The VVT Project: A web-based platform for strategy instruction and research into self-regulated learning of L2 vocabulary

James M. Ranalli
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Bilingual, Multilingual, and Multicultural Education Commons, and the Other Languages, Societies, and Cultures Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Graduate College at Digital Repository @ Iowa State University. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
The VVT Project: A web-based platform for strategy instruction and research into self-regulated learning of L2 vocabulary

by

James M. Ranalli

A dissertation submitted to the graduate faculty in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY

Major: Applied Linguistics and Technology

Program of Study Committee:
Carol A. Chapelle, Co-major Professor
Volker Hegelheimer, Co-major Professor
Roberta J. Vann
Geoffrey F. K. Sauer
Alison L. Morris

Iowa State University
Ames, Iowa
2012

Copyright © James M. Ranalli, 2012. All rights reserved.
Para a Leonor,
que mexeu no destino e mudou a minha sorte
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>vi</td>
</tr>
<tr>
<td>CHAPTER 1: INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>Organization of the dissertation</td>
<td>5</td>
</tr>
<tr>
<td>References</td>
<td>7</td>
</tr>
<tr>
<td>CHAPTER 2: DESIGNING ONLINE STRATEGY INSTRUCTION FOR INTEGRATED VOCABULARY DEPTH-OF-KNOWLEDGE AND WEB-BASED DICTIONARY SKILLS</td>
<td>9</td>
</tr>
<tr>
<td>Abstract</td>
<td>9</td>
</tr>
<tr>
<td>Introduction</td>
<td>9</td>
</tr>
<tr>
<td>Approach</td>
<td>15</td>
</tr>
<tr>
<td>Design</td>
<td>20</td>
</tr>
<tr>
<td>Procedures</td>
<td>27</td>
</tr>
<tr>
<td>Conclusion</td>
<td>37</td>
</tr>
<tr>
<td>Notes</td>
<td>39</td>
</tr>
<tr>
<td>References</td>
<td>39</td>
</tr>
<tr>
<td>CHAPTER 3: ONLINE STRATEGY INSTRUCTION FOR INTEGRATED DICTIONARY SKILLS AND LANGUAGE AWARENESS</td>
<td>46</td>
</tr>
<tr>
<td>Abstract</td>
<td>46</td>
</tr>
<tr>
<td>Introduction</td>
<td>46</td>
</tr>
<tr>
<td>Methods</td>
<td>52</td>
</tr>
<tr>
<td>Results</td>
<td>63</td>
</tr>
<tr>
<td>Discussion</td>
<td>74</td>
</tr>
<tr>
<td>Conclusion</td>
<td>76</td>
</tr>
<tr>
<td>Notes</td>
<td>76</td>
</tr>
<tr>
<td>References</td>
<td>77</td>
</tr>
<tr>
<td>CHAPTER 4: FEEDBACK, MONITORING ACCURACY, AND SELF-REGULATION IN MOTIVATED LEARNING OF L2 VOCABULARY DEPTH OF KNOWLEDGE</td>
<td>84</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Abstract</td>
<td>84</td>
</tr>
<tr>
<td>Introduction</td>
<td>84</td>
</tr>
<tr>
<td>Methods</td>
<td>98</td>
</tr>
<tr>
<td>Results</td>
<td>108</td>
</tr>
<tr>
<td>Discussion</td>
<td>116</td>
</tr>
<tr>
<td>Limitations and suggestions for further research</td>
<td>121</td>
</tr>
<tr>
<td>Notes</td>
<td>122</td>
</tr>
<tr>
<td>References</td>
<td>123</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 5: TASK DEFINITION IN L2 STRATEGY USE: INVESTIGATING DISCREPANT CASES WITHIN A MODEL OF SELF-REGULATED LEARNING</th>
<th>133</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>133</td>
</tr>
<tr>
<td>Introduction</td>
<td>134</td>
</tr>
<tr>
<td>The study</td>
<td>141</td>
</tr>
<tr>
<td>Results</td>
<td>152</td>
</tr>
<tr>
<td>Discussion</td>
<td>165</td>
</tr>
<tr>
<td>Conclusion</td>
<td>169</td>
</tr>
<tr>
<td>Notes</td>
<td>169</td>
</tr>
<tr>
<td>References</td>
<td>170</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>CHAPTER 6: CONCLUSION</th>
<th>177</th>
</tr>
</thead>
<tbody>
<tr>
<td>References</td>
<td>179</td>
</tr>
</tbody>
</table>
ACKNOWLEDGEMENTS

My sincere thanks to the following people who, in various ways, made this project possible or helped me to complete it:

- to the members of my POS committee for their guidance and encouragement throughout the process: Alison Morris, Geoff Sauer, and Roberta Vann, and especially to my co-major professors, Carol Chapelle, for sharing her vast knowledge and experience, and Volker Hegelheimer, for providing the initial inspiration for the whole thing;
- to the ESL instructors at Iowa State University who allowed me to work with them and their composition students: Richmond Dzekoe, Jinrong Li, Zhi Li, Hong Ma, Edna Lima, and especially Hyejin Yang;
- to John Levis, for sponsoring the grant that provided me with a crucial year in which to sort out basic technology and design issues;
- to Mack Shelley, Dave Peters, Fred Lorenz, and Chris Zittnan, for their invaluable help with the statistics;
- to Susan Hunston, for taking the time to provide feedback not only on the first paper but on the VVT course itself;
- to Gordon Bateson, the developer of one of the key technologies used in the project, for his meticulous work and responsiveness;
- to Mike Colvin, for the chats about academic productivity, which gave me great ideas for managing cognitive load during more than a year of intensive analysis, secondary research, and writing;
- to Jorge Rodriguez, for introducing me to the productivity-enhancing powers of yerba maté;
- to Charlie Johnson, for listening and commiserating during all those runs around the lake;
- to Ulugbek Nurmukhamedov, for his constructive advice and contagious enthusiasm;
- to my professors, fellow students and friends in the TESL/AL program, for their camaraderie and good humor, which made Ross Hall an enjoyable place to be;
- and most of all to Leonor Leandro, my wife, whose love and support kept me going.

I also gratefully acknowledge the College of Liberal Arts and Science’s Computational Advisory Committee (LASCAC) for supporting this project with a grant.
ABSTRACT

The focus of this dissertation is a web-based, second language (L2) instructional resource called VVT (Virtual Vocabulary Trainer) designed to teach integrated vocabulary depth of knowledge and dictionary referencing skills to tertiary-level learners of English as a Second Language (ESL). In addition to evaluating the potential of online resources to address long-standing challenges in the field of L2 strategy instruction, the project also targeted research objectives in other related areas. The work is described in four separate articles aimed at peer-reviewed journals in the fields of computer-assisted language learning (CALL), applied linguistics, and teaching English as a Second Language. The first paper discusses the theoretical underpinnings of the VVT course, the procedures followed in developing it, and the materials themselves, while also suggesting general design principles for L2 strategy instruction based on frameworks derived from cognitive theory. The second paper represents the evaluation component of the project and reports how the feasibility of automated, online strategy instruction was investigated by studying the resource’s effectiveness, both actual and perceived, in an experiment involving 64 ESL composition students. In the third paper, the VVT course serves as the platform for a study critiquing a recently proposed, structural model of L2 vocabulary learning, while at the same time adding to the literature on the acquisition of L2 vocabulary depth-of-knowledge features. The final paper uses discrepant cases from the evaluation study, i.e. individual participants whose performances diverged significantly from group norms, to investigate the metacognitive strategy of task definition, the stage of learning in which students develop internal, and potentially idiosyncratic, representations of instructional tasks. The findings provide evidence of the feasibility of automated, online strategy instruction for complementing teacher-led forms, while also shedding light on the challenges many L2 learners face in self-directed learning of vocabulary depth of knowledge. Finally, it demonstrates the potential of an integrative, multicomponential model of self-regulation for researching and theorizing about L2 learning.
CHAPTER 1: INTRODUCTION

The outcome of any serious research can only be to make two questions grow where only one grew before.

– Thorstein Veblen

The focus of this dissertation is a web-based, second language (L2) instructional resource called VVT (Virtual Vocabulary Trainer), which I developed to teach vocabulary learning strategies to tertiary-level learners of English as a Second Language (ESL). I started this dissertation project with one relatively modest goal, i.e. to test the efficacy of this resource, but added more goals as the project expanded over time to encompass not only evaluation of the VVT but more general research objectives in a number of L2 domains. These are described here in four separate but related papers targeted at peer-reviewed journals in the fields of computer-assisted language learning (CALL), applied linguistics, and teaching English as a Second Language. While these papers are written as third-person research reports, this introduction uses a first-person narrative to explain how the project and its objectives evolved.

The VVT project started as an experiment to see whether web-based video tutorials, which are frequently used for learning computer software and related skills, and which I had come to appreciate in a first-year CALL methods course required in my PhD program, could be used for purposes of L2 strategy instruction (SI). I knew from conducting research for my master’s thesis that a number of systemic barriers have prevented SI from being widely adopted in the field of L2 teaching. These include a lack of expertise among teachers, a lack of space in syllabuses and timetables, and a lack of good materials. In addition, my long experience as a teacher of English as a Foreign Language (EFL) in Namibia, Korea, and Sri Lanka had convinced me that 1) learners are largely left to their own devices in the task of acquiring the L2 lexicon; and 2) while they know how to learn basic form-meaning pairings, most are at a loss as to how to learn many aspects of word usage, which is compounded by their underdeveloped skills at exploiting the pedagogical dictionaries designed to help them with this very task. In hopes of addressing both of these problems, I began building this online resource, using the rudimentary technical knowledge I had at the time.
To bolster this knowledge, I took another course, this one introducing principles of instructional design, in which I was exposed to the literature of that field. Here I found not only practical guidance for the development process (through procedures such as rapid prototyping) but also useful theory in the form of instructional design frameworks. In particular, those frameworks based on Information Processing Theory – which differentiated between instruction targeting declarative versus procedural forms of knowledge, or the orchestration of the two – seemed to have special relevance for the VVT project. I believed, again based on my professional experience, that learners’ problems with vocabulary usage stemmed in part from a lack of conceptual understanding of so-called vocabulary depth-of-knowledge features (such as collocation), in conjunction with underdeveloped dictionary referencing skills, and that addressing these different shortcomings required very different types of teaching from each other. Thus, as the VVT prototype took shape, the project also began to expand beyond evaluation of a particular CALL resource to include the identification of instructional design principles that might inform SI for L2 learning more generally.

When an early version of the VVT prototype was complete, I began piloting it with the help of students and instructors in an ESL composition course at Iowa State University. As part of this piloting, I developed an online assessment task to be used as a measure of the learning that students accomplished through the VVT. It required them to identify and correct sentence-based errors involving grammatical collocation, transitivity, or complementation patterns of verbs, nouns and adjectives, with the help of online dictionaries accessed via hyperlinks. This assessment was employed as both a pre-test and post-test.

The preliminary findings were promising. In the absence of instruction, most students performed poorly, but after the VVT course their scores tended to increase substantially. I also began asking them to estimate their scores after each administration of the assessment task. This source of data was suggested based on psychology studies cited in Kohler’s (2002) doctoral dissertation on SI for L2 metacognitive strategies, and it yielded intriguing results. Before the VVT course, most participants predicted that their scores would be much higher than those they actually achieved, while afterward their actual and estimated scores were much better calibrated. The difficulties students had monitoring and evaluating seemed to add an important new dimension to the project, and I therefore began reading the cognitive psychology literature on confidence judgments.
Around the same time, while searching for an analytical framework that could accommodate both the lexical and strategic focuses of the project, I read a theoretical paper by Tseng and Schmitt (2008), which presented the initial outline of a broad, integrative model of L2 vocabulary learning, with emphasis on the construct of motivation. This paper, which was based on structural equation modeling, captured my attention because it presumed L2 vocabulary learning to be a largely self-directed undertaking, and also conceptualized vocabulary knowledge in terms of depth as well as breadth. In addition, the authors concluded on the basis of their findings that achievement in both types of vocabulary learning is contingent upon mastery of learning strategies or tactics, lending support to the VVT project’s primary objective.

However, Tseng and Schmitt’s structural model included a latent variable called self-regulating capacity, which was in essence a volitional construct, and which had been the focus of an earlier paper by the same authors, written with their colleague at the University of Nottingham, Zoltán Dörnyei. This paper had caused a stir in the L2 strategy field (see, for example, Gao, 2007) by dismissing much previous research for its theoretical fuzziness and lack of psychometric validity. The authors argued that study of L2 learning strategies per se should be abandoned in favor of a focus on self-regulation, concluding that “the most important aspect of strategic learning is not the exact nature of the specific techniques that students employ but rather the fact that they choose to exert creative effort in trying to improve their own learning” (Tseng, Dörnyei, & Schmitt, 2006, p. 95). This struck me as a contradiction. Tseng and Schmitt had, in the space of two years, apparently moved from the position that particular strategies are unimportant to one in which it is essential to master them.

To try to understand, I began to immerse myself in the educational psychology literature on self-regulation, and discovered a diverse area of scholarship that went well beyond the arguably narrow interpretation in Tseng et al. (2006). I also found concepts and principles that might explain the monitoring problems revealed in my piloting. I decided that, in addition to using my participants’ performance data to evaluate the efficacy of the VVT course, I would use the same data, contrasted with the participants’ self-assessed scores, as the basis of a study critiquing Tseng and Schmitt’s structural model and, I hope, contributing to the literature on the acquisition of L2 vocabulary depth of knowledge.
The final piece of the puzzle, with respect to this dissertation, was the individual performances that defied my expectations. The research design had evolved to include treatment and comparison groups, and there were always one or two learners in the latter who excelled, even without instruction. I wanted to find out what allowed them to do so. Also, there were always several participants in the treatment group who would demonstrate lower-than-expected performance on the post-test. Often, these students had failed to complete the VVT materials, ostensibly due to poor motivation or low proficiency in another aspect of their English, as evidenced by similar problems they had in completing assignments for the composition course in which the study was based. Naturally, I also wanted to learn from these cases if only so I could modify the instruction to better cater to their needs. Most interestingly, there were occasionally learners who, despite completing the VVT course, had difficulty fully taking on board key aspects of the instruction, and who continued to demonstrate monitoring problems.

These discrepancies, as well as my readings in the area of self-regulated learning, compelled me to conduct case studies of these participants. I decided to focus on the metacognitive strategy of task definition, which previously published case studies had suggested might be problematic for some ESL learners, and which had been identified as a gap in the L2 strategy research. In addition to contributing to the literature in this area, I also saw it as an opportunity to demonstrate the value of a broader, more integrative model of self-regulated learning for researching and theorizing about both L2 strategies and vocabulary learning.

To summarize, over several years of development, piloting, and secondary research, the VVT project took on the following objectives:

- to test whether automated, online materials could be used for purposes of L2 strategy instruction;
- to teach an integrated form of vocabulary depth of knowledge and pedagogical dictionary skills;
- to identify instructional design principles that might be useful for L2 strategy instruction;
- to critique an important new model of L2 vocabulary learning and demonstrate where it needed further elaboration;
• to investigate the challenges involved in self-directed learning of vocabulary depth-of-knowledge features;
• to explain discrepant cases found in the evaluation study;
• to build on previous research in the area of L2 task definition; and
• to illustrate how an integrative model of SRL might be useful for studies of L2 learning.

ORGANIZATION OF THE DISSERTATION

The objectives form the basis of this dissertation, which consists of six chapters: this introductory chapter, followed by four chapters containing the separate papers and a brief concluding chapter. Each article is summarized as follows:

**Paper 1.** Designing online strategy instruction for integrated vocabulary depth of knowledge and web-based dictionary skills, to be published in CALICO Journal

This paper describes the theoretical underpinnings of the VVT course, the procedures followed in developing it, and the materials themselves. I start by outlining the problems that motivated the project, reviewing findings from the L2 lexicography and learning strategy literatures, and then proposing a solution in the form of an automated, online form of strategy instruction integrating web-based dictionary consultation skills with knowledge of the system of lexical description known as *pattern grammar* (Hunston & Francis, 2000). The remainder of the paper is divided into three sections. The first details the theoretical ideas behind the instructional methodology, which comprised design principles aimed at 1) managing demands on learners’ cognitive processing capacities; 2) differentiating materials so as to appropriately target declarative or procedural forms of knowledge or their orchestration; and 3) harnessing the computer’s multimedia capabilities in principled ways. The article goes on to illustrate how these principles were realized in iterative cycles of development involving user testing, evaluation and redesign, and also describes the selection of tools and technologies that were used. Finally, the paper presents examples of the materials along with commentary on how they illustrate the instructional design principles.
**Paper 2.** *Online strategy instruction for vocabulary depth of knowledge and web-based dictionary skills*, being considered for publication in *Language Learning & Technology*

This paper describes the evaluation component of the project. Specifically, it shows how the feasibility of automated, online strategy instruction was investigated by studying the effectiveness, both actual and perceived, of the VVT course as a prototype. I begin by reviewing literature outlining the problems of L2 strategy instruction, as well as the need to help learners make more strategic and self-directed use of CALL resources. After describing the strategy to be focused on and the rationale for selecting it, I discuss the research methodology, showing how the technology-based nature of the project facilitated a rigorous experimental design including random assignment to treatment and control groups and isolation of the effects of instruction. Measures of performance included scores on the pre- and post-test assessment as well as dictionary-use data obtained from video screen-capture footage, while learners’ perceptions of effectiveness were measured using two separate questionnaire instruments. Discussion of the findings includes a recommendation that automated online SI be combined with teacher-led forms to facilitate skill maintenance and transfer to learners’ own L2 writing.

**Paper 3.** *Feedback, monitoring accuracy, and self-regulation in motivated learning of L2 vocabulary depth of knowledge*, to be submitted to *Language Learning*

In this paper, the VVT course serves as the platform for a study proposing elaborations for Tseng and Schmitt’s model of motivated L2 vocabulary learning (Tseng & Schmitt, 2008), while at the same time contributing to the literature on the acquisition of L2 vocabulary depth-of-knowledge features. The model is first described with reference to its relevance for the field, after which two key shortcomings are highlighted: 1) the lack of differentiation between size-related versus depth-related vocabulary learning in terms of the challenges these might pose; and 2) the model’s inability to account for development of the mastery of strategic behavior identified as a requirement for achievement. The latter is ascribed to the lack of structural connections between two appraisal variables and either of two variables representing strategic behavior, which is seen to preclude any recursive, regulatory function in the model. To demonstrate the need for such a function, the study focused on two key forms of feedback presumed necessary for self-directed vocabulary learning: internal feedback via metacognitive monitoring of strategy use, and external feedback via dictionary consultation. On the basis of the findings, I argue for a more variegated,
and less motivation-focused, model of L2 vocabulary learning that takes self-regulation as its organizing principle, and suggest ways that automated, online strategy instruction such as that exemplified in the VVT course might contribute to future L2 strategy and vocabulary research within a self-regulated learning framework.

**Paper 4.** *Task definition in L2 strategy use: Investigating discrepant cases within a model of self-regulated learning*, to be submitted to *TESOL Quarterly*

The final paper addresses discrepant cases from the evaluation study; that is, individual participants in either the treatment or comparison groups whose performances diverged significantly from group norms. Four case studies form the basis of this investigation, which focused on the role of task definition, the metacognitive strategy whereby learners develop internal, and potentially idiosyncratic, representations of tasks, which Wenden (1995, 1998) identified as an important gap in L2 strategy research. The study builds on the work of Abraham and Vann (Abraham & Vann, 1987, 1996; Vann & Abraham, 1990) by using a form of microanalysis of strategy use involving learners’ task products, triangulated with self-report data comprising both questionnaire responses and thinkaloud protocols, and used in conjunction with a conceptual framework to allow conjectures about cognitive and metacognitive processes. The study is positioned within an integrative, multicomponential model of self-regulated learning (Winne & Hadwin, 1998) that views task definition as a process separate from planning, monitoring and evaluating, i.e. the three metacognitive strategies typically cited in L2 strategy research. I attempt to link each discrepant case to the stage of task definition, thus demonstrating the validity of the construct, while also illustrating the value of Winne and Hadwin’s model of SRL for researching and theorizing about L2 learning strategies.

