 (.96)	10 (.77)
60 used texts on campus	7 (.58)	14 (.56)	9 (.69)	8 (.27)	27 (1.08)	12 (.92)
most buy used on campus	10 (.83)	22 (.88)	9 (.69)	13 (.26)	26 (1.04)	9 (.69)
described column colors [x 2]	4 (.33)	8 (.32)	6 (.46)	15 (.40)	40 (1.60)	5 (.38)
described Y-axis	2 (.17)	6 (.24)	3 (.23)	11 (.23)	23 (.92)	2 (.15)
*100 total purchases on campus	3 (.25)	13 (.52)	6 (.46)	9 (.75)	11 (.44)	5 (.38)
*95 total purchases online	3 (.25)	10 (.40)	5 (.38)	9 (.75)	11 (.44)	6 (.46)
Subtotal [Total]	58	138	79 [275]	102	235	82 [419]

Note. ( ) = usage as a percentage of the population. As more than one usage is possible in a single essay, percentages can exceed 100. *= integrations not contained in the model.

As the focus of this study is on what learners incorporate from the expert model, the asterisked integrations were omitted from the Chi-square analysis for goodness of fit. (Notably, the results of the Chi-square test including these tokens yielded the same results as without them. For clarity, though, this is not discussed herein)[4]. Table 5.23 describes the global integrations and can be read the same as Table 5.22 for local integrations.

[4] It is also relevant to state that these two local integrations were also not scored in the percentage overlap with the model, as there was no protocol developed for their scoring according to isomorphic model.

Table 5.23. *Global Integrations by Test and Condition*

Global Integrations	Pretest			Posttest		
	IND n=12	DYD n=25	CON n=13	IND n=12	DYD n=25	CON n=13
1. total population 195	2 (.17)	6 (.24)	2 (.15)	2 (.17)	22 (.88)	2 (.15)
2. labels condition and location	4 (.33)	20 (.80)	9 (.69)	13 (1.08)	40 (1.60)	13 (1.00)
3. overall trend most buy on campus	2 (.17)	1 (.04)	2 (.15)	2 (.17)	22 (.88)	3 (.23)
4. overall trend most buy new	0 (.00)	0 (.00)	1 (.08)	0 (.00)	22 (.88)	0 (.00)
subtotal [Grand]	8	27	14 [49]	17	106	18 [141]

Note. ( ) = usage as a percentage of the population. As more than one usage is possible in a single essay, percentages can exceed 100.

To calculate the Chi-square statistic the total number of the integrations the students incorporated, by their superordinate groupings (i.e., local and global) are required. Table 5.24 shows the exact number and type of data integrations for all students by their condition assignment. A note on the population sizes is needed. Chi-Square is less dependent on the population size than it is on the number of tokens contributed, which must be at least four per cell (Fisher, 1958). That condition is met in both the expected and the observed counts, so a Chi-square is an acceptable test in under these conditions.

Table 5.24. *Local and Global Integrations by Test and Condition*

	Pretest			Posttest		
	IND n=12	DYD n=25	CON n=13	IND n=12	DYD n=25	CON n=13
Integrations	count (%)	count (%)	count (%)	count (%)	count (%)	count (%)
Local	52 (.87)	115 (.81)	68 (.83)	84 (.83)	213 (.67)	71 (.80)
Global	8 (.13)	27 (.19)	14 (.17)	17 (.17)	106 (.33)	18 (.20)
Total	60 (1.00)	142 (1.00)	82 (1.00)	101 (1.00)	319 (1.00)	89 (1.00)

Since we know that the model contained nine local data integrations and four global data integrations (See Table 5.13), we know the students were explicitly exposed to 13 data integrations that could have been extracted from the model. This means for local data integrations an essay should contain 69% references to local elements in the data (i.e., 9/13 = 0.6923), and for global integrations an essay should contain 31% references to global elements in the data (i.e., 4/13 = 0.3076). Knowing this allows us to calculate a Chi-square statistic (x^2) for goodness of fit. As there are two variables (k = 2) being test (i.e., local and global data integrations), the associated degrees of freedom (df) for this test is $df = k – 1$, which means for this data set, df = 2–1 = 1. The Chi-square critical value for a test with df = 1, α = .05, is 3.841 (Fisher, 1958). Therefore, any Chi-square statistic value above 3.841 is considered statistically different from the expected number of data integrations. The way a Chi-square statistic is calculated is by summing the square difference of the observed value and the expected value, and dividing that result by the expected value. This can be seen in the following formula:

$$x^2 = \Sigma \text{ (Observed – Expected)}^2 \text{ / Expected}$$

The model has established an expected frequency for data integrations (local = 69%, and global = 31%), so we can calculate the Chi-square expected values for each category. This is done by multiplying the total number of integrations by the expected percent. Here is an example, using the pretest for the Control condition.

According to Table 5.24 column four, the total number of data integrations for the Control condition was 82. The expected number of integrations is equal to the total number of data integrations, which is 82, times the expected percent of integrations, which is 0.6923; thus the expected local integrations is 82 * 0.6923 = 56.77 local integrations. The same procedure is used for the global integrations, so the expected global integrations is equal to 82 * 0.3076 = 25.22 global integrations. Using the Chi-square formula, we can calculate a residual for each category.

The X^2 residual for local is $(68_{\text{observed}} - 56.77_{\text{expected}})^2 / 56.77_{\text{expected}} = 5.0011$

The X^2 residual for global is $(14_{\text{observed}} - 25.22_{\text{expected}})^2 / 25.22_{\text{expected}} = 4.9938$

The x^2 statistic for the Control condition, then, is equal to the sum of the residuals, $X^2_{local} + X^2_{global} = 5.0011 + 4.9938 = 9.9949$. Using the Chi-square critical value 3.841, for a test with $df = 1$, $\alpha = .05$, means the Control condition's value is greater than the critical value and can be seen as not "a good fit" and significantly different than the expected number of data integrations. Table 5.25 displays the pretest Chi-square statistic for the remaining conditions, with the Chi-square value noted in the final column on the far right of the Table. Table 5.26 shows the Chi-Square statistics for the posttest, also seen in the far right column.

Table 5.25. *Pretest Chi-square for Goodness of Fit for Global and Local Data Integrations*

	Observed			Global		Local		
	global	local	total	exp count	residual	exp count	residual	Chi-square statistic
IND (n=12)	8	52	60	18.46	5.9237	41.54	5.9305	11.8542*
DYD (n=25)	27	115	142	43.68	6.3691	98.31	6.3799	12.7490*
CON (n=13)	14	68	82	25.22	4.9938	56.77	5.0011	9.9949*

Note. exp = expected, *= significant at α = .05 level where the chi-square critical value for 1 *df* is 3.841

In the second and third columns, Table 5.25 provides the observed frequencies (Observed) for data integration type. The expected frequencies (exp) for global integrations can be found in the fifth column, and the seventh column displays the frequencies for local integrations. These expected frequencies represent the number of data integrations that are predicted to be found in the condition's cumulative observed integrations.   The residual column shows the squared difference between observed and expected frequencies divided by the expected. In the final column the Chi-square statistic is noted with an asterisk if the integration ratio is significantly different (at α = .05) from the expected values.