**REFERENCES**


CHAPTER 2: DESIGNING ONLINE STRATEGY INSTRUCTION FOR INTEGRATED VOCABULARY DEPTH-OF-KNOWLEDGE AND WEB-BASED DICTIONARY SKILLS

A paper to be published in CALICO Journal

Jim Ranalli

ABSTRACT

This paper describes the design and development of a CALL resource that teaches aspects of word combinability (i.e. grammatical collocation, transitivity, and complementation) to tertiary-level ESL learners by integrating conceptual understanding with related text-analysis and web-based dictionary skills. The resource delivers an automated, online form of strategy instruction outside of class time, facilitating self-paced learning with sufficient practice and feedback. In addition to the L2 lexicography, vocabulary, and strategy literatures, the project is informed by instructional design principles aimed at 1) managing demands on learners’ cognitive processing capacities; 2) differentiating materials so as to appropriately target declarative or procedural forms of knowledge or the integration and coordination of the two; and 3) harnessing the computer’s multimedia capabilities in principled ways. It seeks to address a number of related issues in second language studies including: the fact that powerful, pedagogical reference tools go largely under-exploited and misused by their target audience; the apparent lack of understanding of collocational relationships among many L2 learners of English; the challenge of providing strategy instruction cost-effectively; and the need to facilitate more strategic and self-directed use of CALL resources.

INTRODUCTION

Imagine a university instructor teaching an undergraduate ESL composition course. While grading the most recent homework assignment (a short research paper about a 20th century technological innovation), she comes across this sentence:

In the 1950s, American scientists warned people for the harm television could do to society.

The instructor must decide how to address the awkward lexical pattern warn somebody for something. Besides fixing the error herself or simply ignoring it, what options exist? This
scenario will be used to illustrate a number of related problems in instruction in English as a second or foreign language (ESL/EFL) to which this paper proposes a unique solution.

First, the conventional use of an error code (e.g. “SVA” to indicate a problem with subject-verb agreement) would be unavailable in this case because, as noted by Ferris (1999, p. 6), lexical errors are “untreatable;” that is, there are no rules or general principles learners could be taught to facilitate self-correction. Yet not addressing such errors seems unreasonable given how common they appear to be. Chan (2010), for example, identified comparatively high percentages of errors such as *The bus was crowded of people* and *They will not listen him* in her corpus of writing by secondary and tertiary ESL learners in Hong Kong.

Fortunately, help is potentially close at hand in the pedagogical dictionaries produced nowadays by major publishers in the field of English Language Teaching (ELT). Such references seek to provide ESL/EFL learners not only with comprehensive coverage of a word’s meanings but also vocabulary depth-of-knowledge features (see discussion in Read, 2004) including morphological, syntactic and derivational behavior, and collocational, stylistic and register constraints – benefits generally not offered by dictionaries for other languages (Frankenberg-Garcia, 2011). Moreover, many ELT dictionaries are now freely available on the web where, for many purposes, they can be more quickly and conveniently consulted than the paper-based versions (Lew, 2011). Figure 1, for example, shows two entries for the verb *warn* in the online version of the Longman Dictionary of Contemporary English (LDOCE), along with their associated syntactic patterns in bold. The instructor in our scenario might therefore consider providing a link to this dictionary page in the margin alongside the original error.
Recourse to dictionaries, however, is problematic. Many learners appear to lack mental models with which to differentiate among dictionary types (e.g. pedagogical versus those designed for native speakers) and the different purposes for which these are appropriate (Frankenberg-Garcia, 2011). Even advanced learners tend to prefer bilingual dictionaries and to focus on L1-L2 translation equivalents over other types of information (Atkins & Varantola, 1998; Frankenberg-Garcia, 2005), while use of monolingual dictionaries can be impeded by L2 processing constraints (Neubach & Cohen, 1988; Tono, 2001). Learners have trouble distinguishing correctly among senses of polysemous words (Chan, 2011; Nesi & Haill, 2002) and tend to focus only on the initial or final entries (Nesi & Tan, 2011). They often ignore explicit usage information presented via symbols or codes and rely instead on examples for guidance (Bogaards & van der Koot, 2002; Chan, 2012; Dziemianko, 2006), which can lead them astray when, for example, they use a word’s semantic associations to deduce facts about its usage.

Figure 1 Entry for warn, with syntactic patterns and collocations in bold, in the online version of the Longman Dictionary of Contemporary English (www.ldoceonline.com)
Difficulty in exploiting dictionaries must be partly attributed to the way they organize and present information (Dziemianko, 2006; Tono, 2011), but lexicographical research also shows that learners lack basic understanding of some vocabulary depth-of-knowledge features, thus undercutting the basis for consultation and exploitation. In a study involving 40 adult ESL learners in the UK, Nesi surmised that her participants had difficulty recognizing usage information in dictionary examples because “the concepts of transitivity and grammatical collocation were very poorly understood” (2000, p. 115). Chan (2012) drew similar conclusions about transitivity and complementation in her investigation of dictionary use among 31 advanced-level ESL learners at a Hong Kong university. And Frankenberg-Garcia characterized the referencing strategies of the 211 Portuguese university students in her study as “disastrous” (2011, p. 119) with respect to prepositional collocations, noting the participants’ apparent inability to distinguish between these and prepositions of time and place.

Studies of dictionary use usually conclude with calls for learner training, with many recent such calls also identifying a need for accompanying language-awareness raising regarding features of word combinability (Chan, 2011; Frankenberg-Garcia, 2011; Laufer, 2011). While L2 strategy instruction has been found to be generally effective in a recent meta-analysis (Plonsky, 2011), the study concluded that the observed effect sizes were not large enough to resolve lingering questions about cost-effectiveness. Conventional forms of strategy instruction require the commitment and training of teachers as well as the provision of already limited class time (Rubin, Chamot, Harris, & Anderson, 2007). In addition, established models of L2 strategy instruction (e.g. Chamot, Barnhardt, Beard El-Dinary, & Robbins, 1999) may not be suited to the present problem insofar as they generally target procedural knowledge, which could be helpful for certain aspects of dictionary use but less so for developing conceptual understanding of linguistic phenomena. The need to provide a sufficient quantity of relevant practice and timely, individualized feedback, which is important for skill acquisition (Carlson, 2003; DeKeyser, 2007), also poses challenges for teacher-led forms of strategy instruction.

Finally, the three aspects of L2 learning suggested in our hypothetical example – vocabulary depth-of-knowledge, dictionary skills, and strategy instruction – are all encountered by learners under conditions of limited cognitive processing capacity. Learning, not just of languages but in general, is constrained by the limitations of working memory, a fact that teachers and instructional designers should make a guiding principle of their work but which is too often
overlooked (Sweller, van Merriënboer, & Paas, 1998). The implications are doubly important for L2 learning, which is mediated by forms that themselves consume additional processing capacity.

So what can be done to help our instructor, and more importantly, the learner who produced the original error? First, the challenging features of word combinability mentioned above – grammatical collocation, complementation, and transitivity – can be subsumed into a single, pedagogically friendly category using the system of linguistic description known as pattern grammar, developed by Hunston and colleagues (Hunston & Francis, 1998, 2000; Hunston, Francis, & Manning, 1997) as part of the COBUILD project. Pattern grammar\(^1\) combines these features under the umbrella term pattern and emphasizes the systematic relationships between patterning and meaning. For example, when used with an -ing pattern, the verb remember can mean “to have a memory of something” (e.g. *I remember growing up in southern California*), but when used with a to-infinitive it can mean “to not forget to do something” (e.g. *Remember to lock the door when you leave*). Many learners may be familiar with a small number of specific patterns such as these, but not as part of a wider linguistic category or concept, which presumably constrains not only the possibility of self-correction but also the chances that patterns may be acquired through noticing.

While pattern grammar posits that all word classes can be described in terms of patterning, Hunston and colleagues have focused their corpus research on the patterns of the more frequent verbs, nouns and adjectives, and this research has informed the design of ELT dictionaries (as illustrated in Figure 1 above). Yet syllabuses and teaching materials have been slow to incorporate patterns, perhaps because of confusion about how to address them. To be sure, the solution cannot lie in direct teaching of a large number of individual patterns. Besides being impractical in most instructional contexts, this would not empower L2 learners to deal with pattern grammar on their own. Rather, a generalized understanding of pattern grammar can be taught that connects the examples learners may already know and uses them in the formation of a metalinguistic concept or schema. At the same time, the instruction can provide text-analysis and referencing skills that will allow learners to identify new, unfamiliar patterns in input and research them efficiently in pedagogical dictionaries.
For ease of description and teaching, this combination of concepts and skills was given the name *strategic pattern knowledge* and made the goal of the instructional design project described in this paper. Strategic pattern knowledge might not only allow learners to self-correct pattern-based errors pointed out by a composition teacher, but also to deal with this lexical phenomenon more independently in other aspects of learning. Importantly, it might go some way toward addressing an insensitivity to collocational relationships that has been cited as a possible reason for L2 learners’ limited knowledge of formulaic sequences (Flowerdew, 2006; Granger, 1998; Wray, 2002). It might also prove an important complement to the item-focused teaching of depth-of-knowledge features advocated by some experts on L2 vocabulary pedagogy (Laufer, 2010; Lewis, 1997; Nation, 2007).

Furthermore, this skills-based form of lexical knowledge can be taught using automated online instructional materials. As outlined in Ranalli (2009), this can reduce the burden that strategy instruction imposes on teachers and timetables. It would also allow strategy instruction to share in the benefits of CALL already enjoyed in other areas of L2 pedagogy, including convenient access, cost efficiency, easier distribution and recycling of materials, situated learning, the use of multimedia, immediate and individualized feedback, and learner empowerment (Reinders & White, 2010). Online instruction also provides unique affordances to tertiary-level ESL composition teaching, which is the context of this project. As noted by Warschauer (2007), the online space is where our students do much of their work, and it has changed not only writing instruction but writing itself. For example, the web provides student writers with convenient access to powerful instructional aids including, as already noted, learner’s dictionaries. Finally, online instruction makes it possible to conduct non-intrusive forms of research concurrently with instruction through the collection of performance and interaction data, which can facilitate evaluation and research.

To achieve these goals, it will be useful to conceptualize strategic pattern knowledge as a *complex cognitive skill*; that is, a skill made up of both declarative and procedural forms of knowledge that must be integrated and coordinated for successful operation of the whole (Van Merriënboer, 1997). This will help to distinguish the instructional strategies needed to develop the different constituent types of knowledge and skill, and these strategies can then be implemented using forms of presentation and tasks that are appropriate to online learning, and which take account of L2 learners’ more limited cognitive processing capacities.
The project has therefore employed concepts and methods from the field of instructional design in the development of a prototype online course, which will now be described. To distinguish the unique characteristics of the resource from the design principles it exemplifies (since both are presented as innovations), the paper will be structured using the three-part hierarchy for analyzing approaches to L2 instruction proposed by Richards and Rodgers (2001), which has served as the basis for other methodological frameworks in CALL (Hampel, 2006; Hubbard, 1992, 1996). Here, approach will refer to the theories and concepts that guided the project. The following section on design will describe the methods and tools used to 1) decompose strategic pattern knowledge into its constituents, 2) set objectives for the instruction, and 3) iteratively create, test and refine the materials. The third section, called procedures, will provide representative samples of the materials along with commentary on how they demonstrate the principles outlined at the beginning. The paper will conclude by discussing implications and next steps for the project.

**APPROACH**

As is the case with some influential theories of L2 learning strategies (O'Malley & Chamot, 1990) and models for L2 strategy instruction (Chamot et al., 1999), the current project is situated in cognitive theory, which views learning as “an active, constructivist process in which learners select and organize informational input, relate it to prior knowledge, retain what is considered important, use the information appropriately, and reflect on the outcomes of their learning efforts” (Chamot, Barnhardt, El-Dinary, & Robbins, 1996, p. 176). There are, however, some key distinctions and changes of emphasis. In their seminal book, O'Malley and Chamot (1990) also describe learning strategies as complex cognitive skills but view these as consisting solely of procedural knowledge once all their constituents have become automatized. In the present conceptualization, some declarative and procedural forms of knowledge must by necessity remain separate (as explained below) but must also be efficiently integrated and coordinated, and a large focus of instructional design thus becomes appropriately targeting the different components and then orchestrating them in a context of limited processing capacity. Three main theoretical inputs, which are roughly contemporary with each other and in fact share some authorship and key ideas, provided guidance in this regard. Each is now briefly summarized and its relevant principles are then outlined and exemplified with reference to the current objectives.
Cognitive Load Theory

Of great relevance to this project and to L2 instruction more generally is cognitive load theory (Sweller et al., 1998; Van Merriënboer & Sweller, 2005), which focuses on the limitations inherent in cognitive processing. Based on a generally accepted two-part division of cognitive architecture into working memory (WM) and long-term memory (LTM), the theory states that only about seven items can be held in WM at any one time, and if these items interact, the number is lower because interaction requires processing. There are apparently no limits to LTM, but its contents are only available for processing when activated in WM. The phenomenon of chunking allows WM capacity to increase by subsuming otherwise disparate factual and episodic knowledge into containers called schemata. Schemata also expand the capacity of WM because they can be organized hierarchically and embedded within each other (Gagné, Yekovich, & Yekovich, 1993). In addition to declarative knowledge, schemata can also include procedural knowledge that, when performed, requires more or less conscious control (and thus processing capacity) depending on the degree to which it has been automatized. According to cognitive load theory, therefore, an important goal of instruction should be to develop learners’ domain-appropriate schemata through both concept formation and rule automation to help ease processing constraints during learning.

Cognitive load is the amount of processing capacity required by a particular learning task. Sweller et al. (1998) identify three types. Intrinsic cognitive load is the amount inherent in the task; it cannot be altered by manipulating instruction. Germane cognitive load is when processing capacity is devoted to schema construction or rule automation; in other words, to learning. (Germane cognitive load is thus the goal and the “sweet spot” of instructional design.) The third type, extraneous cognitive load, is generated when instructional materials require processing that contributes neither to schema construction nor rule automation.

Imagine the instructor in our scenario directing the author of the error to correct it by consulting the word warn in a learner’s dictionary. A certain amount of processing will be required to hold the objective in WM while performing the search (intrinsic cognitive load). If the learner lacks schemata for both pattern grammar and learner’s dictionary entries, he will have insufficient knowledge with which to efficiently categorize the error and organize the information in the dictionary. As a result, he is likely to resort to means-end analysis, a problem-solving strategy.
characteristic of novices in which solutions are tested randomly, one after another, placing increasing demands on WM (extraneous cognitive load). Alternatively, if he has well-developed schemata for the task, the dictionary information can be conceptually ordered and exploited with reference to the error in an efficient way, with some surplus processing capacity perhaps devoted to remembering the particular pattern used in the correction (germane cognitive load).

The pedagogical consequences of schemata being composed of both declarative and procedural knowledge are taken up next.

The 4C/ID Model

The Four Component Instructional Design (4C/ID) Model is, as the name suggests, a framework for instructional design, but it includes theoretical insights regarding the need for multidimensional forms of training to address the different components of complex cognitive skills. The model was developed by Van Merriënboer (Van Merriënboer, 1997; Van Merriënboer, Clark, & De Croock, 2002) for training in modern technical job skills such as air traffic control and computer programming. Such skills are distinguished from industrial-age job skills insofar as they require schema-based understanding of complex systems and advanced problem-solving abilities in addition to psychomotor and lower-order cognitive skills. The rationale for applying this framework to the domain of L2 learning strategies is that, to the extent that a particular strategy requires analyzed knowledge of the L2 (Abraham & Vann, 1996; Bialystok & Ryan, 1985), it will involve conscious processing of a complex system (McLaughlin, Rossman, & McLeod, 1983).

The 4C/ID framework differentiates between forms of knowledge representation by decomposing complex cognitive skills into their constituents, which are then categorized as either recurrent skills or non-recurrent skills. Recurrent skills, based in procedural knowledge, are used for those aspects of a task that are the same from situation to situation. With enough practice, they can become highly automatic such that they require little conscious control. Non-recurrent skills, based in declarative knowledge, can be thought of as “novel” because they allow one to address the unique aspects of a particular task, and “effortful” insofar as they require conscious control (Van Merriënboer et al., 2002, p. 42). In addition to task definition, non-recurrent skills also facilitate monitoring and evaluation of performance. The major differences between recurrent and non-recurrent skills are summarized in Table 1.
Table 1 Comparison of recurrent and non-recurrent constituent skills, following Van Merriënboer (1997)

<table>
<thead>
<tr>
<th></th>
<th>Recurrent skills</th>
<th>Non-recurrent skills</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type of knowledge</td>
<td>Procedural</td>
<td>Declarative</td>
</tr>
<tr>
<td>General characteristics</td>
<td>Fast and accurate, but inflexible</td>
<td>Flexible, but slow and error prone</td>
</tr>
<tr>
<td>Type of processing</td>
<td>Automatic</td>
<td>Controlled</td>
</tr>
<tr>
<td>Contribution to the whole skill</td>
<td>Free up cognitive resources that can be devoted to non-recurrent skill operation by reducing cognitive load</td>
<td>Allow application of the whole skill to new tasks and situations by facilitating task definition as well as monitoring and evaluation of performance</td>
</tr>
</tbody>
</table>

The key to performance of a complex cognitive skill is the successful orchestration of its constituents. For example, our hypothetical composition student may have a well-developed schema that helps him identify potential pattern-grammar errors in text, and another schema that facilitates efficient searching of learner’s dictionaries, but by themselves these do not constitute strategic pattern knowledge. His schemata must be integrated and coordinated so that 1) pattern identification leads to effective dictionary searches, and 2) dictionary data is interpreted accurately with reference to the original context of use.

The four components in the title of the model constitute the four main elements of instruction:

**Just-in-time information.** JIT information is presented immediately before, or at the same time as, learners are expected to perform a task aimed at developing recurrent skills. The temporal contiguity limits the demands on working memory and allows the declarative representation of the procedure to be transformed into a procedural program, which in skill acquisition terms is called *compilation* or *rule automation*.

**Supportive information.** This information helps learners develop schemata in long-term memory. Supportive information may represent a conceptual model of a system or steps in a cognitive strategy for addressing a particular type of problem. To reduce cognitive load, supportive information is not presented while learners are performing a task but before, so that the resulting schema stored in long-term memory can be activated in working memory during performance.
Part-task practice. In part-task practice, constituents of the whole skill are practiced piecemeal, supported by the activation of relevant schemata or the presentation of JIT information. Part-task practice targeting recurrent skills focuses on compilation and strengthening of procedural knowledge, while practice targeting non-recurrent skills focuses on schema construction through induction (mindfully abstracting away from particulars to form a general concept) and elaboration (incorporating new information into existing knowledge).

Learning tasks. These are practice activities focusing on the whole-task or approximations of the whole-task that give learners a chance to develop automaticity in integrating and coordinating the various constituent skills of which a complex cognitive skill is composed. Not only are the contents of JIT and supportive information important but also the manner of presentation, which is the rationale for the third theoretical input to the project.

Multimedia Learning Theory

Multimedia Learning Theory (Mayer, 2002; Mayer & Sims, 1994), which has featured in contemporary L2 studies involving hypermedia annotation (Chun & Plass, 1996; Plass, Chun, Mayer, & Leutner, 1998), proposes that mental representations built on input from both visual and verbal/auditory sources of information lead to better learning than those built on information from either source alone. The theory is based on models of human memory (Baddeley, 1990) and cognition (Paivio, 1991) that postulate separate channels for visual and verbal information. Verbal information can be written or spoken, while visual information can be a static printed image or a dynamic animation or video. Learners actively select, organize and integrate information from both channels to create coherent knowledge representations. A number of empirical studies have shown that multimedia presentations result in knowledge that is better retained and transferred, which Mayer characterizes as “meaningful learning” (2002, p. 101).

The amount of information that can be processed in either channel is limited, however. Mayer and associates have drawn on cognitive load theory in proposing a number of principles for instructional design. Those relevant to the current project are:

Spatial contiguity. Fewer demands are made on working memory when verbal and visual information are integrated, rather than separated, in presentation.
Coherence. Verbal information, images and sounds are instructive to the degree that they are relevant and explanatory. Interesting but unrelated text, background music or decorative designs may impede learning if they divert processing away from the selection, organization or integration of information.

Modality. In the case of animation, related verbal information presented as on-screen text requires the visual channel to process two forms of incoming information, leading to possible overload. If the animation is accompanied by the same verbal information as spoken narration, the load can be distributed between the two channels.

Redundancy. Presentation of the same information via two channels or modes, such as verbal information presented via narration and on-screen text, can divert processing capacity by prompting searches for differences to reconcile. Presentations should therefore aim for complementarity and avoid redundancy. (In the project described here, this had to be balanced against the need for L2 comprehension support, as described below.)

Signaling. Certain information can be highlighted in presentation to guide learners in selecting and organizing, either because that information is important or shows how other information is related.

Personalization. An informal, conversational style in presenting verbal/auditory information is preferable to a detached, formal style because the former can activate a conversation schema and accompanying conversational maxims (Grice, 1975). Thus a learner may attend to narration with greater commitment than if the speaker was perceived as less of an interlocutor.

With these instructional design principles guiding the project, the next step was to operationalize strategic pattern knowledge, elaborate its constituents, and identify forms of presentation and tasks that could be used to teach it.

DESIGN

The current project was both propelled and burdened by a number of innovative elements, including the adaptation of a model for technical skills training to L2 pedagogic purposes and the goal of an automated, online form of strategy instruction. As a result, a great deal of information necessary for all stages of the design was unknown at the beginning, making a linear approach impossible. Instead, as illustrated in Figure 2, the three main design processes overlapped in time
in an iterative and recursive way, with work done in one process often being revised on the basis of products from another. To address these challenges, the project borrowed a number of methods and tools from the field of instructional design, which are described below.

The first process was analyzing the skill of strategic pattern knowledge to identify its components that would require instruction. To do so, principles and techniques were adopted from skill decomposition (Van Merriënboer, 1997), an initial design procedure which breaks complex cognitive skills into their constituents and categorizes them as recurrent or non-recurrent. This was accomplished in a number of ways, for example, by reviewing related research and published instructional materials, and by introspection on the part of the author during performance. The process was also informed by cognitive task analysis (Clark, Feldon, van Merriënboer, & Yates, 2008), which uses techniques such as thinkaloud and observation to uncover implicit or underlying schematic forms of knowledge that may be required for performance. The skill-analysis process revealed, for instance, that efficient use of online dictionaries can be impeded if learners lack the ability to use shortcut keys for copy-and-paste and navigation between application windows. It also established the need to make patterns more relevant by contextualizing them within a superordinate schema for vocabulary depth-of-knowledge features.

Figure 2 Design processes are shown in filled rectangles extending and overlapping in time. Non-timebound methodological and technological inputs, which informed or facilitated the design processes, are shown in unfilled rectangles.
The thinkalouds and observations were conducted as part of rapid prototyping, which is an instructional design approach adapted from the field of software design. Rapid prototyping involves early and frequent tests of models of instructional materials with potential learners, the goal being to identify problems and refine objectives before making large investments of time, effort or funding (Tripp & Bichelmeyer, 1990). Rapid prototyping is suited for projects in which instructional goals are complex and novel, and when the designer is experienced enough to respond intuitively and creatively to issues arising during frequent usability tests.

These tests proved helpful not only for skill analysis but for instructional planning and materials design. They were conducted with the help of students at Iowa State University’s Intensive English and Orientation Program, where the author worked as coordinator of the Language Learning Center. Higher-level students whose vocabulary knowledge approximated that of the eventual target users were identified and encouraged to participate in usability sessions as an alternative form of self-access learning. This also facilitated participatory design (Willis & Wright, 2000), which involves stakeholders in the design process to ensure the eventual product meets their needs.

On the basis of the skill-analysis process, the following objectives were set for the project: to instruct learners in a basic form of strategic pattern knowledge such that, at the end of training, they would be able to 1) identify potential patterns in given sentences; 2) reference them in a learner’s dictionary to determine whether they are used appropriately; and if not 3) use dictionary data to make suitable corrections. An acknowledged limitation is that these objectives do not include transfer of the skill to learners’ own writing. Transfer is a challenge for any instructional intervention (Salomon & Perkins, 1989) and one which the project will address later if the basic concept is proved feasible (see conclusion).

Further skill analysis and prototyping allowed elaboration of the objectives into an instructional plan (Table 2) addressing three schemata that form the basis of strategic pattern knowledge, and which the materials will help learners to construct. For each schema, the recurrent and non-recurrent skills that require training are listed, as well as the supportive and just-in-time information needed to develop those skills. “Word patterns” are the focus of the third schema; the first two, related to “Deep vocabulary knowledge” and dictionaries, are prerequisites.
Table 2 Instructional plan showing targeted schemata along with the recurrent skills, non-recurrent skills, supportive information and just-in-time information addressed in each

<table>
<thead>
<tr>
<th>Schema</th>
<th>Recurrent skills</th>
<th>Non-recurrent skills</th>
<th>Supportive information</th>
<th>Just-in-time information</th>
</tr>
</thead>
</table>
| **Deep vocabulary knowledge** | None             | - Identifying DVK features  
- Categorizing common vocabulary errors according to the DVK feature they represent  
- Locating DVK features in a variety of learner’s dictionaries | - A variety of DVK features exist.  
- DVK features are important for contextually appropriate use of vocabulary.  
- DVK features are best consulted in learner’s dictionaries. | None |
| **Dictionaries**     | - Navigating quickly between windows using shortcut keys and mouse  
- Copying and pasting quickly using shortcut keys and mouse  
- Searching keywords using web browser’s “Find” function  
- Using the “Menu” tool in the LDOCE | - Selecting an appropriate dictionary for one’s purpose  
- Identifying appropriate entry/definition to match sense of word in given context  
- Negotiating long dictionary entries with numerous senses | - Choice of dictionary type depends on one’s purpose.  
- Monolingual learner’s dictionaries are best for encoding purposes.  
- Cognitive strategy for identifying appropriate sense of a polysemous among numerous entries | Visual demonstration of  
- window navigation using shortcut keys and mouse  
- copy-and-paste using shortcut keys and mouse  
- keyword search using web browser’s “Find” function  
- use of the “Menu” tool in the LDOCE |
Word patterns

Further practice with
- navigating quickly between windows using shortcut keys and mouse
- copying and pasting quickly using shortcut keys and mouse
- searching keywords using web browser’s “Find” function
- using the “Menu” tool in the LDOCE

Identifying WPs in sentences and texts
- Referencing WPs in learner’s dictionary to check whether use is appropriate
- Using pattern information to identify appropriate sense of a word
- Using example sentences to distinguish between senses and patterns
- Identifying appropriate pattern for given context
- Using dictionary information to correct WP errors

WPs are common in English with verbs, nouns and adjectives.
- A wide variety of WP types exists.
- WPs help distinguish among the senses of different meanings of frequent words.
- WPs are important for accuracy and fluency.
- Words with similar meanings often share a WP.
- WPs can structure long stretches of text and span intervening words.
- It is helpful just to notice WPs in texts or dictionary entries. Memorization is not necessary.

Visual demonstration of whole-task practice exercise using constituent skills in combination

Note: DVK = deep vocabulary knowledge, WP = word pattern, LDOCE = online version of the Longman Dictionary of Contemporary English.
The next process involved the design of 1) online tasks to train the recurrent and non-recurrent skills and 2) multimedia presentations to convey the supportive and just-in-time information. This was facilitated by the choice of development environment, which had to fulfill rather demanding criteria. In order to support the overlapping design processes, the authoring tools had to be both plastic, i.e. components can be modified quickly and easily, and modular, i.e. one component of an instructional unit can be changed without adversely affecting other components (Tripp & Bichelmeyer, 1990). In addition, they had to be versatile enough to address the project’s content-related goals while accommodating its instructional design principles. For example, the authoring tools had to be able to:

- integrate text, images, audio and video in principled ways and in conjunction with an adequate variety of text-based activity types;
- provide feedback on users’ performance while collecting information about their interaction with the materials;
- facilitate ways of accessing an online learner’s dictionary that were initially scaffolded but later authentic; and
- allow different conditions of access to the materials based on stage of learning and previous performance.

Initially, the author experimented with Flash-based authoring tools such as Adobe Captivate and Rapid Intake ProForm, but these performed poorly on many of the above criteria. For instance, they offered only a limited number of text-based activity types, which were also difficult to integrate with audio and video. The need for the materials to include a large number of elements meant the project files quickly became very large, which slowed prototyping considerably. The resulting Flash objects were themselves very large, taking a long time to load in users’ web browsers and putting those with slower Internet connections at a significant disadvantage. In addition, incompatibilities in the implementation of SCORM standards meant scores and interaction data often failed to transfer to the learning management system.

After much additional experimentation and research, a mix of technologies was selected. These are described in relation to the role they fulfilled.
Presentations

PowerPoint slideshows were the basis of many of the presentations. The slides consisted of text and images (e.g. clipart, Creative Commons-licensed photos, and computer screenshots) arranged in accordance with the modality principle. Simple animation effects were achieved through the use of slide transitions. The slide presentations were video-recorded with Camtasia Studio, which was also used to capture onscreen demonstrations of the recurrent skills. Off-screen behaviors, such as the use of keyboard shortcuts, were filmed with a pocket video camera. Camtasia Studio was also used for post-production. The video footage was integrated with spoken narration that had been recorded and edited in Audacity, the open-source audio software program. The narration was based on scripts that were specially written in accordance with the coherence and personalization principles (Mayer, 2002). These scripts were also the basis for optional captions that learners could use, if needed, for comprehension, but which could also be turned off to avoid the redundancy effect.

Tasks

Tasks were created with Hot Potatoes, an authoring suite for web-based materials based in HTML and Javascript. Although dated (it is no longer being supported by its developers at the University of Victoria), it remains popular among language instructors for the wide variety of text-based activities that can be created from a basic palette of question types: multiple choice, short answer, cloze, matching, ordering and crossword puzzle. Its versatility arises from the fact that reading texts, images, audio, and video can be integrated into activities easily and in a variety of ways. The activities themselves are easy to create and modify and require no scripting skills.

Delivery and administration

Moodle was the main delivery mechanism. This is an open-source learning management system that includes tools for controlling access to materials, facilitating communication with and among learners, conducting assessments, managing grades, and tracking activity. It also allows selective assignment of tasks to specified subgroups of learners, which can facilitate research. In addition, the project relied heavily on QuizPort, a third-party module for Moodle that allows a designer to link Hot Potatoes exercises and other HTML pages into chains or groups; to add
functionality; and to gather data on users’ performance and interactions, thus achieving the hybridization of client- and server-side CALL technologies described by Levy and Stockwell (2006). QuizPort was used to combine related presentations and tasks into tutorials, which became the main unit of instruction. The tutorials make use of a conditional access feature in QuizPort to create adaptive instructional experiences. Learners first make their way through a tutorial in a set, linear sequence, but when a threshold score is achieved, they are given unrestricted access to a menu of all the presentations and tasks.

While this combination of tools took some time for the author to master, it proved to be modular, plastic and versatile enough to support the goals of the project. The materials that resulted, and the way these embody the instructional design principles, are described next.

PROCEDURES

The instructional principles and procedures take shape in a prototype course called VVT (Virtual Vocabulary Trainer) consisting of 10 tutorials and three whole-task practice sequences. For illustration purposes, the tutorials will be described in relation to the schema they target, as shown in Figure 3. In practice, however, the learning management system is used to distribute the instruction over a period of four to eight weeks, depending on how it is incorporated into a particular course.

Figure 3 VVT home page showing tutorials categorized under the three targeted schemata
Because of the cumulative nature of the skills and knowledge addressed, the sequence of the tutorials is fixed. The mastery score for any tutorial is 80 per cent; learners who score less than this must retake the tutorial to get full credit. Eighty per cent is also the cutoff point for allowing open access to the presentations and tasks that make up each tutorial (as shown in Figure 4). The estimated time needed to complete a tutorial ranges from 20-50 minutes, excluding retakes. In total, the VVT course takes most students 8-10 hours to complete.

**Figure 4** On left, the entry page learners see on their first attempt of a tutorial; on right, the open-access menu of presentations and tasks visible after the tutorial has been completed with a score of 80 per cent or higher

**Deep Vocabulary Knowledge**

These tutorials aim at the development of a schema for “deep vocabulary knowledge” (DVK), which refers to a collection of lexical knowledge types that are distinguished from basic vocabulary knowledge (i.e. form-meaning associations), the type typically focused on by learners. The video shown in Figure 5 aims at elaboration of learners’ pre-existing knowledge to develop such a schema. Various forms of lexical knowledge – spoken and written form, collocation\(^2\), part of speech, register, etc. – some of which will already be familiar, are conceptually and spatially associated with each other while the accompanying narration identifies them as features that “students need to know in order to use words accurately and appropriately.” The CC button in the lower right-hand corner is used to toggle the optional
captions, which are provided for learners who need comprehension support, but which can remain hidden so as not to require additional processing. Like all presentations in the course, this one is less than three minutes long, an intentional design feature to embody the coherence principle.

The introductory tutorials in this section focus on establishing learners’ basic understanding of 10 aspects of deep vocabulary knowledge, including word patterns. Later, exercises such as that shown in Figure 6 aim at further refining the overall schema through induction. Learners are presented with examples of errors involving deep vocabulary knowledge and asked to identify the specific DVK feature that each represents. To manage processing demands, relevant images are provided, which support comprehension so that attention can be directed to form. Also, these tasks are limited to error identification and classification; learners are not required to attempt corrections at this stage, which would likely result in cognitive overload and subvert the instructional purpose. After each set of identification/classification tasks, a presentation

![Figure 5 Video presentation aimed at elaboration of a schema for “deep vocabulary knowledge”; button for toggling optional captions is visible in lower right corner](image_url)
encourages learners to mindfully abstract from the particular example to the wider implications for learning and use of that particular DVK feature.

![Image](error-classification-task.png)

**Figure 6** Error classification task used to elaborate learners’ schema for “deep vocabulary knowledge”

**Dictionaries**

The next tutorials in the sequence deal with dictionaries. The first reviews the 10 aspects of DVK and gives learners simple practice in identifying these in a variety of formats used in different learners’ dictionaries. The second aims at elaboration of learners’ existing schema for dictionaries by making key distinctions between corpus-based dictionaries for L2 learners of English, bilingual dictionaries, and dictionaries designed for native-speakers. The presentations provide examples of appropriate applications of the different types (e.g. *Bilingual dictionaries are more suitable for decoding and monolingual learner’s dictionaries for encoding*) and highlight differences in the kinds of information each provide, as shown in Figure 7.
Recurrent skills are also addressed in a tutorial called “Dictionary Tips and Tricks,” which develops learners’ abilities to consult an online learner’s dictionary quickly and efficiently. Three particular subskills are targeted: 1) fast navigation between the source document and dictionary window using shortcut keys; 2) copying and pasting target items from a source document into a dictionary search box using the mouse or shortcut keys; and 3) conducting keyword searches in lengthy dictionary entries using a browser’s “Find” function. Just-in-time information is used to model each technique, as shown in Figure 8. Learners watch the skills being performed and then practice themselves, first in untimed conditions to support compilation, then in timed conditions to promote strengthening. These skills are further reinforced in later tutorials by reminders to use them whenever an exercise entails online dictionary use. Cognitive load is further managed by encouraging learners to situate source and dictionary windows for simultaneous viewing, if screen size allows.
The non-recurrent skill of finding the appropriate sense of a word in a dictionary entry is also covered. This is crucial as the more frequent and useful words of the language are generally polysemous, and searching a large number of senses can be costly in processing terms (which presumably is why learners tend to focus on definitions at the top or bottom of the entry).

Cognitive load is managed in a number of ways. First, learners are reminded to use the subskills for fast, efficient dictionary searches covered in the previous tutorial. Second, practice activities are developmentally sequenced, with easier tasks first, such as that shown in Figure 9, in which the possibilities are narrowed to two numbered definitions. Then, in later tasks involving long entries with multiple senses, learners are shown how to use their web browser’s “Find” function, as well as a unique contextual “Menu” feature in the LDOCE, for keyword searches.

Figure 8  Video presentation modeling the recurrent skill of copy-and-paste using the right mouse button for quick online dictionary consultation
With the groundwork having been laid, the final set of tutorials takes up word patterns as the main focus. They are again contextualized as a component of DVK. Patterns that may already be familiar to learners, such as those that accompany the verb suggest, are made the starting point for the construction of a larger, more detailed schema for word patterns. Presentations in this section provide supportive information such as the fact that patterns in English attach not only to verbs but also to nouns and adjectives (see Figure 10), and that a wide variety of patterns exists. They also aim at the elaboration of a mental model of learner dictionary entries in which word-pattern information is integrated into the information for different senses of a word, as shown in Figure 11 (which also demonstrates the use of signaling). Such a model can reduce cognitive load by helping learners order dictionary information into meaningful categories that can be selectively attended to depending on the particular search.
**Figure 3** Video presentation that helps learners elaborate a mental model of LDOCE dictionary entries. Signaling shows how word-pattern information is integrated into entries for different senses.
This mental model is then further developed by way of dictionary exploitation exercises such as that shown in Figure 12, which is one of several part-task activities addressing the subskills of strategic pattern knowledge. In this exercise, a sentence with a gap or gaps representing a word pattern is provided alongside the LDOCE entry for the word. At this point, neither dictionary searches nor pattern identification are involved to reduce cognitive load. Learners focus on identifying the appropriate sense of the word and the appropriate pattern in the entry. After finding it, they transfer the pattern into the gap-fill, making changes as needed to tense, number, singular/plural, etc. Incorrect submissions are indicated via immediate feedback and hints are available if learners have trouble providing the right response.

![Figure 4](image)

**Figure 4** Part-task practice developing learners’ skills at exploiting dictionary entries for word-pattern information, which is then applied to a specific context

In addition to elaboration, induction is used to build the word-pattern schema. Figure 13 shows two ordered tasks from a longer sequence addressing the non-recurrent skill of identifying patterns in text. In the first, learners find instances of the abstracted patterns provided at the top of the exercise by clicking on the node (i.e. lexical) word of each pattern as they find it in the text. They are encouraged to use their browser’s “Find” function to highlight non-lexical elements of the pattern (e.g. a preposition) to narrow the possibilities and reduce cognitive load. When they have identified the instantiation of each pattern, they move on to the second exercise,
in which the node words are now highlighted and non-lexical elements of each pattern are missing. At this point, dictionary use is encouraged to help fill the gaps. An important feature of word patterns shown here is that they can structure surprisingly long stretches of text and span intervening words.

Finally, the training culminates in whole-task practice, as shown in Figure 1. In each task, learners are provided with a sentence containing a word pattern error that is not highlighted. The sentences have been written to contain at least one distractor pattern in addition to the actual error. Learners must first identify the error by clicking on the node word in the pattern. Guessing is discouraged by significant reductions in points for clicking on any part of the sentence that is correct. Instead, learners must confirm their hypothesis in the dictionary, for which a link is

**Figure 5** Tasks demonstrating the use of induction to further construct learners’ word-pattern schema
provided. Clicking on the actual error makes that part of the sentence editable. Learners then choose the appropriate pattern from the dictionary and modify it to fit the context. Automatization and strengthening are supported by imposing a three-minute time limit for each item. The items are organized into three sets of 15, which are made available at 2-3 day intervals, in accordance with the principle of spaced (versus massed) practice (Baddeley, 1990; Van Merriënboer, 1997).