The result of the Chi-square goodness-of-fit test shows that all conditions are statistically significant. Therefore, we can reject the null hypothesis and conclude that each condition incorporates global and local data integrations that deviate from the expected frequency. This is an important notation, though it is not a specified research question, because it provides evidence that the initial graph descriptions produced by learners in Stage 1 do not display the characteristic ratio of local and global data integrations as available in the expert model. This means that all groups have room to improve on their posttest writing, in terms of matching the model expected frequencies. This brings us to responding to Research Question 2: In terms of local and global data integrations, what are the differences for producing the expected number of data integrations represented in an expert model between the Control, Individual, and Dyad conditions? Table 5.26 provides data to illuminate a response to this question, as it shows the results for the posttest Chi-square tests for goodness of fit; it is read in the same fashion as Table 5.25.

Table 5.26. *Posttest Chi-square for Goodness of Fit for Global and Local Data Integrations*

	Observed			Global		Local		
	global	local	total	exp count	residual	exp count	residual	Chi-square statistic
IND (n=12)	17	84	101	31.07	6.3699	69.92	6.3790	12.7489*
DYD (n=25)	106	213	319	98.12	0.6321	220.84	0.6270	1.2591
CON (n=13)	18	71	89	27.38	3.2114	61.61	3.2175	6.4289*

Note. exp = expected, *= significant at α = .05 level where the chi-square critical value for 1 *df* is 3.841

The information in Table 5.26 shows significant differences between the expected integrations for both the Individual and Control conditions (as noted by Chi-square statistics greater than 3.841). This means the amount of integrations produced by the Control and Individual conditions did not match the expected ratio for data integrations. This indicates that both the Individual and the Control conditions displayed an over or under use of data integration types. In contrast, the Dyad condition had a non-significant Chi-square value ($X^2 < 3.841$, n.s.), and did not deviate significantly from the expected ratio of integrations. Thus, the ratio of data integrations for the Dyads matched the expected frequencies, while the Control and the Individual conditions significantly deviated from the ratio of integrations presented in the model.

Looking at the data in this way, we can realize two things. One, this identifies why it was important in the 37-point coding scheme to not allow points for redundancy of elements. If redundant elements were included in the 37-point scheme, it would have been difficult to see what learners gain from looking at an expert model, as scores might have been artificially inflated. Two, recall that in Table 5.24 the local and global elements were broken out into percentages. Table 5.24 shows that there was little change in the ratio of integrations from the pretest to the posttests for the Individual and Control conditions. The Individuals had 87% local integrations on the pretest and 84% on the posttest. The Control had 83% local integrations on the pretest and 80% on the posttest. The dyads, however, had 81% local integrations on the pretest but reduced that percentage to 67% on the posttest. These ratios demonstrate an increased level (81% – 67% = 14% increase) of global integrations used by the Dyads, whereas the other conditions each increased their global integrations by a mere 3%. From the perspective that global data integrations are complex, the descriptions produced by the Dyads could be defined as more complex.

To answer research question 3— Will learners in the Control, Individual, and Dyad conditions identify the complex global integration found in the interaction of

Graph 1—reconsideration of Table 5.23 is needed. There were four global integrations possible, and one was considered remarkably complex. This global integration was considered complex because it required learners to see through an interaction between bookstore location and type of book to identify that a majority of students purchased a new textbook. To find this trend, learners had to recognize that they need to sum the lowest frequency of one group with the highest frequency of another group to determine the overall trend. This requires the learners ignore the local integrative nature of column graphs (Carswell et al., 1992; Shah, 2002) to grasp the underlying global trend for most books purchased. The results from Table 5.23 have been abbreviated in Table 5.27 below.

Table 5.27. *Results of Target Transfer Item for Complex Global Data Integration*

	Pretest			Posttest		
Global Integrations	IND n=12	DYD n=25	CON n=13	IND n=12	DYD n=25	CON n=13
overall trend most buy new	0 (.00)	0 (.00)	1 (.08)	0 (.00)	22 (.88)	0 (.00)

Note. ( ) = percentages

Table 5.27 shows that in the pretest, no students from the Individual condition or the Dyad condition identified this trend on their first innovation of a response to Graph 1. It also shows that only one learner in the Control condition was able to see this trend and produced a statement addressing this global integration. In the posttest, no learners in the Individual condition were able to see this despite having access to the expert model. The last column also notes that the single learner in the Control condition, who managed to see and produce a statement about this global trend in the pretest, did not produce it on their posttest revisions. In stark contrast to these results, however, 22 of the learners in the Dyad condition identified this global trend in their posttest. As the other conditions showed no variation in their data, and there was no evidence of learners in either the Individual or the Control conditions producing this feature, there is no way to statistically test this change. At a glance, though, the results indicate that Dyads were the only condition to identify this complex global integration on the posttest.

CHAPTER 6. DISCUSSION

This dissertation began with the question: How can feedback become a productive resource for students. To challenge the recent literature on using models as feedback to learners, an experiment with college L2 English students tested the effects of working with a model under individual or collaborative conditions. From the start, it was hypothesized that the treatment conditions with the model would surpass the control condition (who represented task repetition without peer interaction or an expert model). Furthermore, it was hypothesized that working in dyads with a model would afford productive collaborations for acquiring both global and local data integrations, while maintaining overlap with the model content, whereas learning with the model individually would not be as informative for learners. The results of this study provided support for these hypotheses.

The first research question examined if learners working alone with a model or working in dyads with a model would reproduce essential content-level elements that could be traced back to specific language from the expert model. The results showed there were statistically significant differences between the learning gains of all conditions (and the Cohen's d effect sizes were large for these differences). This means the model and the interaction with the model all brought about a significant change to posttest writing. This provides evidence working individually or in dyads might be a useful way of encouraging learner uptake of content overlap with the expert model. Two facets about this study are important to note when understanding why learners in both treatment conditions increased their content overlap with the model. One is the learners in this study did not view their original essays while looking at the isomorphic expert model. The other is the learners completed a comprehension worksheet that provided evidence (with scores averaging 97% for individuals and 98% for dyads) that learners could understand the meaning and purpose of each sentence in the model. In light of these facets, the findings might suggest that individual learners who have a clear understanding of an expert model may be able to transfer content language from an isomorphic model essay to their revisions without comparing the model to their original essays.

Language researchers have noted that it is important, or at least useful, for learners to compare their original writing to the model in order to facilitate language uptake. The findings of this study challenge the need for this in the contexts of a demonstrable task and comprehension of the model. While models can be seen as a resource for the learners, it is important to consider how this study employed a demonstrability element to the model and the tasks. The model used in this study was an isomorph that described a graph, which had concrete numbers and was based on a topic the learners had experiences with (i.e., living conditions). This may have let learners focus on the language in the model, as they could understand the contexts of the prompt. This is not always the case with picture prompts, as learners are sometimes left to make assumptions about the contents of a picture prompt. The comprehension worksheet also provided evidence that learners could comprehend the essential information, used to describe the graph, provided in the model information. This comprehensibility combined with knowledge of the subject may be part of the reason both the individuals and the dyads were able to transfer content language from the model to their revisions. However, the results of the second research question showed that dyads were also able to incorporate the expected number of global and local data integrations, whereas the individuals and the control conditions were not able to do so. One reason for this might be the dyads were working with a topic they were familiar with, had experience with, and they each had an expert solution that may have acted as a safe model to discuss.