**Figure 6** Whole-task practice activity bringing together most of the constituent skills developed in the VVT course

**CONCLUSION**

Since the completion of the prototype, the project has moved to the evaluation stage. At the time of writing, the VVT course has been tested in four different semesters with nearly 200 learners. The findings show the course to be highly effective with the target group and to be positively received by most students (see Ranalli, in preparation).
If institutional support is forthcoming, the next focus will be curricular integration. To date, the materials have been used only as a parallel but discrete component of an ESL composition course at Iowa State University. This piloting has revealed scope for increasing instructors’ familiarity with pattern grammar. In conjunction with this, the project seeks their collaboration in helping learners integrate strategic pattern knowledge into their L2 writing. Teacher support is seen as crucial for transfer and maintenance of the gains achieved through the VVT course, which could be expanded to include other aspects of vocabulary learning, both depth-related (e.g. lexical collocation) and size-related (e.g. recording and reviewing strategies). Tutorials addressing these objectives are already in development.

In addition to describing the creation of a particular course, this paper has proposed principles for the design of online L2 strategy instruction, based on the concepts of cognitive load, complex cognitive skill, and multimedia learning. These principles could prove useful in addressing recently discussed needs to help learners make more strategic and self-directed use of CALL resources (Hauck, 2005; Hauck & Hampel, 2008; Winke & Goertler, 2008), particularly in contexts where processing demands are likely to be high.

Making the case for CALL learner training, Hubbard (2004) wrote:

We should not release our students into powerful learning environments unprepared: It is our responsibility ... to see that they are able to make informed decisions about how to use computer resources effectively to meet their learning objectives. (p. 51)

Without instructional support that includes the development of greater language awareness, resources such as online ELT dictionaries may be as likely to provide students with cognitive overload as they are learning opportunities. The instructional design approach outlined here can be used to identify and help manage the demands such tools make on learners’ cognitive resources, while at the same time developing the provisional resource of their L2. In other words, it might help learners not only to navigate powerful learning environments more effectively, but to refine their learning objectives by means of increased understanding of what languages consist of and how they work.
NOTES

1. The linguistic phenomena addressed in this project are called by different names (e.g. patterns, colligation, grammatical collocation, constructions) depending on the approach to linguistic description one adopts. Pattern grammar has been adopted here because the author has found the notion of pattern to be instructive for ESL/EFL learners and because the system aligns well with pedagogical dictionary data.

2. Patterns are distinguished from collocations as follows: the former are characterized by the (frequent and nonrandom) co-occurrence of a lexical word with a function word, a word class, or a complementation structure; while the latter feature the co-occurrence of two or more lexical words. These categories are obviously not mutually exclusive.

REFERENCES


CHAPTER 3: ONLINE STRATEGY INSTRUCTION FOR INTEGRATED DICTIONARY SKILLS AND LANGUAGE AWARENESS

A paper being considered for publication in Language Learning & Technology

Jim Ranalli

ABSTRACT

This paper explores the feasibility of an automated, online form of L2 strategy instruction (SI) as an alternative to conventional, classroom-based forms that rely primarily on teachers. Feasibility was evaluated by studying the effectiveness, both actual and perceived, of a five-week, online SI course designed to teach web-based dictionary skills integrated with knowledge of lexical patterns involving grammatical collocation, complementation, and transitivity. Sixty-four learners in a U.S. university-based ESL composition course were matched for vocabulary size and then randomly assigned to treatment and comparison groups in a learning management system. The SI treatment comprised multimedia presentations and a variety of practice activities with immediate, user-specific feedback, while the comparison condition involved recurrent use of dictionaries for referencing vocabulary usage features but no instruction. Performance data showed significant gains among the SI group in contrast to the comparison group, as well as a clear need for such training, while perception data indicated that participants felt the instruction was effective and relevant. The findings point to the as yet untapped potential of this form of instruction to address long-standing cost-benefit concerns about SI, as well as to help learners make more strategic use of CALL resources.

INTRODUCTION

Despite its decades-long history as a focus of scholarly interest, L2 strategy instruction (SI) has yet to fulfill its promise. While it is broadly accepted that achievement in L2 learning depends in large part on learners’ own self-initiated and self-directed efforts, projects designed to teach helpful learning behaviors to those learners most in need of them have faced a number of challenges, not least of which are lingering questions about whether classroom-based SI is worth the effort and time. A recent meta-analysis by Plonsky (2011) found an overall small to medium effect for SI studies, with effectiveness moderated by several contextual, treatment, and outcome variables. Studies addressing certain strategy domains (e.g. reading, vocabulary) had larger effect
sizes than others (e.g. grammar, listening), as did studies focusing on fewer rather than more strategies and those involving longer versus shorter treatments, which could “prompt some to question whether the benefits of SI outweigh the costs” (Plonsky, 2011, p. 23).

Teachers, to whom the burden of SI usually falls, may find it conflicts with curricular goals, their teaching beliefs or teaching styles, or the expectations of students and parents (Rubin, Chamot, Harris, & Anderson, 2007). Even when teachers are willing to undertake SI, they may be ill-equipped in terms of skills and knowledge for promoting strategies, instructional materials, and institutional support. In their seminal book on L2 strategies, O’Malley and Chamot (1990) identified teacher preparation for SI as the greatest obstacle facing the field. More recently, Chamot (2005) has discussed the need to investigate the degree to which effective SI is tied to the characteristics and experiences of individual teachers.

Another key factor in the cost-benefit calculus is the methodology used for SI. In their systematic review of 38 SI studies, Hassan et al. (2005) drew a contrast between interventions based on awareness-raising techniques versus those based on behavior-modeling approaches, and called for more research to determine the relationship between this variable and the durability of gains from SI. To the extent that any SI intervention involves skill acquisition and thus will require that learners be given sufficient practice opportunities (DeKeyser, 2007), ideally accompanied by immediate and specific feedback (Carlson, 2003), this constitutes yet another challenge for teacher-led forms of strategy instruction.

Finally, the difficulties of conducting rigorous experimental research on SI have been noted (e.g. Chamot, 2005; Hassan et al., 2005; Plonsky, 2011). Random assignment to groups and isolating the effects of instructional treatments usually require laboratory settings, but classroom-based research involving participants with real-world goals and motivation profiles offers greater ecological validity. The result, according to A. D. Cohen and Macaro (2007), is that “there is still much work to be done on strategy instruction in order to prove to learners, teachers, and the wider SLA research community that such an undertaking in the classroom is worthwhile” (p. 284).

But what if some elements of SI could be taken out of the classroom and delivered through the computer, thereby lifting some of the burden from teachers and timetables? SI could thus emulate other areas of endeavor, such as corporate job skills, in which online instruction has
been used cost-effectively to complement or even replace existing face-to-face training programs (Kim, Bonk, & Zeng, 2005). It could also allow SI to share in the many benefits of CALL already enjoyed by other areas of L2 instruction, including convenient access, easier distribution and recycling of materials, situated learning, the use of multimedia, data on learner performance and progress, immediate and user-specific feedback, and learner empowerment (Reinders & White, 2010). In addition, it could address recently discussed needs to help learners make more strategic and self-directed use of CALL resources (Hauck & Hampel, 2008; Hubbard, 2004; Winke & Goertler, 2008). Finally, through the capacities of learning management systems (LMSs), online SI might also facilitate more rigorous and ecologically valid research designs.

While a considerable amount of scholarship has investigated the use of L2 learning strategies in technology-mediated environments (e.g. Chapelle & Mizuno, 1989; Harris, 2003; Hauck & Hampel, 2008; Huang, Chern, & Lin, 2009; Jamieson & Chapelle, 1987; Li, 2009; Pujolá, 2002; Ulitsky, 2000), few studies to date have evaluated computers as a delivery mechanism for SI. Among those that have, A. D. Cohen and Ishihara (2005), O'Bryan and Hegelheimer (2007), and Sykes and Cohen (2008) were small-scale, preliminary studies without controls, while Kohler (2002) and Dreyer and Nel (2003) combined computer-based SI with face-to-face training. The present investigation was therefore designed to directly test the feasibility of an automated, online form of SI by evaluating the efficacy of a purpose-built, prototype course through a combination of experimental and perception data.

It is axiomatic that the strategies targeted by SI should be appropriate to the given context and intended beneficiaries. Furthermore, while some SI approaches address strategies in a general way, experts have called for greater focus on the strategies needed for specific tasks (A. D. Cohen, 1998, 2007; Oxford, Cho, Leung, & Kim, 2004; Wenden, 1995). The context of the present investigation was a tertiary-level ESL writing course for international students with a strong technology component; one class per week is held in a computer lab and most assignments are distributed and collected via an LMS. The students have been identified in a placement test as needing additional instruction before they can register for the normally required composition courses, with the test guidelines noting that their writing exhibits, among other things, “word-choice errors that obscure meaning” and “serious and frequent errors in […] word form.” The study thus sought to teach a specific form of strategic behavior involving online
dictionaries employed in vocabulary usage tasks, which will be described in the following sections.

**Dictionary use among L2 learners of English**

Today’s L2 learners of English are doubly lucky. First, they have at their disposal powerful lexical resources that have been specially written for them and informed by corpus research – advantages not shared by learners of many other L2s (Frankenberg-Garcia, 2011). In addition, most of the major publishers who produce these dictionaries have now made them freely available online where, for many purposes, they can be consulted more quickly and conveniently than the paper-based versions (Dziemianko, 2010).

However, lexicographical research shows that pedagogical dictionaries go widely under-exploited and often misused. Even advanced learners tend to prefer bilingual to monolingual dictionaries and to use them mostly for referencing L1-L2 translation equivalents (Atkins & Varantola, 1998; Frankenberg-Garcia, 2005). Learners have trouble distinguishing correctly among senses of polysemous words (Chan, 2011; Nesi & Hail, 2002), and often consult only the initial or final entries (Nesi & Tan, 2011). They tend to ignore explicit usage information provided via symbols or codes, relying instead on dictionary examples for guidance (Bogaards & van der Kloot, 2002; Chan, 2012; Dziemianko, 2006), which can lead them astray, for example, when they use a word’s semantic associations to deduce facts about its syntactic behavior.

Some consultation problems must be attributed to the way particular dictionaries organize and present information (Dziemianko, 2011; Tono, 2011), as well as limited attentional resources that learners may be unwilling to expend in the midst of other L2 tasks. Monolingual resources in particular impose additional processing burdens (Neubach & Cohen, 1988; Tono, 2001) that may discourage their use.

But research also shows learners may lack basic understanding of certain lexical features, thus undercutting the basis for consultation. Chan (2012) found that 31 advanced ESL learners at a Hong Kong university had trouble using dictionaries to identify and correct lexical errors involving complementation, transitivity and prepositional collocation, with some indicating in thinkalouds that they did not regard these issues as important for grammaticality, while in a focus group others acknowledged never reading usage information in dictionaries. Similarly, Frankenberg-Garcia (2011) elicited preferred reference types for a variety of vocabulary-use
tasks among 211 learners of English at a Portuguese university; most selected a grammar book, as opposed to any type of dictionary, as the best reference for searching prepositional collocations, suggesting an inability to distinguish between these common lexical features and prepositions of time and place. In a study involving 40 adult learners in UK language schools, Nesi (2000) surmised that many participants were unable to exploit usage information in dictionary examples because “the concepts of transitivity and grammatical collocation were very poorly understood” (p. 115).

**Integrating dictionary skills with knowledge of lexical patterns**

Studies of pedagogical dictionary use often conclude with calls for learner training, but such training has tended to make particular dictionaries and their data categories the starting point, rather than authentic language-use problems of the type that should prompt consultation (Lew, 2011). Thus, more recent studies (Chan, 2011, 2012; Frankenberg-Garcia, 2011; Laufer, 2011) have proposed improving learners’ language awareness as part of dictionary-skills training. This was the approach adopted in the present study, whose goal was to teach the strategic use of pedagogical dictionaries for addressing errors involving grammatical collocation, complementation, and transitivity, which appear to be common in some ESL writing (see Chan, 2010).

As a preliminary step, these linguistic features were subsumed into a single, pedagogically friendly category using the system of linguistic description known as pattern grammar, proposed by Hunston and colleagues (Hunston & Francis, 1998, 2000; Hunston, Francis, & Manning, 1997) as part of the COBUILD project. Pattern grammar unites grammatical collocation, complementation, and transitivity under the umbrella term pattern and emphasizes the relationships between patterning and meaning. For example, when used with an -ing pattern, the verb remember can mean “to have a memory of something” (e.g. *I remember growing up in southern California*), but when used with a to-infinitive it can mean “to not forget to do something” (e.g. *Remember to lock the door when you leave*).

Specific patterns such as these may be familiar to many learners, but not as part of a wider linguistic category or concept, because they are typically taught in an ad hoc way, if at all (Hunston et al., 1997). And yet patterns can be seen to feature prominently in today’s pedagogical dictionaries. For example, the entry for the verb *warn* in the online version of the
Longman Dictionary of Contemporary English (LDOCE) shown in Figure 1, highlights several patterns in bolded blue letters, organized according to the separate senses to which they correspond.

![Figure 7 Entry for warn, with lexical patterns in bold, in the online version of the Longman Dictionary of Contemporary English (www.ldoceonline.com)](image)

To make it easier to discuss the dictionary-skills and language-awareness aspects of the study, they were combined into a single construct called *strategic pattern knowledge*. This was operationalized as the ability to use a pedagogical dictionary to identify and correct errors involving lexical patterns. While imparting strategic pattern knowledge might not directly enhance learners’ productive vocabulary skills, it could empower them to deal with pattern grammar more independently in their learning and use of English – for example, in writing and revising. It might also go some way toward addressing an insensitivity to collocational relationships that has been blamed for L2 learners’ poor knowledge of formulaic sequences (Granger, 1998; Laufer, 2011; Wray, 2002).
The present study thus took the form of an experiment contrasting an automated, online SI condition with a comparison condition that involved practice with online dictionaries but no explicit instruction. In line with the task-based nature of the strategy instruction, the evaluation employed a task-based form of assessment administered as a pre-test and post-test. To further enhance validity, strategic pattern knowledge was assessed using both product and process data (Abraham & Vann, 1996), with collection of the latter facilitated by the computer-based nature of the inquiry. A matched randomized block design was used to control for vocabulary size because a moderate but significant correlation with treatment effects had been observed in a pilot study. Perception data was also elicited because learner attitudes toward the usefulness and relevance of specific strategies can influence SI outcomes (Rees-Miller, 1993; Wenden, 1991).

The research questions were:

1. How is the development of strategic pattern knowledge affected by membership in an automated, online SI condition versus an online dictionary practice condition?
2. How do participants perceive the SI materials in terms of efficacy and instructional value?

METHODS

Overview

The study took place in Fall 2011, beginning in the second week of the 15-week semester, as illustrated in Figure 2. The researcher met with each of three different sections of the composition course on regularly scheduled computer lab days to explain the purpose of the study and recruit participants. Students were told gift cards would be awarded to randomly selected participants at the end of the semester, but regardless of participation, all students were expected to complete the treatments as part of the course requirements. After being registered in the project LMS, the students were shown video instructions for the pre-test twice and then had 20 minutes to complete it. A brief, online biodata questionnaire was then assigned for homework. The following week, each section again met in the computer lab to complete an online version of the Vocabulary Levels Test (described below), whose results were used to determine matched pairs for random assignment to groups created in the LMS. One group was randomly selected to work through the SI materials during weeks 4-8 while the other was assigned the dictionary
practice activity. In week 9, computer lab days were again used for the post-test (also preceded by a showing of the video instructions twice), which concluded the experiment phase of the study. The two groups then switched conditions for weeks 10-14 so that everyone could experience the SI materials. Perception data was gathered using short questionnaires administered during the SI treatment and a longer evaluation at the end of the semester (see descriptions below).

* Procedures conducted in a computer lab.

**Figure 2** Overview of procedures showing co-extension of online, LMS-based research components with timeline for the face-to-face composition course from which participants were recruited

**Participants**

The sample ($N = 64$) consisted mostly of students who spoke Mandarin Chinese as their L1 (76.5%). While not very linguistically diverse, such a sample is broadly representative of the current international student body at Iowa State University (according to 2011 figures) and other public universities across the U.S., which is the population to which this study seeks to generalize. After matched random assignment to the strategy instruction (SI) group or dictionary practice (DP) group on the basis of vocabulary size, analysis showed the two groups did not differ significantly in this dimension. As shown in Table 1, the resulting groups were also similar in terms of gender balance, variety of L1s, average time spent in U.S., age, and TOEFL score. Group size ($N = 32$) was more than sufficient to meet the target ($N = 6$) as determined by
an a priori power analysis calculated in G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) using an effect size of 1.32 (Cohen’s $f$) obtained from pilot-study data, a desired power level of .95 and a significance level of .05.

**Table 1** Biodata for the treatment and comparison groups

<table>
<thead>
<tr>
<th></th>
<th>SI group</th>
<th>DP group</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex</td>
<td>Female 14</td>
<td>Female 12</td>
</tr>
<tr>
<td></td>
<td>Male 18</td>
<td>Male 20</td>
</tr>
<tr>
<td>L1</td>
<td>Chinese 26</td>
<td>Chinese 23</td>
</tr>
<tr>
<td></td>
<td>Malay 2</td>
<td>Arabic 4</td>
</tr>
<tr>
<td></td>
<td>[an African language]* 1</td>
<td>Korean 3</td>
</tr>
<tr>
<td></td>
<td>Hindi 1</td>
<td>Malay 1</td>
</tr>
<tr>
<td></td>
<td>Korean 1</td>
<td>Thai 1</td>
</tr>
<tr>
<td>Time in U.S. (months)</td>
<td>$M = 7.5, SD = 14$</td>
<td>$M = 5.7, SD = 8.4$</td>
</tr>
<tr>
<td>Age</td>
<td>$M = 21.6, SD = 5.9$</td>
<td>$M = 20.1, SD = 2.7$</td>
</tr>
<tr>
<td>TOEFL score (iBT)</td>
<td>$M = 76.7, SD = 10.2$</td>
<td>$M = 81.2, SD = 9.8$</td>
</tr>
</tbody>
</table>

*Note: SI = strategy instruction, DP = dictionary practice; * language not specified for reasons of confidentiality;*

**Treatments**

**Strategy instruction (SI)**

Participants in this condition were provided with strategy instruction in the use of learner’s dictionaries for identifying and correcting pattern-grammar errors. The instruction took shape in an automated, online course called VVT (*Virtual Vocabulary Trainer*), which lasted five weeks and consisted of 11 tutorials organized around three topics: vocabulary depth of knowledge (see discussion in Read, 2004), dictionaries, and pattern grammar. The first and third topics were given the more accessible titles *Deep Vocabulary Knowledge* and *Word Patterns*, respectively.

The course is based on instructional design principles that generally aim at 1) managing the demands made on learners’ working memories to facilitate efficient learning, and 2) differentiating appropriately among instruction targeting domain-relevant declarative knowledge (e.g. conceptual understanding of pattern grammar), procedural knowledge (e.g. syntactic processing of sentences to quickly identify patterns), or the successful integration and coordination of the two (see Ranalli, in press, for a detailed description).
The tutorials consisted of video presentations and text-based practice activities designed to help learners:

- differentiate among dictionary types and choose the appropriate type for their purpose;
- use keyboard shortcuts and keyword search tools to speed up online dictionary searches, thus reducing demands on working memory;
- select the appropriate sense of a word from among those provided in dictionary entries;
- identify patterns in sentences and longer stretches of text;
- identify pattern information in dictionary entries;
- choose the appropriate pattern for a given context of use; and
- use pattern information in dictionaries to complete or correct sentences.

Additional activities were included to check learners’ understanding of the material and review previously learned concepts and skills.