Collaborative writing researchers have chance that learners will shy away from interaction in collaborative writing tasks if they feel their response might be deemed wrong or not important (Storch, 2013), this study used an isomorphic model (an expert model that can be seen as a "correct solution") to foster a discussion of the language in the model. The fact that the model could be seen as a solution might have pushed learners to consider the content of the model as important. This may have pushed dyads to engage in deeper discussion about the content of the model, which may have caused learners to discuss in general terms their opinions about both the model and the prompt. This points to one common benefit noted by collaborative learning specialists, discussing solutions with a partner encourages a deeper understanding of the material and helps learners focus on the abstractions of the material that make their knowledge more applicable to novel situations. While this may seem like a finding that adds to current knowledge, what is important about this finding is the interaction phase came after receiving the expert model as a learning resource. This is different from common modes of collaboration with innovation task (e.g., Schwartz et al. 2005), which puts the interaction phase first. Naturally, the more language the learners notice in the model, the more likely they will recall that language and content elements at a later time (Schmidt, 1990; Schmidt and Frota, 1986), but in the case of the deep structures (i.e., the global and local data integrations) the noticing of global structures seemed to be evident only in the dyad condition.

Another perspective on this is that learners in the Individual condition had only the model and prompt to compare to their individual experiences. The learners in the dyad condition, however, were able to provide their own contrasting cases to each other. Identifying how one learner wrote something different on their pretest, might have caused the other learners to notice language in their peer's dialogue that overlapped with their writing experiences. This may have helped the learners engrain the content language of the model in terms of their recollection of their pretest essay, but also allowed the learners to discuss the topic of living on and off campus.

During the Stage 2 treatment phase, the Dyad classes were buzzing with discussions about where one student lived and whether they lived with a roommate. Laughter filled the classroom and their excitement for the discussion portion of the task was palpable. Commonly there were vocalized expressions of shock in learning that one member of a dyad lived in the same dorm (and sometimes on the same floor) as the other. These lively interactions pushed students to explain their living conditions to other students (and often relating their friend's living conditions as well). Doing so may have driven members to broaden their approach to interpreting the data in the model. This would make their discussion of the model more conceptual and less restricted to the exact data presented to them.

In the Dyad condition, to maintain their conversations and idea exchanges, learners needed to consider the global data integrations to couch the details of the local integrations in the model. Put metaphorically, it would have been difficult for the dyads to sustain a 15-minute discussion by only discussing the trees (i.e., the local integrations) without discussing the forest (i.e., the global integrations). This finding supports the research on collaborative learning that shows learners who interact with partners to discuss, explain, and negotiate a meaning on a shared topic develop a deeper understanding of the materials than those working alone (e.g., Barron, 2000). Another point to consider is the comparison of the Control to the Individuals.

The learners in the Individual condition did not show a balance of global and local traits, nor did the Control. If the model was the source of obtaining the balance in posttest writing for the Dyads, the Individual group should have also acquired the global integrations. This means, working individually with the model cannot be shown as affording any more awareness of the global underlying structures as task repetition can. This does not mean that the models were useless for the Individuals. It may mean, as other researchers have suggested (e.g., Hanaoka, 2007), models are useful for identifying surface features. This notation must be taken with caution because, different from the previous studies on model feedback, no learners in any condition had access to their original texts. This may bring to light the potential of having learners write a response to a difficult task (as noted by the abysmal 37-point overlap scores ranging from .09 to .14 on the pretest, see Table 5.17), and then considering a solution without considering their mistakes. While this runs counter to the idea that learner feedback should come as soon after the mistake has been made so as not to miss the "window of opportunity for making a cognitive comparison" (Doughty, 2001, p. 254), it also seem to be fruitful to delay the feedback. The findings of this study suggests that delaying feedback for learners working in groups might be beneficial as they will have a chance to test their language hypothesis by defending their hypothesis with the other learners. This process of negotiation helps the learner focus on making their reason for their language decision clear, if not succinct (similar to the Schwartz, 1995 study).

From a practical perspective, if research on feedback is to mirror what happens practically, learners cannot know when they will have a model to work with. Thus, it makes logical sense to have feedback come when the learners are prepared to learn from it. From this dissertation's data, comparisons on the timing of the feedback cannot be made. However, from a theoretical point, it seems that dyads that were pushed to consider a model from multiple perspectives were better prepared to transfer what they gleaned from the model when compared to their counterparts who sat in silence, reading, and pondering the intricacies of the graph and its description. This means that the collaborative interactions played more of a determining factor for the use of certain feedback elements despite its delayed timing.

Returning to the results of the 37-point content analysis, this dissertation found strong evidence for learning from models in both Dyads and Individual conditions (i.e., they each beat the Control condition). While this may provide results supporting collaboration with models in the L2 writing classroom, it is important to note that it is not claimed that models *naturally* support productive collaboration. The supporting task that required learners to engage with one another, and to not only share their ideas but to listen to their partner's ideas, was no doubt crucial to the results found in the analysis.

Recall that when learners were asked to identify and describe the important elements of the model on the worksheet in Stage 2, telling why each sentence was important, both treatment conditions scored between .92 and 1.0 on all eight sentences from the model, and on their responses to the two discussion questions (see Table 5.16). Since the ability to identify and comprehend trends in data can be interpreted as understanding a graph (Carswell et al., 1992), the worksheet served a useful purpose in showing the learners understood the contents of the model. Since the model contained the solutions learners needed to excel in the posttest stage of the study, the analysis of the comprehension worksheet might have shown that all learners were prepared to maximally learn from the model. However, by having learners write a description of a graph that contained a complex facet (i.e., the interaction effect), the results describe a different picture of how collaborative tasks with models might prepare a learners for transferring what they see into what they know.

Clearly, the comprehension worksheet demonstrated a weakness in asking learners to merely identify that they understand what is in a model. Just as Bransford and Schwartz (1999) noted with the significant difference in performance scores on the true and false test for psychology students, SPS assessments of comprehension knowledge are not useful for showing when students are ready to learn. The results of the posttest writing showed that learners who discussed the model were more prepared to apply the information in the model when compared to learners who simply described their understanding of the model. It is likely, then, that the interaction provided a rich environment for giving and receiving explanations, which have been connected to learners achieving an understanding of the deeper constructs behind the materials (Chi et al. 1994; Webb, 1982). This is an important point because while this study provides evidence for both collaboration with models, and viewing models individually, it also provides support for and extends previous L2 writing research that shows models can provide learners with information about content. In this case, it provides evidence for the incorporation of data integrative points, through structured tasks while examining an expert model.

As Hatano and Inagaki (1986) suggested, emblematic of successful learning is the ability to apply ones knowledge to a novel situation (i.e., adaptive expertise). This perspective of adaptive expertise might be a positive one for L2 writing researchers to adopt. Flexible approaches to conceptualizing learning gains may lead to new classroom procedures, including adaptations of approaches to assessing learning gains (e.g., truth-wins criterion) and incorporating theories form educational contexts outside of language teaching (e.g., PFL). In language learning research, to my knowledge, this is the only dissertation (if not experiment) that has employed stringent truth–wins criteria to evaluate the outcome gains of collaboration (for one dissertation in education that accomplishes this, see Sears, 2006).

Future work could investigate if dyads with a model are differently prepared than dyads who work with a graph but no model. Varying the content and topic of graphs might also be of interest to researchers concerned with background knowledge. For example, individuals and dyads might be compared on data they collect as compared to individuals and dyads given simulated data on a fictitious topic to summarize. With an eye towards grammar, a model with restricted interpretations, such as a pie chart, might be one way of using models to help learners focus on sentence level concerns. Technology might also be explored for its role in affecting output in a study that investigates dyads conversing in a chat room about the data versus face-to-face discussions. This investigation might having interesting results since learners in the online condition will have to type their responses, but in reduced formats common to chatrooms.

In conclusion, this dissertation adds to the literature of feedback studies in L2 writing by demonstrating the potential relevance of collaborative learning in light of content development of underlying structures. In particular, the results suggested one way by which teachers might tap the benefits of collaboration in the language classroom through structure tasks that encourage dialogue on a common topic. The findings also suggested that collaboration with a model (but without the learners' original essay) could promote productive interactions that lead to a more robust understanding of the underlying concepts and structures in the expert model.

REFERENCES

REFERENCES

Adams, R. (2003). L2 output, reformulation and noticing: implications for IL development. *Language Teaching Research*, *7*(3), 347–376. doi:10.1191/1362168803lr127oa

Allwright, R. L., Woodley, M. P., & Allwright, J. M. (1988). Investigating reformulation as a practical strategy for the teaching of academic writing. *Applied Linguistics*, *9*(3), 236-256.

Andersson, J., & Rönnberg, J. (1995). Recall suffers from collaboration: Joint recall effects of friendship and task complexity. *Applied Cognitive Psychology, 9*, 199-211.

Antil, L., Jenkins, J., Wayne, S., & Vadasy, P. (1998). Cooperative learning: Prevalence, conceptualizations, and the relation between research and practice. *American Educational Research Journal, 35,* 419-454.

Ashman, A. F., & Gillies, R. M. (1997). Children's cooperative behavior and interaction in trained and untrained work groups in regular classrooms. *Journal of School Psychology, 35*(3), 261–279.

Barron, B. (2000). Problem solving in video-based microworlds: Collaborative and individual outcomes of high-achieving sixth-grade students. *Journal of Educational Psychology, 92*, 391-398.

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

Bazeley, P., & Jackson, K. (Eds.). (2013). *Qualitative data analysis with NVivo*. Los Angeles, CA: Sage.

Biber, D., Nekrasova, T., & Horn, B. (2011). The Effectiveness of Feedback for L1 English and L2 Writing Development: A Meta Analysis. *ETS Research Report Series*, *2011*(1), i–99.

Bitchener, J. (2008). Evidence in support of written corrective feedback. *Journal of Second Language Writing, 17,* 102–118.

Bitchener, J. (2009). Measuring the effectiveness of written corrective feedback: A response to "Overgeneralization from a narrow focus: A response to Bitchener (2008)". *Journal of Second Language Writing, 18,* 276–279.

Bland, M. J. & Altman, D. G. (1955). Multiple significance tests: The Bonferroni method, *British Medical Journal*, *310*(6973), 170-170.

Bransford, J.D., & Schwartz, D.L. (1999). Rethinking transfer: A simple proposal with multiple implications. In A. Iran-Nejad and P. D. Pearson (Eds.), *Review of Research in Education, 24,* 61-100. Washington, D.C.: American Educational Research Association.

Campbell, C. (1990). Writing with others' words: Using background reading text in academic compositions. In B. Kroll (Ed.), *Second language writing* (pp. 211–230). Cambridge, UK: Cambridge University Press.

Carswell, C. M., & Wickens, C. D. (1987) Information integration and the object display: An interaction of task demands and display superiority, *Ergonomics*, *30*(3), 511-527.

Carswell, C. M., Emery, C., & Lonon, A. M. (1992). Stimulus complexity and information integration in the spontaneous interpretation of line graphs. *Applied Cognitive Psychology, 7*(4), 341–357.

Chi, M. T. H., de Leeuw, N., Chiu, M-H, & Lavancher, C. (1994). Eliciting self-explanations improves understanding, *Cognitive Science*, *18*, *439-477.*

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive Science, 5*(2), 121–152.

Cleveland, W. S., & McGill, R. (1985). Graphical Perception and Graphical Methods for Analyzing Scientific Data, *Science*, *229*(4716), 828–833.

Cohen, A. (1989) Reformulation: A technique for providing advanced feedback in writing, *Guidelines, 11,* 1–9.

Cohen, E. G. (1994). Restructuring the classroom: Conditions for productive small groups. *Review of Educational Research, 64*, 1-35.

Cohen, E., Lotan, R., Scarloss, B., Schultz, S., & Abram, P. (2002). Can groups learn? Teachers College Record, 104, 1045–1068.

Cohen, J. (1977). *Statistical power analysis for the behavioral sciences*. New York: Academic Press.

Coleman, E. (1998). Using explanatory knowledge during collaborative problem solving in science. *The Journal of the Learning Sciences, 7*, 387-427.

Dennis, A., & Valacich, J. (1993). Computer brainstorms: More heads are better than one. *Journal of Applied Psychology, 78*, 531-537.

Detterman, D. K. (1993). The case for the prosecution: transfer as an epiphenomenon. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: intelligence, cognition, and instruction* (pp. 1–24). Norwood: Ablex.

Detterman, D. K., & Sternberg, R. J. (Eds.). (1993). *Transfer on trial: Intelligence, cognition, and instruction*. Norwood, NJ: Ablex.

Dillenbourg P. (1999) What do you mean by collaborative leraning?. In P. illenbourg (Ed) *Collaborative-learning: Cognitive and Computational Approaches* . (pp.1-19). Oxford, UK: Elsevier.

Doughty, C. (2001). Cognitive underpinnings of focus on form. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 206–257). Cambridge: Cambridge University Press.

Ede, L. S., & Lunsford, A. A. (1990). *Singular Texts/plural Authors: Perspectives on Collaborative Writing*. Carbondale & Edwardsville, IN: Southern Illinois University Press.

Eschholz, P. A. (1980). The prose models approach: Using products in the process. In T. R. Donovan & B. W. McClelland (Eds.), *Eight approaches to teaching composition* (pp. 21–36). Urbana, IL: National Council of Teachers of English.

Fernández Dobao, A., & Blum, A. (2013). Collaborative writing in pairs and small groups: Learners' attitudes and perceptions. *System*, *41*(2), 365–378. doi:10.1016/j.system.2013.02.002

Ferris, D. (2004). The ''Grammar Correction'' Debate in L2 Writing: Where are we, and where do we go from here? (and what do we do in the meantime ...?), *Journal of Second Language Writing, 13*, 49–62.

Ferris, D. (2006) Does error feedback help student writers? New evidence on short- and long-term effects of written error correction, In K. Hyland and F. Hyland (Eds.). *Feedback in Second Language Writing: Contexts and Issues*. Cambridge: Cambridge University Press.

Ferris, D. (2011). *Treatment of error in second language student writing*. Ann Arbor, MI: University of Michigan Press.

Ferris, D. R. & Roberts, B. (2001) Error feedback in L2 writing classes: How explicit does it need to be?, *Journal of Second Language Writing, 10,* 161–184.

Fisher, R. A. (1958). *Statistical Methods for Research Workers*. New York: Hafner.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1-38.

Gillies, R. (2004). The effects of cooperative learning on junior high school students during small group learning. Learning and Instruction, 14(2), 197–213.