Figure 3 shows a screenshot from one of several video presentations aimed at elaborating learners’ schema for pattern grammar by connecting examples they may already know, and then associating this provisional schema with the need for correctness in academic writing. Figure 4 shows excerpts from a sequence of practice activities which further contribute to elaboration of a pattern-grammar schema. In the exercise above, learners study the abstracted patterns given at the top and then click on the instantiation of each pattern in the paragraph, which they locate with the help of their browser’s “Find” tool. The follow-on activity below features the same instantiated patterns now highlighted but missing components, which learners must then supply using dictionary data.
When I first come to USA, I had to spend a lot of time to learn the culture of this country. Because I had watched many American film on TV and at the cinema, I believe I know many things about American culture. It seems like a good strategy of finding out about American idea and state of mind. However, in the real life, I find that I cannot know many thing about American culture from just watch movies. The life is much more complex than as Hollywood like to show, and many people in real life are not like actor or actress. They have more side to see a thing and more
Figure 4 Practice activities designed to elaborate learners’ schema for word patterns

Dictionary practice (DP)

Participants in this condition were assigned a recurring activity in the project LMS requiring them to add entries to a personal vocabulary database. The activity was an online version of the paper-based task that normally constitutes the vocabulary component of the composition course in question. Learners identify words in their course readings or other sources that they want to learn, then research and record lexical features and other information they think will help in using and remembering these words. The model provided to participants is shown in Figure 5. An introductory video explained the purpose of the activity and showed which dictionaries to use and how to create an entry. The database entries were graded each week but the researcher gave no feedback regarding pattern grammar.
Measures of knowledge and strategic ability

Vocabulary Levels Test (VLT)

Vocabulary size was measured using the Vocabulary Levels Test (Nation, 1990). A more recent version (Schmitt, Schmitt, & Clapham, 2001) was selected to reduce the chances that participants had previous experience of particular test items. The test was adapted for online administration and included four frequency levels: the 2000- and 3000-word levels, which assess knowledge of high-frequency words; the 5000-word level, which represents a boundary between low- and high frequency words; and an academic word level, which tests knowledge of the Academic Word List (Coxhead, 1998). Each level contains a selection of 60 words, 30 of which are actually tested. The words are presented in six-item lists positioned next to three synonyms or glosses. Test-takers select a word from the list to best match each synonym/gloss by inputting the corresponding number into the appropriate gap (Figure 6). One point is awarded for each correct response, and the scores are totaled for a maximum possible of 30 on each level and 120 on the test as a whole. Reliability is reported in Table 2.
Figure 6 Screenshot of online version of the Vocabulary Levels Test adapted from Schmitt et al. (2001)

Table 2 Reliability for the online version of the Vocabulary Levels Test

<table>
<thead>
<tr>
<th>Level</th>
<th>Number of items</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>30</td>
<td>.857</td>
</tr>
<tr>
<td>3000</td>
<td>30</td>
<td>.878</td>
</tr>
<tr>
<td>5000</td>
<td>30</td>
<td>.866</td>
</tr>
<tr>
<td>AWL</td>
<td>30</td>
<td>.923</td>
</tr>
<tr>
<td>Test as a whole</td>
<td>120</td>
<td>.964</td>
</tr>
</tbody>
</table>

*Note*: Number of test-takers: 64

Strategic pattern knowledge task

Strategic pattern knowledge was assessed using an online task, which can best be described via example. Learners are presented with a web page in the LMS containing a sentence such as *In the 1950s, American scientists warned people for the harmful effects of television* (Figure 7). In both written and video-based instructions, they have been told the sentence contains a content word (i.e. verb, noun, adjective or adverb) used incorrectly; specifically, there is a problem with
the pattern of the content (or “key”) word that must be identified and corrected using an online dictionary; in this case, the verb *warn*. The learners are to copy and paste the sentence into an editing box, make an appropriate correction, and submit their answer. They are told not to change the key word in any way but to focus on the words around it, adding, deleting or changing them as needed but without altering the meaning of the sentence. A table of links to various dictionaries is provided.

![Image](image-url)

In the sentence below, there is a problem with the pattern of one of the content words. (Content words include nouns, verbs, adjectives and adverbs.) You must identify which word is being used incorrectly, and then rewrite the sentence in the space provided to make it correct. You can COPY and PASTE the sentence into the editing box to save time. You should not change the content word itself in any way. Instead, focus on the words surrounding it. You may need to add, delete or change the surrounding words. Do NOT change the meaning of the sentence.

Use one of these dictionaries, or any other online dictionary, to help you. You may NOT use handheld electronic dictionaries/translation, paper dictionaries, mobile phones, or any resource besides online dictionaries.

In the 1950s, American scientists warned people for the harmful effects of television.

**Answer:**

![Editing Box](editing-box-url)

**Figure 7** Item from the strategic pattern knowledge task showing instructions, dictionary links, sentence to be corrected, and editing box

While they could attempt a correction using existing knowledge, the sentences have been specially selected to contain lexical patterns that are likely to be unfamiliar, so the appropriate course of action is to use one of the pedagogical dictionaries for learners of English, since the other dictionary types (bilingual dictionaries or those designed for native speakers of English) are not reliable sources of this sort of information. After finding the entry for the appropriate word (see Figure 1 above), they should identify the numbered sense that corresponds to the usage in the sentence, select an appropriate pattern from among those provided – in this case, *warn somebody about something* or *of something* – and then return to the task page and make the required changes.
The task, which was used as both a pre-test and post-test, consisted of 10 items presented in randomized order. The sentences were found in the Longman Learners Corpus by searching high-frequency node words featured in pattern-grammar references (Francis, Hunston, & Manning, 1996, 1998). The sentences were edited to make the context clear and to isolate the pattern error as the only infelicity. A list of acceptable corrections was compiled using information from pedagogical dictionaries and the Corpus of Contemporary American English (Davies, 2008), and by surveying eight native speakers of English who have master’s degrees or higher in TESOL or applied linguistics and three or more years of ESL/EFL teaching experience.

Item characteristics are summarized in Table 3. Four items dealt with verb patterns, four with noun patterns, and two with adjective patterns. Most corrections involved changes to prepositions, usually by substitution but in some cases by addition or removal. Some substitutions involved complement clauses: an -ing form in the case of one verb pattern, and a that-clause in the case of two adjective patterns. In several cases, different types of correction were possible.

<table>
<thead>
<tr>
<th>Node word type</th>
<th>Number of items</th>
<th>Correction type required</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>4</td>
<td>substitute complement clause</td>
<td>3*</td>
</tr>
<tr>
<td>Noun</td>
<td>4</td>
<td>substitute preposition</td>
<td>7*</td>
</tr>
<tr>
<td>Adjective</td>
<td>2</td>
<td>add preposition</td>
<td>1†</td>
</tr>
<tr>
<td></td>
<td></td>
<td>remove preposition</td>
<td>2†</td>
</tr>
</tbody>
</table>

Note: 10 items total; * either correction type was acceptable in the case of two items; † either correction type was acceptable in the case of one item.

Polytomous scoring was used to allow differentiation between the ability to identify an error and the ability to correct it. One point was awarded if the participant made any change to the grammatical component of the pattern, i.e. to any part other than the lexical node word. Such a change was interpreted as evidence that the error had been located, so the variable yielded by this score was called ID. An additional point was awarded if the change matched an item in the list of acceptable responses, yielding another variable called CORRECT. With 10 points possible for both ID and CORRECT, the maximum score on the task was 20. Pre-test reliability (Cronbach’s alpha) for ID was .70 and for CORRECT was .71, while the post-test coefficients were .85 and .83, respectively, indicating consistency between the two measures.
Dictionary use measures

In addition to providing performance scores, the strategic pattern knowledge task was also the source of process data assessing use of dictionaries. Each participant’s onscreen performances on both administrations of the task were video recorded using Camtasia, a screen-capture program. The video files were then imported into Transana, a software tool for analyzing audio and video data, where they were coded and transformed into two quantitative measures of referencing ability:

1. **Lookups.** These were counts of the number of individual dictionary searches performed by each participant. Searches were divided into two categories: *learner-dictionary lookups* (i.e. in monolingual, pedagogical dictionaries) and *non-learner dictionary lookups* (i.e. bilingual dictionaries, “native speaker” dictionaries, and non-dictionary tools such as search engines). When screen-capture footage showed a particular word, phrase, or sentence being searched, a lookup was recorded in the appropriate category, except in cases where the same search had already been conducted in the same resource in the same administration of the task.

2. **Lookup points.** These were points achieved on the task that were attributed to use of a learner dictionary, another type of reference, or else to a participant’s use of pre-existing knowledge, intuition, or guesswork. If screen-capture footage showed explicit information in an online resource being used in a correct or partially correct response, the resulting points were coded either as *learner dictionary lookup points* or *non-learner dictionary lookup points*, depending on the type of resource. If language data used in a correct or partially correct response could not be traced to an online resource, those points were attributed to pre-existing knowledge, intuition, or guesswork and coded as *null lookup points*.

Questionnaires

The online instruments used to elicit perception data were 1) a three-item post-tutorial questionnaire administered multiple times, immediately after the participants’ first completed attempt at each tutorial; and 2) a 25-item end-of-course evaluation. The post-tutorial questionnaire, consisting of closed-ended questions, was designed to capture basic “snapshots” of perceived levels of interest, challenge and usefulness of the tutorials while the instruction progressed. The end-of-course evaluation, which included both closed- and open-ended items, sought more detailed, summative perceptions of learning achieved through the materials. Both
instruments were administered anonymously, and separately from those procedures that were screen-captured, to encourage more accurate, less self-protective responses (Dörnyei & Taguchi, 2003).

RESULTS

Research Question 1: Effects of treatment condition

Pattern identification and correction

Descriptive statistics for ID and CORRECT scores are reported in Table 4, along with combined totals and effect sizes. The averages showed that, across time and conditions, more points were achieved by identifying errors than by correcting them, which makes sense given the nature of the task and scoring method, i.e. while every error that is corrected must necessarily be identified, the converse is not true. (However, as will be shown, more errors were both identified and corrected than were identified alone.)

At pre-test, the SI group demonstrated a slight edge over the DP group in both ID and CORRECT, as well as more dispersion in those scores, while at post-test the gaps in scores widened substantially and variation decreased. Using either the general index in J. Cohen (1988) or a more conservative scale recently proposed specifically for L2 research (Oswald & Plonsky, 2010), the effect sizes for pre-to-post differences on either the total combined score, or the ID and CORRECT scores individually, would be characterized as large for the SI group and small for the DP group. The post-test, between-group differences for both measures would also be considered large effects.
Correlational analysis showed a strong, positive relationship between ID and CORRECT \((r = .9, p < .001)\), so a multivariate analysis of variance with these scores as separate dependent variables was not appropriate. Thus the scores were combined and the totals were submitted to a univariate analysis of variance (ANOVA) using the GLM procedure in SPSS. The ANOVA employed a randomized complete block design, with treatment condition and matched pair (based on VLT scores) as the between-subjects factors, and time as the within-subjects factor. Because time had only two levels, sphericity was not a concern, but analysis of histograms and Q-Q plots indicated non-normality in both the pre- and post-test data. Base-10 logarithmic transformations, which are recommended in cases of substantial positive and negative skew (Tabachnick & Fidell, 2007), failed to resolve these issues, so it was decided to continue with the analysis while acknowledging the loss in statistical power. The ANOVA showed a significant main effect for time, \(F(1,31) = 50.83, p < .001\), and a significant effect for the interaction of time and group \(F(1,31) = 25.7, p < .001\), but no effect for the interaction of time and matched pair, \(F(1,31) = 1.11, p < .391\).

Because the ordinal variable of matched pair represented vocabulary size only indirectly, the influence of this construct was further investigated by regressing the combined total scores (ID + CORRECT) on VLT scores. For the sample as a whole, vocabulary size could explain a small but significant amount of variation on pre-test performance, \(R^2 = .22, p < .001\). Thus, as the scatterplot in Figure 8 indicates, a larger vocabulary size did not provide participants with much advantage on the task at pre-test.
At post-test, regression analyses suggested VLT score was a better predictor for the SI group, $R^2 = .36, p < .001$, compared to the DP group, $R^2 = .11, p = .06$. However, the scatterplot showed two outliers in the SI group falling relatively near the regression line (Figure 9). When these data were removed from the analysis, the results changed considerably, $R^2 = .07, p = .176$.

Treatment effects, then, were not significantly related to vocabulary size, contrary to findings in the pilot study, which was based on a smaller sample ($N = 38$).
Figure 9 Relationship between total score (ID + CORRECT) and Vocabulary Levels Test score for both groups at post-test

Dictionary use measures

Lookups. Descriptive statistics and effects sizes for dictionary lookups are reported in Table 5. At pre-test, the averages were similar, with both groups conducting less than three lookups in learner dictionaries and less than four in non-learner dictionaries. Patterns diverged markedly at post-test, however, with the SI group’s learner dictionary lookups increasing to an average of 9.53 (nearly one per item on the task), while among the DP group the increase is less than half a lookup. Use of non-learner dictionaries decreased for both groups, but again the change was more substantial among the SI group, with such lookups nearly dropping off completely. For both categories of lookup, effect sizes for the within-group, pre-post differences would again be considered large for the SI group and small for the DP group, while between-group differences at post-test represent large effects.
Table 5 Descriptive statistics and effect sizes for dictionary lookups

<table>
<thead>
<tr>
<th></th>
<th>Learner dictionary lookups</th>
<th>Non-learner dictionary lookups</th>
<th>Total lookups (learner + non-learner)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SI</td>
<td>DP</td>
<td>SI</td>
</tr>
<tr>
<td></td>
<td>N = 32</td>
<td>N = 32</td>
<td>N = 32</td>
</tr>
<tr>
<td>Pre-test mean</td>
<td>2.63 (3.91)</td>
<td>2.75 (3.94)</td>
<td>3.97 (4.74)</td>
</tr>
<tr>
<td>Post-test mean</td>
<td>9.53 (4.6)</td>
<td>3.34 (4.71)</td>
<td>0.16 (0.52)</td>
</tr>
<tr>
<td>Mean change</td>
<td>6.9</td>
<td>0.59</td>
<td>-3.81</td>
</tr>
<tr>
<td>Pre-Post d</td>
<td>1.62</td>
<td>0.14</td>
<td>-1.13</td>
</tr>
<tr>
<td>Between groups d (Post)</td>
<td>1.33</td>
<td>1.01</td>
<td></td>
</tr>
</tbody>
</table>

Note: SI = Strategy Instruction, DP = Dictionary Practice; standard deviations in parentheses

To test for significance, the data were fitted to a generalized linear model using the logarithm as the link function and the Poisson distribution as the probability distribution, which is appropriate in cases where data represent counts and the number of observations is not fixed (Larson-Hall, 2010). Using the GENLIN procedure in SPSS, a Poisson regression was modeled to represent the mean rate at which participants performed lookups, with learner dictionary lookups as the response variable, group and time as predictors, and VLT scores as a covariate. Non-learner dictionary lookups were not included in the model as this would have violated the assumption of independence and because learner dictionary lookups was the variable of primary interest. A test of the full model against an intercept-only model was statistically significant, χ²(4) = 49.2, p < .001. Pairwise comparisons showed the mean difference in learner dictionary lookups for the SI group on the post-test was significant compared to the DP group on the post-test (mean difference = 6.0, 95% CI 2.44, 9.57, p < .001), and compared to the SI group on the pre-test (mean difference = 6.7, 95% CI 3.22, 10.17, p < .001).

Dictionary lookup points. Descriptive statistics and effect sizes for dictionary lookup points are reported in Table 6. Pre-test averages again showed similarities between groups, with the majority of points achieved without the aid of a reference, while learner dictionary lookups accounted for less than two points and non-learner dictionary lookups for about half a point, on average. On the post-test, however, the SI group averages showed a sixfold increase in points from learner dictionary lookups, with null lookup points reduced by about one third and non-learner dictionary lookup points dropping off completely. For the DP group, by contrast, learner dictionary lookup points increased and non-learner dictionary and null lookup points decreased...
by much smaller margins. Large effect sizes for learner dictionary lookup points were observed in the within-group, pre-post difference for the SI group, as well as for the between-group difference at post-test.

**Table 6** Descriptive statistics and effect sizes for lookup points

<table>
<thead>
<tr>
<th></th>
<th>Learner dictionary lookup points</th>
<th>Non-learner dictionary lookup points</th>
<th>Null lookup points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SI N = 32 DP N = 32</td>
<td>SI N = 32 DP N = 32</td>
<td>SI N = 32 DP N = 32</td>
</tr>
<tr>
<td>Pre-test</td>
<td>1.78 (3.63) 1.16 (2.69)</td>
<td>0.59 (1.39) 0.44 (1.04)</td>
<td>4.5 (3.54) 4.44 (2.72)</td>
</tr>
<tr>
<td>Post-test</td>
<td>11.06 (6.02) 2.94 (5.3)</td>
<td>0 (0) 0.41 (1.01)</td>
<td>2.81 (3.54) 3.88 (2.65)</td>
</tr>
<tr>
<td>Mean change</td>
<td>9.28 1.78</td>
<td>-0.59 -0.03</td>
<td>-1.69 -0.56</td>
</tr>
<tr>
<td>Pre-Post d</td>
<td>1.87 0.42</td>
<td>-0.6 -0.03</td>
<td>-0.48 -0.21</td>
</tr>
<tr>
<td>Between groups d (Post)</td>
<td>1.43</td>
<td>-0.57</td>
<td>-0.34</td>
</tr>
</tbody>
</table>

*Note:* maximum score for any type of lookup point = 20; SI = Strategy Instruction, DP = Dictionary Practice; standard deviations in parentheses

To test for significance, the mean change scores (see Table 6) were submitted to a two-sample Hotelling’s $T^2$ test using the R statistical software package. Hotelling’s $T^2$ is a generalization of Student’s $t$ for multivariate data, which is used when there are multiple dependent variables likely to be correlated with each other, and when the independent variable comprises only two groups (Tabachnick & Fidell, 2007). The test showed a significant difference between SI and DP score vectors, $T^2(3,60) = 12.98, p < .001$. As a post-hoc measure, the mean change scores for each category of lookup point were submitted to Welch two-sample $t$-tests, which returned a statistical result in the case of learner dictionary lookup points, $t(52.43) = 6.22, 95\% \text{ CI } 5.08, 9.92, p < .001$.

It is notable that, despite the difference in magnitude, post-test gains for both groups are attributable to increased successful exploitation of learner’s dictionaries. To determine whether the DP group’s gains were statistical, another two-sample Hotelling’s $T^2$ test was run using only this group’s learner dictionary lookup points, non-learner dictionary lookup points and null lookup points as dependent variables and time as the independent variable. No difference was found, $T^2(3,60) = 1.03, p = .385$. Regardless, these results merit closer scrutiny. Examination of the raw data showed that 78% of the DP group’s post-test learner dictionary lookup points were contributed by only six individuals, whose average total post-test score, $M = 15.7, SD = 3.1$, 

greatly exceeded that of the remainder of the group, \( M = 5.27, SD = 2.6 \). Likewise, these individuals accounted for most of the learner dictionary lookups at post-test, \( M = 10.8, SD = 1.7 \), versus \( M = 1.62, SD = 3.2 \) for the rest of the group. These participants appear to have acquired some degree of strategic pattern knowledge independently of the strategy instruction. This finding will be addressed further in the discussion.

Finally, to determine whether the treatments had differential effects on participants’ ability to identify errors versus their ability to correct them, the number of responses where errors were simply identified versus those where errors were both identified and corrected were tallied for both groups with reference to dictionary type, and proportions were calculated (Table 7). Overall, items where errors were both identified and corrected were more numerous than those where they were only identified, with the exception of pre-test items answered by the DP group using non-learners dictionaries.

**Table 7** Frequencies and proportions for response type by category of reference

<table>
<thead>
<tr>
<th></th>
<th>SI Learner dictionary</th>
<th>SI Non-learner dictionary</th>
<th>SI No use of references</th>
<th>DP Learner dictionary</th>
<th>DP Non-learner dictionary</th>
<th>DP No use of references</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre</td>
<td>Identified only</td>
<td>7</td>
<td>3</td>
<td>34</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Identified and corrected</td>
<td>28</td>
<td>8</td>
<td>55</td>
<td>16</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Proportion of identified only to identified and corrected</td>
<td>0.25</td>
<td>0.375</td>
<td>0.618</td>
<td>0.313</td>
<td>1.5</td>
</tr>
<tr>
<td>Post</td>
<td>Identified only</td>
<td>33</td>
<td>0</td>
<td>26</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Identified and corrected</td>
<td>161</td>
<td>0</td>
<td>32</td>
<td>43</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Proportion of identified only to identified and corrected</td>
<td>0.205</td>
<td>0</td>
<td>0.813</td>
<td>0.186</td>
<td>0.6</td>
</tr>
</tbody>
</table>

Note: *SI* = Strategy Instruction, *DP* = Dictionary Practice

These data show again that post-test gains for both groups are attributable to increased use of learner’s dictionaries and, moreover, that while more items overall were answered correctly by the SI group, the proportions in the learner dictionary category were similar, i.e. 0.205 for the SI group versus 0.186 for the DP group. To determine whether these proportions were statistically different, the data from the learner dictionary category were fitted to another generalized linear model using GENLIN in SPSS, with the logit as the link function and the binomial as the
distribution, which is appropriate for binary data with a fixed number of observations (Larson-Hall, 2010). With response type (i.e. identified only versus identified and corrected) as the dependent variable and group as the predictor, a pairwise comparison showed the difference to be non-statistical, \( p = .818 \). Thus, those DP group participants who discovered the strategy for themselves seemed as adept in both identifying and correcting errors as their SI group counterparts. This issue will also be addressed in the discussion.

**Research Question 2: Participant perceptions**

Cumulative data for the post-tutorial questionnaires are provided in the pie charts in Figure 10, which represent averaged ratings of the perceived levels of interest, challenge, and usefulness of the SI tutorials based on 465 responses. Analysis showed that, on the whole, the tutorials were viewed positively, with a majority of participants rating them either *interesting* or *very interesting* (57%), *challenging* or *very challenging* (54%), and *useful* or *extremely useful* (66%).
Perceptions of learning achieved through the materials were measured by the end-of-course evaluation, in which participants rated statements on a scale from 1 (strongly disagree) to 6 (strongly agree). Table 8 shows responses to statements regarding the SI materials that have been rank ordered according to the combined percentage of those who responded either agree or strongly agree. Among the 58 respondents, clear majorities felt the tutorials had improved their
understanding of dictionary types (81%) and their ability to exploit a dictionary for word-pattern information (78%), as well as convinced them to use learner dictionaries more often in the future (78%). Most also agreed they had learned things they didn’t know about word patterns (86%), had improved their ability to identify word patterns in sentences (84%), and would pay more attention to word patterns in the future while writing in English (84%). However, only a smaller majority said the tutorials had increased their interest in vocabulary learning (59%), while nearly half felt the tutorials had made English vocabulary seem more complicated and difficult (47%).

Table 8 Rank ordering of participants’ perceptions of learning outcomes as elicited by the end-of-course evaluation

<table>
<thead>
<tr>
<th>1. Because of the VVT Tutorials, ...</th>
<th>Participants responding agree or strongly agree (N = 58)</th>
</tr>
</thead>
<tbody>
<tr>
<td>... I am able to identify word patterns in sentences more easily.</td>
<td>49 (84%)</td>
</tr>
<tr>
<td>... I will pay more attention to word patterns when I write.</td>
<td>49 (84%)</td>
</tr>
<tr>
<td>... I have more confidence in my ability to use new vocabulary correctly.</td>
<td>47 (81%)</td>
</tr>
<tr>
<td>... I better understand the differences between types of dictionaries (e.g. monolingual, bilingual, native-speaker).</td>
<td>47 (81%)</td>
</tr>
<tr>
<td>... I better understand the connection between word usage and word meaning.</td>
<td>45 (78%)</td>
</tr>
<tr>
<td>... I am able to find word pattern information in the dictionary more efficiently.</td>
<td>45 (78%)</td>
</tr>
<tr>
<td>... I will use online monolingual learner's dictionaries (e.g. Longman) more often.</td>
<td>45 (78%)</td>
</tr>
<tr>
<td>... learning English vocabulary now seems more complicated and difficult.</td>
<td>27 (47%)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>2. The VVT Tutorials ...</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>... taught me things I didn't know about word patterns.</td>
<td>50 (86%)</td>
</tr>
<tr>
<td>... should be part of English 101B.</td>
<td>48 (83%)</td>
</tr>
<tr>
<td>... helped me use monolingual learner's dictionaries (e.g. Longman) more effectively.</td>
<td>46 (79%)</td>
</tr>
<tr>
<td>... taught me the importance of learning more about words I am already familiar with.</td>
<td>44 (76%)</td>
</tr>
<tr>
<td>... improved my ability to edit and correct my own writing.</td>
<td>41 (71%)</td>
</tr>
<tr>
<td>... were easy to use.</td>
<td>39 (67%)</td>
</tr>
<tr>
<td>... made me feel more interested in learning English vocabulary.</td>
<td>34 (59%)</td>
</tr>
</tbody>
</table>
In addition to Likert-scale items, the end-of-course evaluation included open-ended questions asking what participants liked most about the VVT course, how it could be improved, and why it should or should not be included in the ESL curriculum at Iowa State University. Responses to these questions were imported into Weft QDA, an open-source software tool for qualitative data analysis, where they were coded using an inductive approach to identify common themes. Themes commented on by five or more participants are reported below.

Regarding what they liked most, 10 of the 58 respondents (17%) mentioned the video presentations, nine (15%) mentioned the dictionary training, and five (9%) mentioned learning about word patterns. Example comments included:

*I liked the video part. It was easy to understand and learn.*

*I liked "how to use the dictionary correctly". I know more about how to use the dictionary. And the kind of words ... the meaning, usage.*

For suggested improvements, 10 participants (17%) mentioned technical issues and design limitations, for example:

*Use a user friendly interface, maybe. For example, attractive graphics and design it is better to divide homework into more pieces in order to avoid the operating mistakes on computer which can ruin the homework by the only attempt.*

Finally, with respect to curricular integration, three participants (5%) opposed the idea, citing the time required to complete the tutorials and perceived lack of instructional value. For example:

*I do not think VVT is very useful. We spent much time on VVT course. However I felt like I did not learn any useful stuff.*

By contrast, 50 participants (86%) were in favor, with perhaps the most articulate statement of support being:

*I strongly agree. Actually I even think there should be a class ONLY for vocabulary usage like the VVT Course. This kind of class is essential for English learners. It really helps students to use English properly. Without this kind of knowledge (how to use words correctly and as naturally as native speaker) English learners will end with broken English I believe.*
DISCUSSION

This study investigated the efficacy of a prototype online SI course designed to teach tertiary-level ESL learners the integrated knowledge and skills needed to address a specific vocabulary usage task involving pedagogical dictionaries. Following the intervention, SI group participants displayed significant improvements in their ability to select the correct dictionary for the task, and to exploit that dictionary in both identifying and correcting errors involving lexical patterns. In addition, most participants reported that they found the online materials interesting, useful and appropriately challenging, felt their vocabulary knowledge and dictionary skills had benefitted from them, and believed they should be integrated into the ESL curriculum. Together, the findings provide evidence of the potential for some aspects of L2 strategy instruction to be delivered in an automated, online form.

In contrast to SI research generally, the effect sizes documented in this study were large. While the instructional materials and technology-facilitated enhancements to the research design may partly account for these large effects, the sample’s generally poor performance on the pre-test must be taken into account, as should the fact that the SI treatment itself was comparatively long, focused on a single set of strategic behaviors, and in a strategy domain that has been shown to be a more fruitful focus for SI – all of which represent advantages relative to some previous intervention studies (see Plonsky, 2011). Regardless, the findings also demonstrate how technology can be used to increase the validity of SI research.

In addition, the study has addressed calls for integrating dictionary skills training with language awareness, in response to research showing that learners lack conceptual understanding of complementation, transitivity, and grammatical collocation. (Indeed, in the absence of strategy instruction, most participants had trouble identifying the majority of errors, and thus made changes to parts of the sentences that were correct, usually creating additional errors in the process.) While some L2 vocabulary experts have proposed addressing such gaps in learners’ lexical knowledge through intensive, item-focused instruction (e.g. Laufer, 2005), the present research shows how an approach based on strategies, metalinguistic concepts, and pedagogical references might prove an effective complement or alternative, particularly given that strategic gains were seen to accrue largely independently of vocabulary size.
The fact that a small cohort of DP group participants appeared to have developed strategic pattern knowledge through dictionary practice alone, with similar mastery of the subskills involved, calls to mind the “good language learner” in early L2 strategy research (Rubin, 1975). As noted by Tseng and Schmitt (2008), some learners are able to discover effective new strategies on their own through metacognitive monitoring and evaluation of strategy use. Assuming such learners will exist in any population, the question becomes, how to cater to their more advanced needs at the same time as the majority’s more fundamental ones?

Automated, online SI represents one possible answer. Because the instruction is self-paced, high ability learners can move quickly through those assignments that represent little challenge for them, while others can take the time they need to gain control over the basic behaviors. This appeared to be the case in the VVT course, as log files showed considerable variation in the amount of time it took learners to reach criterion level on the tutorials. Teachers, meanwhile, can use performance data from the online materials to identify those learners who will need extra help during class, versus those who are ready for more advanced consolidation or extension tasks.

Needless to say, there will be continued need for teachers in the SI process, as there are limits to what can be accomplished through technology alone. Indeed, the greatest limitation of this investigation is that it did not evaluate whether strategic gains transferred to participants’ L2 writing or persisted over time. It was assumed, in fact, that in the absence of complementary, teacher-led instruction to facilitate transfer and maintenance, such an effort was unwarranted. Future work should therefore investigate blended approaches that can take best advantage of the respective strengths of computers and teachers, with the work of the latter focused on refinement and contextualized applications of skills developed by the former.

Other limitations of the study suggest focuses for further research more specifically related to pattern grammar and dictionary use. The findings are likely to have been influenced by the predominance of Chinese-speaking learners in the sample. Although this was the population of interest, studies with greater representation of learners from other educational backgrounds, particularly those less skilled in syntactic processing as a result of years of exam-oriented English instruction, are warranted. The contribution of skill acquisition-oriented practice and feedback should also be more directly investigated. For example, eye-tracking could provide
finer-grained analyses of treatment effects on dictionary consultation behavior, as some screen-capture footage showed untrained participants apparently overlooking pattern information even when they managed to find the correct entry. Finally, case studies could help identify the specific factors that allow individuals to bootstrap their way into strategic pattern knowledge.

CONCLUSION

As noted, lifting some of the burden of L2 strategy instruction from teachers and timetables could shift the cost-benefit balance toward including it in syllabuses and language programs. While a great deal of attention has been devoted to the search for CALL applications that can provide more effective instructional opportunities for learners, few commensurate efforts have tried to harness the power of technology in enhancing learners’ ability to learn. There is clearly scope for increasing such efforts, assuming worthwhile strategic goals that lend themselves to automated, online SI can be identified. The possibilities are limited only by the imaginations of SI researchers, instructional designers and technology-minded teachers.

NOTES

1. While task is typically used in second language acquisition to mean a learning activity involving a primary focus on meaning and a lack of restrictions on learners’ linguistic choices (see Ellis, 2003), the much broader conceptualization adopted here is “a structured plan for the provision of opportunities for the refinement of knowledge and capabilities entailed in a new language and its use” (Breen, 1989, p. 187).

2. The SI group mean was 89.1 ($SD = 21.7$) and the DP group mean was 89.3 ($SD = 20.8$), out of a possible 120 on the Vocabulary Levels Test. The nonparametric Mann-Whitney U Test, used because of strongly negative skews in both distributions, showed no statistical difference, $U = 510.5$, $p = .984$.

3. The improvement in internal consistency makes sense assuming generally poor performance on the pre-test compared to greater variation in performance on the post-test (following treatment) among which the instrument could then differentiate.

4. In 12 cases, it was unclear to the researcher whether information from Chinese-English dictionaries had been used to achieve correct or partially correct responses, so a PhD student
in applied linguistics whose first language is Mandarin Chinese was asked to review the relevant video excerpts to make a determination.

5. The benchmarks in J. Cohen (1988) are .20 for a small effect, .50 for a medium effect, and .80 for a large effect, while Oswald and Plonsky (2010) propose .40, .70 and 1.00, respectively.

6. These two, low-performing SI participants had not completed most of the VVT assignments, as well as many assignments in the composition course, which one ended up dropping in the tenth week while the other failed. Although outliers in statistical terms – as was the DP group participant who scored 18 on the pre-test and 20 on the post-test – similar performances had been observed in previous rounds of piloting, and thus they were deemed representative of the target population. Furthermore, because removal of their data did not change the overall pattern of findings, their data was included in the major statistical tests in the study, i.e. the comparisons of means.

7. According to Larson-Hall (2010), Chi-Square is not appropriate in such cases because individuals are contributing more than one observation to particular cells in the contingency table, violating the assumption of independence.

8. This comment refers to the fact that participants were told certain tutorials had to be completed successfully in a single attempt, although in fact additional attempts were often allowed at the researcher’s discretion. This had become necessary after it was discovered that, in tasks involving multiple-choice questions, some participants were selecting responses at random to quickly identify the correct ones, then restarting the activity. Obviously, this is a bug that needs to be remedied.

REFERENCES


CHAPTER 4: FEEDBACK, MONITORING ACCURACY, AND SELF-REGULATION IN MOTIVATED LEARNING OF L2 VOCABULARY DEPTH-OF-KNOWLEDGE

A paper to be submitted to Language Learning

Jim Ranalli

ABSTRACT

L2 vocabulary depth-of-knowledge is required for contextually appropriate use, and yet the question of how learners acquire such knowledge has so far inspired relatively few empirical studies. Tseng and Schmitt’s (2008) structural model of motivated L2 vocabulary learning provides a useful framework for considering this question by 1) acknowledging the largely self-directed nature of L2 vocabulary learning, and 2) relating it to motivation and strategic behavior. This model, however, requires elaboration regarding the differential challenges posed by size-versus depth-related learning, as well as the conditions under which strategic mastery can be developed to facilitate achievement. This study reviewed research regarding two key forms of feedback presumed necessary for self-directed vocabulary learning: internal feedback via monitoring of strategy use, and external feedback via dictionary consultation. Sixty-four students in a tertiary-level ESL writing course were matched for vocabulary size and then randomly assigned to a naturalistic “discovery” condition or an explicit training condition, both administered via a learning management system. Data included actual scores and the participants’ post-dicted scores on size- and depth-related measures, as well as dictionary-use data gathered via screen-capture software. Significant miscalibrations were observed in monitoring of performance on the depth-related task in the absence of explicit instruction.

INTRODUCTION

It is not uncommon to hear learners of English lament the gap between knowing what a word means and knowing how to use it, with the former generally considered much easier than the latter. Among L2 researchers, this distinction is often characterized as vocabulary size versus vocabulary depth. Size boils down to connecting word forms and meanings, and size-related research investigates such topics as lexical coverage, i.e. the numbers of words one needs to
know to perform certain tasks in the L2 (e.g. Hu & Nation, 2000; Laufer & Ravenhorst-Kalovski, 2010; Staehr, 2009). Depth has been conceptualized in various ways but perhaps most relevant for L2 pedagogy is the notion of comprehensive word knowledge (Read, 2004), which goes beyond meaning to include other lexical features such as collocation and register constraints, and syntactic, derivational and morphological behavior. Depth-of-knowledge features can be viewed as requirements for contextually appropriate use, hence the increasing interest in depth in line with contemporary demands for instruction focusing on communicative rather than formal competence.

In addition to the issues of how to assess depth-of-knowledge (e.g. Nurweni & Read, 1999; Wesche & Paribakht, 1996) and how depth-of-knowledge is organized in the mental lexicon (e.g. Henriksen, 1999; Meara, 2009), researchers have turned their attention to how such knowledge develops in L2 learners. There is general agreement that acquisition is incremental; a single exposure will never result in knowing all one needs to know about a word. Beyond this, however, there is a lack of consensus. A longstanding and still influential view, particularly among teachers, is that vocabulary is best learned implicitly, incidentally and in context, for example, through extensive reading. The idea is that, by means of repeated, varied exposures to a particular item, learners will gradually accumulate a comprehensive understanding of its forms, meanings and uses. This view, based on L1 studies and stated most explicitly in Sternberg (1987), has been referred to as the default hypothesis because it is “justified in negative terms” (Laufer, 2010, p. 15); that is, the number of words learned is too great to be accounted for by instruction alone. However, L2 research into noticing and attention suggests that relying on incidental learning will, at the very least, be inefficient because of the low frequency and salience of many lexical items in input (Hulstijn & de Graaff, 1994; Schmidt, 1994). Laufer (2005) surveyed empirical research showing that an intentional and even decontextualized focus on individual words is a beneficial and necessary component of instruction; while a recent study by Elgort (2011), involving a lexical decision task and various priming measures, demonstrated that deliberate vocabulary learning can result in both explicit and implicit knowledge. This is not to suggest that incidental learning through exposure is unimportant but that it must be supplemented by word-focused instruction, a view endorsed by many experts on L2 vocabulary pedagogy (e.g. Lewis, 1997; Nation, 2007; Schmitt, 2008).
It is clear, however, that the burden of word-focused instruction cannot be shouldered by teachers alone. Using the same mathematical logic underlying the default hypothesis, we can see there are simply too many words, and too much to know about them, to be addressed in the classroom, particularly when we consider the many other things teachers are expected to cover and the idiosyncratic lexical needs, beyond a certain frequency threshold, of individual learners. Some, if not most, of the responsibility for L2 vocabulary depth-of-knowledge must therefore fall to learners themselves. But how prepared are they to go about this learning independently? What obstacles do they face, and what support do they need?

A useful framework for considering these questions is a process model of L2 vocabulary learning proposed by Tseng and Schmitt (2008). In addition to the fact that it attempts a broad and integrative view in a field characterized by narrow focuses, this model is noteworthy for assuming a central role for learners in self-directing their own acquisition of L2 vocabulary, both in terms of size and depth. It also takes into account key individual difference factors that other L2 research has shown to greatly influence achievement: namely, motivation and strategic behavior. Furthermore, it makes a key distinction between quantity- and quality-oriented strategy use, reflecting empirical findings that it is not how many or how often learning strategies are employed, but the skill and situational appropriateness with which this occurs, that determines success (e.g. Ehrman, Leaver, & Oxford, 2003; Gardner, Tremblay, & Masgoret, 1997; Gu, 2003; Politzer & McGroarty, 1985; Vann & Abraham, 1990). This model, however, is provisional and requires further elaboration, particularly in terms of explaining: 1) the differential challenges that might be posed by size-related versus depth-related learning; and 2) the manner in which learners can progress from quantity- to quality-oriented strategy use. This paper seeks to demonstrate that, while Tseng and Schmitt’s model is an important and long-overdue first step toward a comprehensive theory of L2 vocabulary learning, it requires additional components to have explanatory and predictive power regarding these two issues. A review of research was conducted regarding two key forms of feedback presumed necessary for self-directed vocabulary learning: internal feedback through monitoring and evaluation of strategy use, and external feedback through consultation of pedagogical dictionaries. Based on this review, hypotheses were generated and tested in a hybridized experiment combining elements of classroom- and laboratory-based research. Before describing the experiment and its findings, further discussion of Tseng and Schmitt’s model is required.
Tseng and Schmitt’s model of motivated L2 vocabulary learning

Based on structural equation modeling (SEM), this model depicts vocabulary learning as a cyclical process of six latent variables (Figure 1). The process begins with learners’ Initial Appraisal of Vocabulary Learning Experience (IAVLE) in terms of the value, interest, effort or desire perceived in a particular learning task. If, on the basis of this appraisal, sufficient motivation is generated, then learners set goals, form intentions, and initiate effort, at which point the cycle moves to the next variable, Self-regulating Capacity in Vocabulary Learning (SRCvoc). SRCvoc constitutes volition that can sustain effort in the face of challenges, such as flagging energy or competing interests. SRCvoc has a direct influence on the following component, Strategic Vocabulary Learning Involvement (SVLI), which represents the quantity dimension of learning strategy use; that is, volition can help determine how many, and how frequently, strategies are employed. Interestingly, Tseng and Schmitt found insufficient statistical evidence to support a direct path between SVLI and learning outcomes, as represented by the variable Vocabulary Knowledge (VOCkno). Rather, the relationship between SVLI and VOCkno was found to be mediated by an additional strategic component, Mastery of Vocabulary Learning Tactics (MVLT), which represents the quality dimension of strategy use. In other words, success in vocabulary learning will not simply result from the application of strategies. It matters how well and how appropriately learning strategies are applied. Finally, the variable Post-appraisal of Vocabulary Learning Tactics (PAVLT) closes the loop and sets the stage for subsequent learning experiences. In PAVLT, learners reflect on their vocabulary learning and make causal attributions, which “exert a critical influence on subsequent expectancy for success, self-efficacy belief, achievement behaviors, and emotional responses” (2008, p. 368) as initiated in later instances of IAVLE.
Figure 1 Tseng and Schmitt’s structural model of motivated L2 vocabulary learning (2008, p. 381), as revised following SEM analysis. **Latent variables:** IAVLE = initial appraisal of vocabulary learning experience; SRCvoc = self-regulating capacity in vocabulary learning; SVLI = strategic vocabulary learning involvement; MVLT = mastery of vocabulary learning tactics; VOCkno = vocabulary knowledge; PAVLT = post-appraisal of vocabulary learning tactics. * Standardized path coefficients, $p < .05$.