Glaser, B. G. (1978). *Theoretical sensitivity: Advances in the methodology of grounded theory*. Mill Valley, CA: Sociology Press.

Guirao, J. C., de Larios, J. R., & Coyle, Y. (2015). The use of models as a written feedback technique with young EFL learners. *System*, *52*, 63–77.

Hanaoka, O. (2006a). Exploring the role of models in promoting noticing in L2 writing. *JACET Bulletin*, *42*, 1–13.

Hanaoka, O. (2006b). Noticing from models and reformulations: A case study of two Japanese EFL learners. *Sophia Linguistica*, *54*, 167–192

Hanaoka, O. (2007). Output, noticing, and learning: An investigation into the role of spontaneous attention to form in a four-stage writing task. *Language Teaching Research*, *11*(4), 459–479.

Hanaoka, O., & Izumi, S. (2012). Noticing and uptake: Addressing pre-articulated covert problems in L2 writing. *Journal of Second Language Writing*, *21*(4), 332–347. doi:10.1016/j.jslw.2012.09.008

Hatano, G., & Inagaki, K. (1986). Two courses of expertise. In H. Stevenson, H. Azuma, & K. Hakuta (Eds.), *Child development and education in Japan* (pp. 262-272). New York, NY: Freeman.

Hedgcock, J. & Lefkowitz, N. (1994). Feedback on feedback: Assessing learner receptivity to teacher response in L2 composing, *Journal of Second Language Writing, 3*, 141–63.

Horowitz, D. M. (1986). What professors actually require: Academic tasks for the ESL classroom. *TESOL Quarterly, 20*(3), 445-462.

Hyland, F. (1998). The impact of teacher written feedback on individual writers, *Journal of Second Language Writing*, *7*(3), p. 255-286.

Hyland, K. & Hyland, F. (2006). Feedback on second language students' writing, *Language Teaching, 39,* 83–101.

Izumi, S., & Bigelow, M. (2000). Does output promote noticing and second language acquisition? *TESOL Quarterly*, *34*, 239–278.

Izumi, S., Bigelow, M., Fujiwara, M., & Fearnow, S. (1999). Testing the output hypothesis: Effects of output on noticing and second language acquisition. *Studies in Second Language Acquisition*, *21*, 421–452.

James, M. A. (2006a). Transfer of learning from a university content-based EAP course. *TESOL Quarterly, 40*, 783–806.

James, M. A. (2006b). Teaching for transfer in ELT. *The ELT Journal, 60*, 151–159.

James, M. A. (2008). The influence of perceptions of task similarity/difference on learning transfer in second language writing. *Written Communication, 25*, 76–103.

James, M. A. (2009). ''Far'' transfer of learning outcomes from an ESL writing course: Can the gap be bridged? *Journal of Second Language Writing, 18*(2), 69–84.

James, M. A. (2010). An investigation of learning transfer in English-for-general-academic-purposes writing instruction. *Journal of Second Language Writing*, *19*(4), 183-206.

Janssen, J., Erkens, G., Kirschner, P. A., & Kanselaar, G. (2010). Effects of representational guidance during computer-supported collaborative learning. *Instructional Science*, *38*, 59–88.

Johnson, D. W., & Johnson, R. T. (1989). Cooperation and competition: Theory and research. Edina: Interaction Book Company.

Johnson, D. W., & Johnson, R. T. (1999). Making cooperative learning work. *Theory into practice*, *38*(2), 67-73.

Johnson, D. W., & Johnson, R. T. (2009). An educational psychology success story: Social interdependence theory and cooperative learning. *Educational Researcher, 38*, 365-379.

Kepner, C. G. (1991). An experiment in the relationship of types of written feedback to the development of second language writing skills. *Modern Language Journal*, *75*, 305–313.

Kerr, N. L., & Bruun, S. E. (1983). Dispensability of member effort and group motivation losses: Free-rider effects. Journal of Personality and Social Psychology, 44(1), 78–94.

King, A. (1999). Discourse Patterns for Mediating Peer Learning. In A. O'Donnell and A. King (Eds.), *Cognitive perspectives on peer learning* (pp. 87-115). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Kosslyn, S. M. (1994). *Elements of Graph Design*. New York, NY: Freeman.

Koyama, D. (2014). [Review of the book *Collaborative Writing in L2 Classrooms*, by Neomy Storch]. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *14*(1), 39–43.

Koyama, D., & Sun, A. (2012, September). *A two-year longitudinal study of complexity, accuracy, and fluency: Problematizing complexity in EFL university writing exams*, Paper presented at The Symposium on Second Language Writing, West Lafayette, IN.

Kruger, A. C. (1992). The effect of peer and adult-child transactive discussions on moral reasoning. *Merrill-Palmer Quarterly*, *38*(2), 191-211.

Lamm, H., & Trommsdorf, G. (1973). Group versus individual performance on tasks requiring ideational proficiency (brainstorming): A review. *European Journal of Social Psychology, 3*, 361-388.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159-174.

Lapkin, S., Swain, M., & Smith, M. (2002). Reformulation and the learning of French pro-nominal verbs in a Canadian French immersion context. *The Modern Language Journal*, *86*, 485–507.

Larsen Freeman, D. (2013). Transfer of learning transformed. *Language Learning*, *63*(s1), 107–129.

Latané, B., Williams, K., & Harkins, S. (1979). Many hands make light the work: The causes and consequences of social loafing. *Journal of Personality and Social Psychology*, *37*, 822-832.

Laughlin, P. R., & Ellis, A. L. (1986). Demonstrability and social combination processes on mathematical intellective tasks. Journal of Experimental Social Psychology, 22, 177–189.

Laughlin, P. R., Carey, H. R., & Kerr, N. L. (2008). Group-to-individual problem-solving transfer. Group Processes and Intergroup Relations, 11(3), 319–330.

Laughlin, P. R., Hatch, E. C., Silver, J. S., & Boh, L. (2006). Groups perform better than the best individual on letters-to-numbers problems: Effects of group size. Journal of Personality and Social Psychology, 90(4), 644–651.

Laughlin, P. R., Zander, M. L., Knievel, E. M., & Tan, T. K. (2003). Groups perform better than the best individuals on letters-to-numbers problems: Informative equations and effective strategies. *Journal of Personality and Social Psychology*, *85*(4), 684–694.

Leki, I. (1998). *Academic writing: Exploring processes and strategies*, Second edition. New York, NY: Cambridge University Press.

Leki, I., & Carson, J. (1997). Completely different worlds: EAP and the writing experiences of ESL students in university courses*. TESOL Quarterly, 31,* 39–69. doi: 10.2307/3587974

Long, M. (1996). The role of the linguistic environment in second language acquisition. In Ritchie,W.C. and Bhatia,T.K. (Eds.), *Handbook of language acquisition. Vol. 2: Second language acquisition*. New York: Academic Press, 413–68.

Lorge, I., & Solomon, H. (1955). Two models of group behavior in the solution of eureka type problems. *Psychometrika*, 20(2), 139–148.

Martinez Esteban, N., & Roca de Larios, J. (2010). The use of models as a form of written feedback to secondary school pupils of English. International Journal of English Studies, 10, 143–170.

McNemar, Q. (1955). *Psychological Statistics.* New York, NY: Wiley.

Norris, J. M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics*, *30*(4), 555-578.