It can be seen that, in addition to the heuristics already mentioned, this model allows connections to be made between L2 vocabulary learning and motivational constructs that are integral to contemporary educational psychology and many theories of self-regulation, such as self-efficacy, goal orientation, outcome expectations and task interest/value (Zimmerman, 2000, 2006). In this sense, self-regulation refers to the “self-generated thoughts, feelings, and actions that are planned and cyclically adapted to the attainment of personal goals” (Zimmerman, 2000, p. 14). It is important to distinguish this sense, which conforms more closely to the general usage of the term in educational psychology, from the sense embodied in the latent variable SRCvoc, which, as noted, is a volitional construct. This latter, narrower sense is associated with the work of Kuhl (Kuhl, 2000; Kuhl & Goschke, 1994), who pioneered a modern psychological theory of volition,
which in turn has strongly influenced the work of Dörnyei (Dörnyei, 1994, 2001, 2003; Dörnyei & Otto, 1998) on motivation for L2 learning. Dörnyei has in fact collaborated with Tseng and Schmitt to import the construct of self-regulating capacity to L2 studies (Tseng, Dörnyei, & Schmitt, 2006) and Dörnyei’s process model of L2 motivation provides theoretical underpinnings to the present model.¹

Despite this distinction, the present model might be assumed to function similarly to contemporary models of self-regulated learning (SRL), such as Boekaerts and Niemivirta (2000), Winne and Hadwin (1998), and Zimmerman (2000), which also integrate motivational, strategic and metacognitive components and utilize recursive processes – that is, output from one part becoming input for another – to drive certain model dynamics. While recursiveness is indeed a shared feature, the components of Tseng and Schmitt’s model that recursiveness can explain are a point of difference, as will be shown.

In this model, the culminating variable of PAVLT is based on three, Likert scale-based indicators: “skilled feeling […], helpless feeling […], and satisfied feeling of using vocabulary learning tactics” (2008, p. 387). Learners in this stage should “know how to adaptively attribute their vocabulary learning success or failure to its proper cause” (2008, p. 389), as this is predicted to influence subsequent iterations of the cycle. However, no paths are specified between the appraisal variables and those representing strategic behavior. In fact, an earlier version of the model hypothesized causal relationships between IAVLE and both SVLI and MVLT (Figure 2) but the SEM analysis found IAVLE to have virtually no predictive power over either. This means evaluations of success or failure, or attributions of success or failure to particular causes, are not seen as directly contributing to subsequent modifications of strategic behavior. Nor is any reciprocal relationship hypothesized between SVLI and MVLT. Rather, it is presumed that “the repeated appropriate usage of tactics (as governed by SVLI) eventually leads to mastery over those tactics (MVLT)” (2008, p. 366).
Thus it can be seen that, although the model goes some way toward explaining the motivational dynamics of L2 vocabulary learning, it does not adequately capture the behavioral dynamics; it demonstrates a need for strategic expertise but does not explain how such expertise can develop. This is problematic, especially given the authors’ claims that, while teachers have an important role to play in preparing learners for strategic mastery, once it is achieved in some measure, learners themselves will “know how to discover vocabulary learning tactics from a variety of learning activities” (2008, p. 389). This sort of strategic “bootstrapping,” while possible, presents considerable challenges to learners in terms of obtaining sufficient practice with appropriate feedback, remembering how learning took place, and reasoning about the factors that affected it (Winne, 1997). This argues a need for greater elaboration on how expertise can develop from inexpert strategy use. Although not in the model itself, Tseng and Schmitt do take up the matter in the discussion, asserting that the shift from SVLI to MVLT is facilitated by “the successful
functioning of metacognitive regulation and involvement, such as implementing conditional knowledge, planning, monitoring, and evaluation” (2008, p. 384). But this is as far as it goes. Monitoring has been recognized as a prototypical feature of L2 learning strategy use (A. D. Cohen, 2007), but there is a tendency to treat monitoring and evaluation as primarily issues of self-understanding and metacognitive control (Abraham & Vann, 1996), which has been criticized by some applied linguists for neglecting the role of non-strategic forms of knowledge, such as task knowledge (Wenden, 1995) or language knowledge (Abraham & Vann, 1996). Depending on the task, monitoring and evaluation might require reference to standards informed by non-strategic forms of knowledge to be determinative of success. In line with this view, it is argued here that, in the same way that quality matters in the application of strategies, it also matters in learners’ self-assessments of the learning that results from strategy use. The quality of both the inputs to, and the products of, monitoring and evaluation – that is, the quality of feedback – will determine the extent to which appropriate adjustments can be made to future strategy use so as to bring performance closer in line with goals and standards. As noted by Locke and Latham, whose work on goal setting has been incorporated into many contemporary motivational theories,

… people need summary feedback that reveals progress in relation to their goals. If they do not know how they are doing, it is difficult or impossible for them to adjust the level or direction of their effort or to adjust their performance strategies to match what the goal require (2002, p. 708). Thus, a model of L2 vocabulary learning that posits a central role for learner self-direction and a need for strategic mastery, and which leaves open the possibility of such mastery arising through self-discovery, must contain some explicit, recursive mechanism based on behavioral feedback to have explanatory and predictive power. This paper attempts to demonstrate the need for quality-oriented feedback with reference to self-directed learning of L2 vocabulary depth-of-knowledge. Two sources of feedback are considered: one internal, in the form of evaluations arising from metacognitive monitoring of one’s learning or performance; and the other external, in the form of pedagogical dictionary use. Literature in these two areas will now be reviewed.

**Internal feedback from metacognitive monitoring**

In second language studies, feedback is generally conceptualized as external. In L2 writing research, for example, it is seen originating with instructors, peers, or computers (Hyland &
Hyland, 2006). In line with the premise of Tseng and Schmitt’s model, however, it is assumed that a great deal of vocabulary learning and use will occur in contexts that do not benefit from external sources, and which thus will compel learners to rely on feedback generated internally, in the form of judgments based on metacognitive monitoring. These judgments may be broadly characterized as reflecting size-related concerns, e.g. *Am I sure of the meaning of this word?* or depth-related concerns, e.g. *Am I using this word correctly?* The accuracy of the responses that learners generate to these self-posed questions will help determine what modifications, if any, they make to related strategic behavior and what learning and language use may result.

Psychologists have studied the accuracy of self-estimates of abilities for nearly a century (Ackerman & Wolman, 2007). More recent studies foregrounding the role of metacognition have operationalized such judgments in various ways, often according to the time at which they occur. For example, judgments of learning (JOLs) are subsequent recollections of recently studied information, whereas ease of learning judgments (EOLs) refer to those made prior to study about the perceived relative ease of learning information (Schraw, 2009). Confidence judgments demonstrate one’s confidence in the accuracy of learning or performance and are measured online (while a test is ongoing) or retrospectively (following a test). The choice of measure will depend on the purpose of the research; *local* confidence measures, which focus on metacognitive accuracy with respect to individual test items, are appropriate when comparing performance across domains or test sections; while *global* confidence judgments are usually assessed retrospectively to investigate the effects of an intervention (Nietfeld, Cao, & Jason, 2005). The term *calibration* is often used to describe the relationship between confidence and actual performance, and calibration measures, such as correlation, are used to show how aware individuals are of what they do and do not know. A related measure is *bias*, which can be operationalized using a simple linear model, \( C = c - p \), where \( C \) is calibration, \( c \) is confidence (or self-assessed score) and \( p \) is performance (or actual score). A positive calibration score indicates positive bias or overconfidence, a negative score indicates negative bias or underconfidence, and a score of zero indicates perfect calibration (Phakiti, 2005).

Research shows that metacognitive monitoring is difficult and miscalibration is the rule rather than the exception. Numerous factors have been found to moderate calibration, including individual level of confidence, item difficulty, general versus domain-specific knowledge, learners’ inferential processes, type of feedback, and the time and type of measure with which
calibration is assessed (Stone, 2000). One fairly consistent finding is a hard-easy effect (Gigerenzer, Hoffrage, & Kleinbölting, 1991), in which overconfidence tends to increase with item difficulty. Another is that greater knowledge tends to increase calibration (Björkman, 1992; Nietfeld & Schraw, 2002), while among subjects who lack knowledge or skills, miscalibrations tend to be considerable in the direction of overestimation (Lichtenstein & Fischhoff, 1977). Dunning and his associates (Dunning, 2005; Ehrlinger, Johnson, Banner, Dunning, & Kruger, 2008; Kruger & Dunning, 1999) have shown that, across a variety of social and intellectual skill areas, overestimation increases as actual ability decreases, such that the poorest performers are usually the least calibrated, tending to think they have done much better than they actually have. In cases where subjects possess expertise, however, calibration is higher, with an interesting corollary being that top performers are often not the most accurate self-assessors. Rather, they tend to slightly underestimate themselves, perhaps because of a “false-consensus effect” (Kruger & Dunning, 1999; L. Ross, Greene, & House, 1977) in which subjects assume that, because a task is relatively easy for them, it must likewise be easy for most others, so they see themselves performing more closely to an imagined average.

While perhaps disconcerting for psychologists, miscalibration can serve educational purposes. Overconfidence may motivate learners to more readily undertake challenging but attainable tasks (Bandura, 1986) or to persist in learning (Stone, 2000), while underconfidence may prompt overlearning, leading to better retention (Dunlosky & Rawson, 2012). And yet, consistent overestimation suggests that self-regulation is being undermined because “update information from the mismatch between standards and products is not altering conditions as needed” (Greene & Azevedo, 2007, p. 349).

In one of the few L2 studies of calibration, Phakiti (2005) had 295 EFL learners at a Thai university self-assess their performance on the English Placement Test (EPT) created at the University of Michigan. This test consists of sections for listening, vocabulary, grammar, and reading. Phakiti was interested in calibration across these domains, so his analysis focused on local measures, which he called “single-case confidence,” although he did also include a global measure referred to as “relative-frequency confidence.” Single-case confidence scores, assessed using a confidence measure after every item on the test, indicated miscalibration throughout the EPT, with bias taking the form of overconfidence (Table 1). Relatively better calibration was observed on the vocabulary section, with an overall correlation of $r = .502$, $R^2 = .25$, $p < .01$, and
a calibration score of +5.84%. (Calibration was highest for reading and lowest for grammar.)

Implications for the present study are limited, however, because the vocabulary section of the EPT combines items requiring more size-oriented knowledge with those requiring more depth-oriented knowledge.

**Table 2** Single-case confidence correlation coefficients and calibration scores from Phakiti (2005)

<table>
<thead>
<tr>
<th>EPT section</th>
<th>Performance/confidence correlation coefficients</th>
<th>Calibration score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Listening</td>
<td>.383*</td>
<td>R² = .15</td>
</tr>
<tr>
<td>(k = 20)</td>
<td></td>
<td>+8.94%</td>
</tr>
<tr>
<td>Grammar</td>
<td>.407*</td>
<td>R² = .17</td>
</tr>
<tr>
<td>(k = 30)</td>
<td></td>
<td>+9%</td>
</tr>
<tr>
<td>Vocabulary</td>
<td>.502*</td>
<td>R² = .25</td>
</tr>
<tr>
<td>(k = 30)</td>
<td></td>
<td>+5.84%</td>
</tr>
<tr>
<td>Reading</td>
<td>.435*</td>
<td>R² = .21</td>
</tr>
<tr>
<td>(k = 20)</td>
<td></td>
<td>+2.68%</td>
</tr>
<tr>
<td>Overall test</td>
<td>.540*</td>
<td>R² = .29</td>
</tr>
<tr>
<td>(k = 100)</td>
<td></td>
<td>+6.4%</td>
</tr>
</tbody>
</table>

N = 295, *p < 0.01

In terms of relative-frequency confidence for the EPT as a whole, Phakiti reported a near-perfect calibration score of -0.08%, although this aggregate figure is misleading. As the plot of the actual data in Figure 3 indicates, such high calibration must have resulted from a balancing out of roughly equal instances of under- and overestimation among the sample. (Correlation was not reported for this data.) This is why some researchers of calibration (Pieschl, 2009; Weingardt, Leonesio, & Loftus, 1994) emphasize the need to use calibration graphs, such as Figure 3, in conjunction with scores and correlation coefficients when reporting results.

![Figure 3 Calibration on the English Placement Test (N = 295) from Phakiti (2005, p. 43)
While primarily concerned with implications for language testing, Phakiti identified several needs for L2 classroom research, including metacognitive training to help learners more accurately judge the likelihood of success before making high-stakes decisions about language proficiency exams or courses of study. It should be noted that the context of this research was a test to facilitate placement decisions preceding formal, tertiary-level English instruction, in which learners’ metacognitive self-assessments presumably would be supported with external feedback from instructors, summative assessments, etc. Once again, to the degree that L2 vocabulary learning takes place outside classrooms, such external forms of feedback will often be unavailable. However, a powerful, potential source of external feedback for self-directed vocabulary learning does exist in the form of pedagogical dictionaries, which will now be discussed.

**External feedback through pedagogical dictionary consultation**

In terms of reference tools, L2 learners of English have never had more support for learning and using vocabulary depth-of-knowledge. Contemporary pedagogical learners dictionaries written especially for them seek to provide not only comprehensive coverage of a word’s meanings but also its morphological, syntactic and derivational behavior, collocational, stylistic and register constraints, as well as frequency information, special usage tips and naturalistic examples, all supported by corpus data – advantages not generally shared by learners of other languages (Frankenberg-Garcia, 2011). Moreover, the major ELT publishers have made versions of these dictionaries freely available on the web where for many purposes they can be more quickly and conveniently consulted than the paper-based versions (Lew, 2011a). In addition, specialized dictionaries addressing specific dimensions of lexical knowledge, such as collocation, are now available (Rizo-Rodríguez, 2004).

And yet, learners’ skills in using dictionaries appear not to have kept pace with these developments. Research indicates pedagogical dictionaries of English remain generally underexploited and often misused by their target audience. Learners appear to lack mental models with which to differentiate among dictionary types (e.g. pedagogical versus those designed for native speakers) and the different purposes for which these are appropriate (Frankenberg-Garcia, 2011). Even higher-proficiency learners tend to prefer bilingual dictionaries and to use them mostly for referencing L1-L2 translation equivalents (Atkins &
Varantola, 1998; Frankenberg-Garcia, 2005). Learners have trouble distinguishing correctly among senses of polysemous words (Chan, 2011; Nesi & Haill, 2002) and tend to focus only on the initial or final entries (Nesi & Tan, 2011). They often ignore explicit usage information presented via symbols or codes and rely instead on dictionary examples for guidance (Bogaards & van der Kloot, 2002; Chan, 2012; Dziemianko, 2006), which can lead them astray when, for example, they use a word’s semantic associations to deduce facts about its syntactic behavior.

Problems in consultation can partly be attributed to the way particular dictionaries organize and present information (Dziemianko, 2011; Tono, 2011), as well as limited attentional resources that learners may be unwilling to expend in the midst of other language tasks; monolingual resources in particular impose additional processing burdens that can be disincentives to their use (Neubach & Cohen, 1988; Tono, 2001). But lexicographical research also suggests learners lack basic understanding of some depth-of-knowledge features, thus undercutting the basis for consultation and exploitation. Nesi surmised that the 40 adult ESL learners in her UK-based study were unable to exploit usage information in examples because “the concepts of transitivity and grammatical collocation were very poorly understood” (2000, p. 115). Laufer found that the 95 Israeli high school learners of English in her study failed to look up half of the collocations in her elicitation task either because they could not identify them in dictionaries or because they thought they knew them when they did not, leading the author to assert that many learners are “probably not aware of the nature of collocations” (Laufer, 2011, p. 44). Frankenberg-Garcia (2011) had 211 learners of English at a Portuguese university complete a task eliciting knowledge of lexical and prepositional collocations, connotation, irregular past tense forms, and spelling. For each item, the participants were told to select the type of reference they would consult if existing knowledge failed them. Of the 1215 responses in which they claimed to know the answer, they were wrong more than two-thirds of the time (833 responses). Moreover, their selections of references for dealing with collocation and connotation were “disastrous” (Frankenberg-Garcia, 2011, p. 119), insofar as most passed over monolingual pedagogical dictionaries, the only type among those offered that reliably provides such information. (Interestingly, 65% of the sample selected a grammar book as the best reference for prepositional collocations, implying an inability to differentiate between these lexical features and prepositions of time and place.) Such findings align with those of researchers who study formulaic language
and who assert a general insensitivity to collocational relationships among learners of English (Flowerdew, 2006; Granger, 1998; Wray, 2002).

In a recent study that encapsulates many of these issues, Chan (2012) investigated dictionary use among 31 advanced ESL learners at a Hong Kong university. The participants were asked to make grammaticality judgments about 10 sentences – nine of which included a lexical error involving either complementation, transitivity, prepositional collocation, or countability – and to make corrections as needed. They were provided with dictionary packets containing excerpts from the Cambridge Advanced Learner’s Dictionary, 3rd edition to consult while completing the task. As in previous research, many of Chan’s participants ignored explicit usage information and focused instead on examples, which often led them to make inappropriate generalizations and apply erroneous corrections. Thinkaloud data indicated that many appeared not to regard transitivity as important for grammaticality, and in a post-task focus group some acknowledged never reading grammatical information in dictionaries. Chan concluded that if these participants experienced such problems “it is doubtful whether learners at less advanced levels would be able to assimilate the usage information … and apply it” (2012, p. 19). Among the pedagogical implications she drew was the need to integrate dictionary skills instruction with “grammar training” (2012, p. 21), which echoes calls in other recent lexicographical research for increasing learners’ language awareness of depth-of-knowledge features (Chan, 2011; Frankenberg-Garcia, 2011; Laufer, 2011; Lew, 2011b).

**Literature summary and research questions**

To sum up, achievement in L2 vocabulary learning in Tseng and Schmitt’s model is contingent upon the development of strategic mastery, which in turn is contingent upon monitoring and evaluation of strategy use. Existing research suggests important forms of internal and external feedback that could provide the basis for such monitoring and evaluation may be inadequate when it comes to self-directed learning of vocabulary depth-of-knowledge features, especially those involving word combinability (i.e. collocation, transitivity, and complementation patterns). This study sought empirical evidence regarding these matters by comparing learners’ actual performance with their confidence in performance on size- and depth-related measures, and by repeating administration of the depth-related measure following an intervention. In this intervention, a naturalistic “discovery” condition, in which participants were presented with a
need for, and repeated opportunities to notice, aspects of word combinability via a recurring
dictionary-use task, was contrasted with an explicit instruction condition in which participants
received direct teaching in aspects of word combinability and related dictionary use. Three
research questions guided the investigation:

1. How is monitoring accuracy affected by type of vocabulary measure (size-related versus
depth-related)?
2. How does treatment condition (discovery or explicit instruction) affect monitoring
accuracy, as well as actual performance, on a depth-related measure?
3. How does treatment condition (discovery or explicit instruction) affect use of
pedagogical dictionaries on a depth-related measure?