O‘Donnell, A. M & Dansereau, D. F. (1992). Scripted cooperation in student dyads: A method for analyzing and enhancing academic learning and performance. In R. Hertz-Lazarowitz & N. Miller (Eds.), *Interaction in cooperative groups: The theoretical anatomy of group learning* (pp.120-141). NewYork: Cambridge University Press

O’Donnell, A. M. (1996). Effects of explicit incentives on scripted and unscripted cooperation. Journal of Educational Psychology, 88, 74–86.

O’Donnell, A. M. (2006). The role of peers and group learning. In P. A. Alexander & P. H. Winne (Eds.), Handbook of educational psychology (2nd ed., pp. 781–802). Mahwah: Lawrence Erlbaum Associates.

O’Donnell, A. M., Dansereau, D. F., Hall, R. H., & Rocklin, T. R. (1987). Cognitive, social/affective, and metacognitive outcomes of scripted cooperative learning. *The Journal of Educational Psychology, 79*, 431-437.

Olivera, F., & Straus, S. G.(2004). Group-to-individual transfer of learning: Cognitive and social factors. *Small Group Research*, *35*, 440–465.

Pai, H. H., Sears, D. A., & Maeda, Y. (2014). Effects of small-group learning on transfer: A meta-analysis. *Educational Psychology Review*, *27*(1), 79-102.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal Teaching of Comprehension-Fostering and Comprehension-Monitoring Activities, *Cognition and Instruction, Vol. 1*, 117-175.

Phelps, E., & Damon, W. (1989). Problem solving with equals: Peer collaboration as a context for learning mathematics and spatial concepts. *Journal of Educational Psychology, 81*, 639-646.

Pica, T. (1994). Research on negotiation: What does it reveal about second-language learning conditions, processes, and outcomes?, *Language Learning, 44*, 493–527.

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language research and instruction. In G. Crookes & S. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 9–34). Clevedon, England: Multilingual Matters.

Qi, D.S., & Lapkin, S. (2001) Exploring the role of noticing in a three-stage second language writing task. *Journal of Second Language Writing*, *10*, 277–303.

Raimes, A. (1985). What unskilled ESL students do as they write: A classroom study of composing. *TESOL Quarterly*, *19*, 229-258.

Rau,W., & Heyl, B. S. (1990). Humanizing the college classroom: collaborative learning and social organization among students. *Teaching Sociology, 18*(2), 141–155.

Robinson, P. (1995). Attention, memory, and the “noticing” hypothesis. *Language learning*,*45*(2), 283-331.

Rutherford, W. (1987). *Second Language Grammar: Learning and Teaching.* London: Longman.

Sachs, R., & Polio, C. (2007). Learners’ uses of two types of written feedback on a L2 writing revision task. Studies in Second Language Acquisition, 29, 67–100.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11,* 206–226.

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and awareness in foreign language learning* (Technical Report No. 9, pp. 1–63). Honolulu: University of Hawaiʻi.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 3-32). Cambridge: Cambridge University Press.

Schmidt, R., & Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237–326). Rowley, MA: Newbury House.

Schwartz, D. L. (1995). The emergence of abstract representations in dyad problem solving. *Journal of the Learning Sciences*, *4*(3), 321-354.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition & Instruction*, *16*(4) 475–522.

Schwartz, D. L., & Martin, T. (2004). Inventing to prepare for future learning: The hidden efficiency of encouraging original student production in statistics instruction. *Cognition and Instruction*, *22*(2), 129-184.

Schwartz, D. L., Chase, C. C., Oppezzo, M. A., & Chin, D. B. (2011). Practicing versus inventing with contrasting cases: The effects of telling first on learning and transfer. *Journal of Educational Psychology*, *103*(4), 759.

Schwartz, D., Bransford, J. & Sears, D. (2005). Efficiency and innovation in transfer. In J. Mestre (Ed.), *Transfer of learning from a modern multidisciplinary perspective.* CT: Information Age Publishing.

Schwartz, D.L. & Bransford, J.D. (1998). A time for telling. *Cognition and Instruction, 16*, 475-522.

Sears, D. A. (2006). Effects of innovation versus efficiency tasks on collaboration and learning (Doctoral dissertation, Stanford University, 2006). *Dissertation Abstracts International, 67*, 1652A.

Sears, D. A., & Reagin, J. M. (2013). Individual versus collaborative problem solving: divergent outcomes depending on task complexity. *Instructional Science*, *41*(6), 1153-1172.

Semke, H. D. (1984). Effects of the red pen. *Foreign Language Annals*, *17*, 195–202.

Shah, P. (2002). Graph comprehension: The role of format, content and individual differences. In *Diagrammatic representation and reasoning* (pp. 173-185). Springer London.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*(1), 43.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction, *Educational Psychology Review*, *14*(1), 47–69.

Sheppard, K. (1992). Two feedback types: Do they make a difference? RELC Journal, 23, 103–110.

Slavin, R. (1996). Research for the future: Research on cooperative learning and achievement: What we know, what we need to know. *Contemporary Educational Psychology, 21,* 43-69.

Steiner, I. D. (1966). Models for inferring relationships between group size and potential group productivity. *Behavioral Science*, *11*(4), 273 – 283.

Steiner, I. D. (1972). Group process and productivity. New York: Academic press.

Storch, N. (1998). A classroom-based study: insights from a collaborative text reconstruction task. *ELT Journal*, *52*(4), 291–300. doi:10.1093/elt/52.4.291

Storch, N. (1999). Are two heads better than one? Pair work and grammatical accuracy. *System*, *27*(3), 363–374. doi:10.1016/S0346-251X(99)00031-7

Storch, N. (2001a). Comparing ESL learners' attention to grammar on three different classroom tasks. *RELC Journal*, *32*(2), 104–124. doi:10.1177/003368820103200207

Storch, N. (2001b). How collaborative is pair work? ESL tertiary students composing in pairs. *Language Teaching Research*, *5*(1), 29–53. doi:10.1177/136216880100500103

Storch, N. (2002a). Patterns of Interaction in ESL Pair Work. *Language Learning*, *52*(1), 119–158. doi:10.1111/1467-9922.00179

Storch, N. (2002b). Relationships formed in dyadic interaction and opportunity for learning. *International Journal of Educational Research*, *37*(3–4), 305–322. doi:10.1016/S0883-0355(03)00007-7

Storch, N. (2005). Collaborative writing: Product, process, and students' reflections. *Journal of Second Language Writing*, *14*(3), 153–173. doi:10.1016/j.jslw.2005.05.002

Storch, N. (2013). *Collaborative Writing in L2 Classrooms.* Bristol, UK: Multilingual Matters.

Storch, N. & Wigglesworth, G. (2010). Learners' processing, uptake and retention of corrective feedback on writing. Case studies, S*tudies in Second Language Acquisition 32,* 303–334.

Storch, N., & Wigglesworth, G. (2010). Learners' processing, uptake, and retention of corrective feedback on writing. *Studies in Second Language Acquisition*, *32*(02), 303–334. doi:10.1017/S0272263109990532

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.),*Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *The Canadian Modern Language Review, 50,* 158–164.

Swain, M. (1994). Three functions of output in second language learning. In G. Cook and B. Seidlhofer (Eds.). *Principle and Practice in Applied Linguistics.* Oxford: Oxford University Press.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principles and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 125–144). Oxford: Oxford University Press.

Swain, M. (1998). Focus on form through conscious reflection. In C. Doughty & J. Williams (Eds.), *Focus on form in classroom second language acquisition* (pp. 64–81). New York: Cambridge University Press.

Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics, 16,* 371–391.

Tardy, C. (2006). Appropriation, ownership, and agency: Negotiating teacher feedback in academic settings. *Feedback in second language writing*, 60-78.

Teasley, S., & Roschelle, J. (1993). Constructing a joint problem space: The computer as a tool for sharing knowledge. In S. P. Lajoie & S. J. Derry (Eds.),*Computers as cognitive tools* (pp. 229–258). Hillsdale, NJ: Lawrence Erlbaum Associates.

Tocalli-Beller, A., & Swain, M. (2005). Reformulation: The cognitive conflict and L2 learning it generates. *International Journal of Applied Linguistics*, *15*(1), 5–28.

Truscott, J. (1996). The case against grammar correction in L2 writing classes, *Language Learning, 46,* 327-369.

Truscott, J. (1999). The case for 'The case against grammar correction in L2 writing classes': A response to Ferris, *Journal of Second Language Writing, 8,* 111-122.

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

Watanabe, Y., & Swain, M. (2007). Effects of proficiency differences and patterns of pair interaction on second language learning: collaborative dialogue between adult ESL learners. *Language Teaching Research*, *11*(2), 121–142. doi:10.1177/136216880607074599

Watson, C. (1982). The use and abuse of models in the ESL writing class. *TESOL Quarterly, 16,* 5–14.

Webb, N. M. (1989). Peer interaction and learning in small groups. *International journal of Educational research*, *13*, 21-39.

Webb, N. M. (1992). Testing a theoretical model of student interaction and learning in small groups. In R. Hertz-Lazarowitz & N. Miller (Eds.) *Interaction in cooperative groups: The theoretical anatomy of group learning*, (pp.102–119). New York: Cambridge University Press.

Webb, N. M. (2009). The teacher's role in promoting collaborative dialogue in the classroom. *British Journal of Educational Psychology, 79*, 1–28.

Webb, N. M., & Farivar, S. (1994). Promoting helping behavior in cooperative small groups in middle school mathematics. *American Educational Research Journal*, *31*(2), 369–395.

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

Webb, N. M., Troper, J. D., & Fall, R. (1995). Constructive activity and learning in collaborative small groups. *Journal of Educational Psychology, 87,* 406-423.

Weinberger, A., Ertl, B., Fischer, F., & Mandl, H. (2005). Epistemic and social scripts in computer supported collaborative learning. *Instructional Science*, *33*, 1–30.

Williams, J. (2001). Learner-generated attention to form. *Language Learning, 49*, 583–625.

Williams, J. (2012). The potential role(s) of writing in second language development. *Journal of Second Language Writing, 21*, 321–331.

Williams, K. D. (2001). *Ostracism: The power of silence.* New York, NY: Guilford Press.

Yang, L., & Zhang, L. (2010). Exploring the role of reformulations and a model text in EFL students' writing performance. *Language Teaching Research, 14*(4), 464–484.

Zamel, V. (1983). The composing processes of advanced ESL students: Six case studies. *TESOL quarterly*, *17*, 165-188.

Zamel, V. (1985). Responding to student writing, *TESOL Quarterly, 19*(1), 79–101.

APPENDICES

Appendix A Stage 1 Directions

INDIVIDUAL and DYAD: INSTRUCTIONS

Explain project as introducing how to describe data.
Solicit participate with IRB forms. (No extra-credit, so dropout now, please) 5 minutes.

Set up the description of the graph as an important but complex task. Tell them about how their notes will help them think about their message, but will also help me make teaching materials later. The more you note the better! 5 minutes.

Show them some sample note taking ideas, use the directions from the top of the Notes sheet. 8 minutes.

Distribute Graph 1, and writing paper for Stage 1 writing ….12 minutes
Give more time for notes, in case they think of something novel.

8 minutes for note taking on drafts and refining notes.
(NO FURTHER WRITING ON THE PROMPT!)

Appendix B Graph 1, used in Stage 1 and Stage 3[5]

Please describe the information in the graph for Figure B.1 below. Your description should be six to nine sentences in length and about 150 words. You can write more or less, just make sure you feel your description is complete.

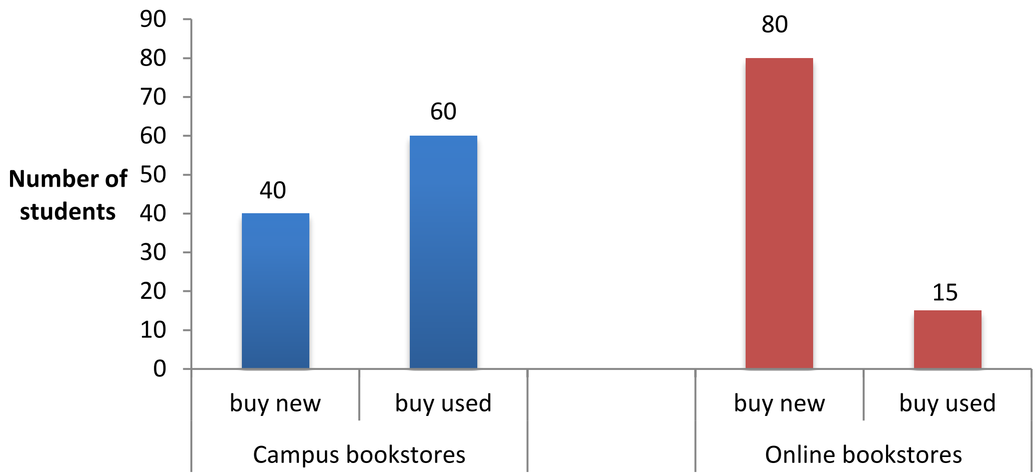

Figure B.1. *The number of students from one physics class who bought a new or used physics textbook by bookstore location*

.

[5] Dissertation formatting rules required this "Figure 1" be relabeled "Figure B.1." The version learners worked with was labeled "Figure 1."

Appendix C Writing sheet for all groups

.

Appendix D Notes sheet (Individual and Dyad conditions)

On this sheet, please take notes on language or concepts you think you can improve or want to know. You can use the phrases we practiced in class, such as: "I don't know how to say X in English", "I wrote X, but I'm not sure if this is correct", "What is the correct tense for X?" and "I'm not sure whether the graph is describing X". You can also make other notes on this sheet, but your notes should be about the language and ideas you want to use in the description of the graph.

.

Appendix E Stage 2 Directions (Individual and Dyad)

First: Read the sample paragraph.
After you have considered each sentence, **highlight any words or phrases** you think are useful for describing graphs like this one.

READ THIS: It is very important for you to think carefully about the graph and its sample description. You should try to learn the structure and the language in the sample essay. It is important you think about any differences you may have about your own perspectives and those in the sample essay. Your goal is to understand how each sentence in the sample contributes to the overall description of the graph.

Try to understand the sample essay so you feel comfortable writing your own graph description at the end of the session.

DYAD: INSTRUCTIONS

First: Read the sample paragraph.
After you have considered each sentence, **highlight any words or phrases** you think are useful for describing graphs like this one.

[Students should individually look at sample graph and model description] about 5 minutes later

Now you will work with a partner. To help you and your partner's understanding of the sample essay, talk about each numbered sentence one by one.

- Please take turns **reading aloud** each sentence of the sample paragraph, **and discuss each sentence** with your partner.
- Explain to your partner what each sentence means to you, and how it relates to other sentences in the description. The person who reads aloud can do this **after** the partner explains their ideas first, but you should **both explain your ideas**.

READ THIS: It is very important for you to work together as a team. You should help each other learn the structure of the sample, and learn how to use the language in the sample essay. It is important to talk over differences in your ideas or explanations and try to reach an understanding of how each sentence contributes to the overall description of the graph.

Try to understand the sample essay so both of you feel comfortable writing your own graph description at the end of the session..

Appendix F Stage 2 Graph 2 (Individual and Dyad)[6]

This is a sample graph description

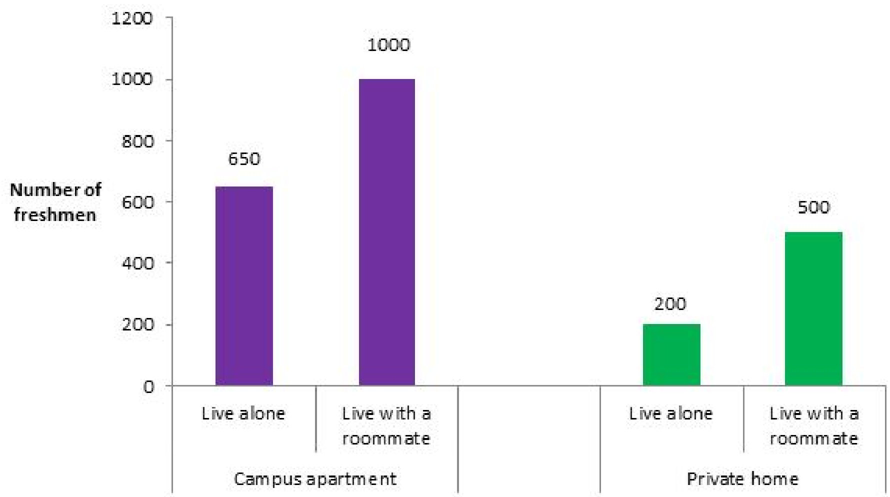

Figure F.1. *The number of freshmen who live alone or with a roommate by student living location*

[1] Figure F.1 is a grouped-column graph that shows the survey results for the living situation of 2,350 freshmen who live alone or with a roommate by student living location. **[2]** The number of freshmen is on the vertical axis and both the living location and the living situation are on the horizontal axis. **[3]** The purple columns represent the freshmen who live in campus apartments, and the green columns represent the freshmen who live in private homes. **[4]** For campus apartments, 1000 freshmen live with a roommate and 650 freshmen live alone. **[5]** This means a majority of freshmen in campus apartments live with a roommate. **[6]** For private homes, this pattern is the same with 500 freshmen living with a roommate and only 200 living alone. **[7]** If the two living locations are compared, it can be seen that living in campus apartments is more common than living in a private home. **[8]** Regardless of the living location, however, living with a roommate is more popular than living alone.

[6] Dissertation formatting rules required this “Figure 1” be relabeled “Figure F.1.” The version learners worked with was labeled “Figure 1.”

Appendix G Stage 2 worksheet used in the Individual condition

Now you will see a sample essay. Please read each sentence carefully. While you read, think about each numbered sentence one by one, and consider how it relates to the other sentences in the sample paragraph. Use the table below to organize your thoughts.

Sentence	The purpose of this sentence is to… (be specific and use details)
[1]	
[2]	
[3]	
[4]	
[5]	
[6]	
[7]	
[8]	

Answer the following questions; be sure you write down your responses:

1) Do you think the data in the graph is believable? Why or why not?

2) If you could add something to (or, remove something from) the sample graph description, what would you add or remove?.

Appendix H Stage 2 worksheet used in the Dyad condition

Now you will see a sample essay. Please read each sentence carefully. While you read, think about each numbered sentence one by one, and consider how it relates to the other sentences in the sample paragraph. Use the table below to organize your thoughts.

Sentence	The purpose of this sentence is to… (be specific and use details)
[1]	
[2]	
[3]	
[4]	
[5]	
[6]	
[7]	
[8]	

Answer the following questions and share your responses with your partner. Be sure you write down your responses, and any responses your partner has that you think are useful.

1) Do you think the data in the graph is believable? Why or why not?

2) If you could add something to (or, remove something from) the sample graph description, what would you add or remove?.

Appendix I Worksheet and activity for Control condition used in Stage 2.

Directions for teacher:

Call out question: “Do you think online bookstores will replace face-to-face bookstores?” Field responses by students, connecting their ideas to the notion that both schools and students have begun to rely on technology for their school needs, whether for buying textbooks or for course webpages. Next call out question: “Do you use electronic texts, such as online textbooks or electronic texts, in any of your courses; how about course management platforms, such as Canvas and BlackBoard?” After students finish writing responses, have whole class discussion, then distribute the Graph 1 and writing paper.

Now we will complete an activity: Distribute “What is your perspective?” Worksheet (excerpted below).

Sample of worksheet

What is your perspective?

1. *Do you use electronic texts, such as online textbooks or electronic texts, in any of your courses?*

   “yes”

   - *Are you satisfied with your experience? How could electronic texts be improved?*

   “no”

   - *If you could use electronic textbooks in your coursework, what courses would you like to use electronic textbooks? Why do you think electronic textbooks would be useful for learning?*

2. *Are there some courses that seem like they have content that* ***might not be better*** *in the digital format than in the traditional paper format? What are they and why do you think so?*

3. *What tools (e.g., class announcements, discussion boards) and applications (e.g., safe-assign, online quizzes, gradebook) found on course management platforms, such as Canvas and BlackBoard, are useful for your learning?*

4. *How do you think electronic textbooks and ebooks (*ebooks are digital versions of novels and short stories) will change libraries?*

5. *Do you think technology will completely replace traditional modes of entertainment (e.g., electronic books replace paperbacks, streaming movies and music online replace DVDs and CDs)?.*

VITA

VITA

EDUCATIONAL BACKGROUND

Purdue University, West Lafayette, IN **December 2015**
Ph.D. English/Second Language Studies

The University of Hawaii, Manoa, Honolulu, HI **December 2007**
M.A. Second Language Studies

ACADEMIC PUBLICATIONS (since 2012)

Koyama, D., Sun, A., & Ockey, G.J. (forthcoming, 2016). The effects of item preview on video-based multiple-choice listening assessments, *Language Learning & Technology, 20*(1).

Ockey, G.J., **Koyama, D.**, Setoguchi, E., & Sun, A. (2015). Validity of the TOEFL iBT speaking section for Japanese university students. *Language Testing, 32*(1), 39–62.

Koyama, D. (2014). [Review of the book Collaborative Writing in L2 Classrooms, by Neomy Storch]. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *14*(1), 39–43.

Ockey, G.J., **Koyama, D.**, & Setoguchi, E. (2013). Stakeholder input and test design: A case study on changing the interlocutor familiarity facet of the group oral discussion test, *Language Assessment Quarterly, 10*(3), 292–308.

Koyama, D., & Setoguchi, E. (2012). The effect of opening gambit types on group oral proficiency scores, *Studies in Linguistics and Language Teaching, 22,* 112–12.