It was hypothesized that learners would demonstrate less accurate monitoring and greater
overconfidence on the depth-related measure than the size-related measure, and that, in the
absence of explicit instruction, they would not show significant improvements in monitoring
accuracy or actual ability following the treatment; neither would they exhibit any significant
improvements in their task-related dictionary use.

METHODS

Context

The context of the current study is a large public research university in the Midwestern U.S. Like
many such universities, its international student population at present consists mostly of students
from mainland China (54%, according to 2011 figures). To enroll, they and all other
international students whose L1 is not English must have achieved an iBT TOEFL score of 71
(or the equivalent), which represents the minimum proficiency deemed necessary for tertiary-
level studies in the English medium. Upon entering, they take an English placement test to
determine whether they will need supplementary ESL instruction in areas such as writing,
reading or listening. The present study was based in one such course, an intermediate-level ESL
composition course that focuses on sentence-level mechanics, paragraph structure and discourse
patterns used in academic writing. The vocabulary focus of this course consists of a recurring
dictionary practice activity in which students look up usage information in references of their
choosing (described below). Other than this, there is no formal vocabulary component in the syllabus, nor is there any vocabulary course in the university’s ESL curriculum as a whole. Thus, the present context can be characterized as one in which it is assumed that learners have acquired sufficient knowledge and skills to self-direct most of their own learning of English vocabulary. Another important contextual feature is a strong reliance on technology in the composition course in question; one class per week is held in a computer lab and most assignments are distributed and collected via a learning management system (LMS). These technologies were exploited in several ways to facilitate the research design, as will be described.

Participants

The sample \(N = 64\) was drawn from three sections of the composition course in question and consisted mostly of students who spoke Mandarin Chinese as their L1 (76.5%). The participants were organized into matched pairs (Rencher, 2002) based on an administration of the Vocabulary Levels Test (described below), because a moderate but significant correlation had been observed between vocabulary size and treatment effects in a pilot study, and because paired observations were required for a separate but related investigation that was part of the same larger research project. After matching, members of each pair were randomly assigned to the discovery group or the explicit instruction (hereafter “instruction”) group using the RAND function in Microsoft Excel. Subsequent analysis showed the two groups did not differ significantly in terms of vocabulary size\(^4\). The groups were also similar in gender balance, variety of L1s, average time spent in U.S., age, and TOEFL score, as shown in Table 2.

<table>
<thead>
<tr>
<th>Table 3 Biodata for participants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Discovery</strong></td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>Female 12</td>
</tr>
<tr>
<td>Male 20</td>
</tr>
<tr>
<td>Chinese 23</td>
</tr>
<tr>
<td>Arabic 4</td>
</tr>
<tr>
<td>Korean 3</td>
</tr>
<tr>
<td>Thai 1</td>
</tr>
<tr>
<td>Malay 1</td>
</tr>
<tr>
<td>Time in U.S. (months)</td>
</tr>
<tr>
<td>5.7 (8.4)</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>20.1 (2.7)</td>
</tr>
<tr>
<td>TOEFL score (iBT)</td>
</tr>
<tr>
<td>81.2 (9.8)</td>
</tr>
<tr>
<td><strong>Instruction</strong></td>
</tr>
<tr>
<td>Sex</td>
</tr>
<tr>
<td>Female 14</td>
</tr>
<tr>
<td>Male 18</td>
</tr>
<tr>
<td>Chinese 26</td>
</tr>
<tr>
<td>Malay 2</td>
</tr>
<tr>
<td>Hindi 1</td>
</tr>
<tr>
<td>Spanish 1</td>
</tr>
<tr>
<td>[a Bantu language] 1*</td>
</tr>
<tr>
<td>Korean 1</td>
</tr>
<tr>
<td>Time in U.S. (months)</td>
</tr>
<tr>
<td>7.5 (14)</td>
</tr>
<tr>
<td>Age</td>
</tr>
<tr>
<td>21.6 (5.9)</td>
</tr>
<tr>
<td>TOEFL score (iBT)</td>
</tr>
<tr>
<td>76.7 (10.2)</td>
</tr>
</tbody>
</table>

**Data sources**

*Vocabulary Levels Test (VLT)*

Vocabulary size was measured using the Vocabulary Levels Test (Nation, 1990). A more recent version (Schmitt, Schmitt, & Clapham, 2001) was selected to reduce the chances that participants had previous experience of particular test items. The test was adapted for online administration in the LMS used in the study. The adapted version consisted of four frequency levels: the 2000- and 3000-word levels, which assess knowledge of high-frequency words; the 5000-word level, which represents a boundary between low- and high frequency words; and an academic word level, which tests knowledge of the Academic Word List (Coxhead, 1998). Each level contains a selection of 60 words, 30 of which are actually tested. The words are presented in six-item lists positioned next to three synonyms or glosses. Test-takers select a word from the list to best match each synonym/gloss by inputting the corresponding number into the appropriate gap. One point is awarded for each correct response, and the scores are totaled for a maximum possible score of 30 on each level and 120 on the test as a whole. Reliability for the measure is reported in Table 3.

**Table 4 Reliability for the online version of the Vocabulary Levels Test used in the study**

<table>
<thead>
<tr>
<th>Level</th>
<th>Number of items</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>30</td>
<td>.857</td>
</tr>
<tr>
<td>3000</td>
<td>30</td>
<td>.878</td>
</tr>
<tr>
<td>5000</td>
<td>30</td>
<td>.866</td>
</tr>
<tr>
<td>AWL</td>
<td>30</td>
<td>.923</td>
</tr>
<tr>
<td>Test as a whole</td>
<td>120</td>
<td>.964</td>
</tr>
</tbody>
</table>

*Note: Number of test-takers: 64*

Participants’ confidence scores on the VLT were collected via survey-type questions immediately following each word-level page in the assessment. Upon completion of a level, the participants moved on to the self-assessment page where they were asked to post-dict their score on the preceding assessment page from a point range of 0-30. They could return to the assessment page if they wished to better gauge their performance. Item-by-item (i.e. local) confidence measures were not employed because they were more likely to have introduced a method effect and because global measures were deemed more appropriate to the research aims.
Nevertheless, global measures for each frequency level were collected to investigate the presence of a hard-easy effect in the participants’ calibrations on a measure of vocabulary size.

*Pattern Identification and Correction Task (PICT)*

The depth-related measure created for the study was used as both a pre-test and post-test. It resembled that used in Chan (2012) in that it included 10 sentence-based items, each incorporating a word specially chosen to represent a challenging area of usage, i.e. grammatical collocation, complementation, or transitivity. These different areas were subsumed into a single, pedagogically friendly category using the system of description known as *pattern grammar*, which was developed by Hunston and colleagues (Hunston & Francis, 1998, 2000; Hunston, Francis, & Manning, 1997) as part of the COBUILD project. Pattern grammar combines these more familiar lexical features under the umbrella term *pattern* and emphasizes the systematic relationships between patterning and meaning. Hunston and Francis (1998, 2000) assert that all classes of words can be described in terms of their patterns, but for pedagogical and lexicographical purposes they have focused their corpus research on the patterns of the more frequent verbs, nouns and adjectives of English.

The pattern identification and correction task (PICT) differed from Chan’s assessment in that grammaticality judgment tasks preceding the dictionary consultation/sentence correction tasks were not included. Participants were informed that each sentence contained one lexical error to be identified and corrected. And whereas Chan (2012) provided paper packets of reference material excerpted from a single pedagogical dictionary, the PICT more closely resembled the assessment in Frankenberg-Garcia (2011) by giving participants a range of reference types from which to choose. In the present study, these references were web-based, and participants had authentic access to them while completing the assessment, which was administered online through the project LMS. This facilitated analysis of both the ability to select an appropriate reference for one’s purpose, and the ability to exploit the selected reference for the necessary information, as described below. It was also thought to better represent current patterns of reference usage among students in the composition course in question and, more generally, among tertiary-level ESL learners in U.S. universities, which is the population to which this study seeks to generalize.
The assessment presented the 10 items in randomized order, each on its own page (as shown in Figure 4). Instructions at the top told participants to copy and paste the sentence into the editing box, identify the error, and then correct it using information from an online resource. The reference selection was provided via a table of eight hyperlinks: two to monolingual pedagogical dictionaries; three to bilingual dictionaries (for Chinese, Korean and a variety of Romance languages); and three to dictionaries designed for native speakers of English.

In the sentence below, there is a problem with the pattern of one of the content words. (Content words include nouns, verbs, adjectives and adverbs.) You must identify which word is being used incorrectly, and then rewrite the sentence in the space provided to make it correct. You can COPY and PASTE the sentence into the editing box to save time. You should NOT change the content word itself in any way. Instead, focus on the words surrounding it. You may need to add, delete or change the surrounding words. Do NOT change the meaning of the sentence.

Use one of these dictionaries, or any other online dictionary, to help you. You may NOT use handheld electronic dictionaries/translation, paper dictionaries, mobile phones, or any resource besides online dictionaries.

In the 1950s, American scientists warned people for the harmful effects of television.

Answer:

Figure 4 Item from the PICT showing instructions, reference selection, sentence to be corrected, and editing box

The sentences were found in the Longman Learners Corpus by searching high-frequency node words featured in pattern grammar references (Francis, Hunston, & Manning, 1996, 1998). The sentences were edited to make the context clear and to isolate the pattern grammar error as the only infelicity. A list of acceptable corrections was compiled using information from the pedagogical dictionaries included among the reference links, as well as from the Corpus of Contemporary American English (Davies, 2008), and by surveying eight native speakers of English who have master’s degrees or higher in TESOL or applied linguistics and three or more years of ESL/EFL teaching experience. Item characteristics are summarized in Table 4. Four items dealt with verb patterns, four with noun patterns, and two with adjective patterns. Most corrections involved changes to prepositions, usually by substitution but in some cases by
addition or removal. Some substitutions involved complement clauses: an –ing form in the case of one verb pattern, and a that-clause in the case of two adjective patterns. In several cases, different types of correction were possible.

**Table 5** Test items \((k = 10)\) characterized by node word type and type of correction required

<table>
<thead>
<tr>
<th>Node word type</th>
<th>Number of items</th>
<th>Correction type required</th>
<th>Number of items</th>
</tr>
</thead>
<tbody>
<tr>
<td>Verb</td>
<td>4</td>
<td>substitute complement clause</td>
<td>3*</td>
</tr>
<tr>
<td>Noun</td>
<td>4</td>
<td>substitute preposition</td>
<td>7*</td>
</tr>
<tr>
<td>Adjective</td>
<td>2</td>
<td>add preposition</td>
<td>1†</td>
</tr>
<tr>
<td></td>
<td></td>
<td>remove preposition</td>
<td>2†</td>
</tr>
</tbody>
</table>

*Note: Max. possible score = 20; * either correction type was acceptable in the case of two items; † either correction type was acceptable in the case of one item*

For scoring, one point was awarded if the participant made any change to the “grammatical” component of the pattern (i.e. any part other than the lexical node word), for example, by substituting a preposition. Such a change was interpreted as evidence that the error had been located. An additional point was awarded if the change matched an item in the list of acceptable corrections. This polytomous approached allowed some differentiation, and therefore meaningful variation, between the ability to identify an error and the ability to correct it. The points achieved on the 10 items were summed to obtain each participant’s score out of a possible 20.

Reliability was determined using the phi coefficient as an index of dependability, which is appropriate for criterion-referenced tests (Bachman, 2004). The specific technique was Cronbach’s alpha calculated for both the pre-test and post-test administrations of the instrument with scores from the discovery and instruction groups (total \(N = 64\)). The combined pre-test data for both groups returned a value of \(\alpha = .744\), while the post-test value was \(\alpha = .865\).^5

Confidence scores on the PICT were collected via a single-question online survey instrument, which participants were directed to immediately after completing the assessment. Instructions at the top of the survey form explained the grading scheme and told participants to select a particular score from a range of 0-20 to represent how well they thought they had performed. Because of technology limitations, they could not review their responses on the PICT during this self-assessment.
Screen capture footage

In addition to the performance and confidence scores, the PICT was also the source of process data assessing use of dictionaries. Each participant’s performances on both administrations of the measure were recorded using the video screen capture program Camtasia. The video files were then imported into Transana, a software tool for analyzing audio and video data, where they were coded and transformed into two quantitative measures of referencing ability:

1. **Lookup counts.** These were counts of the number of individual dictionary searches performed by each participant. Searches were divided into two categories: *learner-dictionary lookups* (i.e. in monolingual ELT dictionaries) and *non-learner dictionary lookups* (i.e. bilingual dictionaries, “native speaker” dictionaries, and non-dictionary tools such as search engines). When screen-capture footage showed a particular word, phrase, or sentence being searched, a lookup was recorded in the appropriate category, except in cases where the same search had already been conducted in the same resource in the same administration of the assessment.

2. **Lookup points.** These were points achieved on the pre-test/post-test that were attributed to use of a learner dictionary, another type of reference, or else to a participant’s use of pre-existing knowledge, intuition, or guesswork. If screen-capture footage showed explicit information in an online resource being used in a correct or partially correct response, the points from that response were coded either as *learner dictionary lookup points* or *non-learner dictionary lookup points*, depending on the type of resource. If language data used in a correct or partially correct response could not be traced to an online resource captured in the video footage, the corresponding points were attributed to pre-existing knowledge, intuition, or guesswork and coded as *null lookup points.*

Treatments

*Discovery condition*

Participants in the discovery group were assigned a recurring activity in the project LMS requiring them to add entries to a personal vocabulary database over a five-week period. The activity was an online version of the paper-based task that normally constitutes the vocabulary component of the composition course in question. Learners are supposed to identify words in their course readings or other sources that they want to learn, then research and record lexical
features and other information that will help in using and remembering these words. The “add entry” form in the database included fields for information about the original context of use, pronunciation, part of speech, additional meanings, example sentences, derivatives, frequency and “usage information.” (A sample entry provided to discovery participants as a model is shown in Figure 5.) The researcher made comments and suggestions on participants’ database entries every week but gave no feedback regarding usage information, which was the field where the participants might be expected to include pattern grammar data.

A brief video, provided via a link in every weekly database assignment, explained the purpose of the activity and showed how to create an entry. The video instructions recommended the use of monolingual learner’s dictionaries and, regarding usage information, told learners to provide “common collocations or patterns” used with their chosen word. The instructions also emphasized the need to focus on information that would help them avoid common vocabulary mistakes such as those they had seen on the first (pre-test) administration of the PICT. In this way, discovery group participants were given an incentive, as well as repeated opportunities, to develop an understanding of pattern grammar and related referencing skills if they did not already possess them.
**Instruction condition**

Participants in the instruction condition were provided with an integrated form of language awareness and dictionary skills training, as discussed in the review of lexicographical research above. This instruction was delivered via the project LMS in an automated, online course called VVT (Virtual Vocabulary Trainer), which lasted five weeks and consisted of 11 tutorials organized around three themes: *Deep Vocabulary Knowledge, Dictionaries,* and *Word Patterns.* The first theme contextualizes pattern grammar as one of several kinds of vocabulary depth-of-knowledge that are important for “accurate, appropriate use” and thus sets the stage for development of the second and third themes. The VVT course is based on instructional design principles that aim to 1) reduce the burden on L2 learners’ processing capacities while using monolingual references, and 2) differentiate appropriately among instruction targeting declarative knowledge (e.g. conceptual understanding of pattern grammar), procedural knowledge (e.g. the use of shortcut keys for fast dictionary searches), or the integration and coordination of the two. The course was developed as part of a separate study to investigate the feasibility of an online form of L2 strategy instruction (Ranalli, in preparation).

The VVT tutorials consisted of multimedia presentations and accompanying tasks. The presentations, which incorporate text, images, animation, live-action video and voice-over narration, were used to develop conceptual knowledge and to model the target skills. The accompanying tasks included a variety of text-based activities designed to develop learners’ abilities in:

- differentiating among dictionary types and choosing the appropriate type for one’s purpose;
- selecting the appropriate sense of a word from among those provided in dictionary entries;
- identifying pattern-grammar information in learner dictionaries;
- identifying patterns in sentences and longer stretches of text;
- using pattern information from the dictionary to complete or correct sentences;
- choosing the appropriate pattern for a given context of use; and
- using keyboard shortcuts and keyword search tools to reduce cognitive processing demands and speed up dictionary searches.
Procedures

As illustrated in Figure 6, the study began in the second week of the semester. The researcher met with each of the three sections of the composition course on a regularly scheduled computer lab day and explained the purpose of the study. To encourage participation, students were told a $30 gift card would be awarded to three randomly selected participants at the end of the semester. Regardless of participation, however, they were expected to complete the treatments as part of the course requirements. Following recruitment, all students were registered in the project LMS. They were shown a short video giving instructions for the PICT two times and told they would be taking the test again later in the semester to see how much they had learned. They were then given 20 minutes to complete the pre-test administration of the instrument under the supervision of the researcher and their composition instructor, immediately after which they post-dicted their score. A biodata questionnaire was assigned for homework. The following week, each section again met in the computer lab to take the VLT. They were given 45 minutes to complete the test and to post-dict their score on each section. The results of the VLT were used to determine matched pairs for random assignment to the instruction group (Group 1) and discovery group (Group 2). Corresponding groups were then created in the project LMS. During weeks 4-8, Group 1 worked through the VVT materials while Group 2 was assigned the recurring discovery activity. In week 9, regularly scheduled computer lab days for each section were again used for the post-test administration of the PICT, which was also preceded by a showing of the video instructions twice. The participants again had 20 minutes to complete the post-test, and they post-dicted their score immediately afterward. This concluded the experiment phase of the study. In the remaining weeks of the semester, the treatment conditions were switched so Group 2 could also experience the VVT training.
RESULTS

Performance and confidence data are reported first to answer Research Question 1 regarding calibration across types of measure and Research Question 2 regarding group differences on the depth-related measure after treatment. This is followed by reporting of the process data to answer Research Question 3 regarding use of pedagogical dictionaries.

Calibration across types of vocabulary measure

Table 5 presents descriptive statistics for performance, confidence and calibration figures from the VLT and pre-test administration of the PICT. The sample was reduced by one for the VLT because one participant had neglected to provide confidence scores for any of the levels. As can be seen, overall performance as a percentage was much higher on the VLT than the PICT, while the margin of miscalibration on the VLT was much smaller. In addition, the negative bias evident in the VLT averages indicates a trend toward slight underestimation, as opposed to substantial positive bias, or overestimation, in the miscalibration figures for the PICT.
Table 5 Descriptive statistics for the VLT and pre-test administration of the PICT

<table>
<thead>
<tr>
<th></th>
<th>VLT</th>
<th>PICT</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N = 63</td>
<td>N = 64</td>
</tr>
<tr>
<td>Max. possible</td>
<td>score = 120</td>
<td>score = 20</td>
</tr>
<tr>
<td>Mean performance</td>
<td>89.27 (21.28)</td>
<td>6.44 (4.44)</td>
</tr>
<tr>
<td>Performance %</td>
<td>74.39%</td>
<td>32.20%</td>
</tr>
<tr>
<td>Mean confidence</td>
<td>83.38 (20.31)</td>
<td>12.45 (3.66)</td>
</tr>
<tr>
<td>Δ Confidence - performance</td>
<td>-5.89</td>
<td>6.01</td>
</tr>
<tr>
<td>Miscalibration %</td>
<td>-4.90%</td>
<td>30%</td>
</tr>
</tbody>
</table>

Note. Standard deviations in parentheses.

These trends are more clearly seen in the scatterplots in Figure 7. On the VLT, most performance scores occur on the right side of the graph, above the 50% mark, while on the PICT most occur on the left, below the 50% mark. The majority of the VLT data points fall slightly below the identity line, which represents perfect calibration, whereas the majority of PICT data points fall above this line, many of them quite far above, indicating substantial overestimation. Moreover, the regression line for the VLT data is nearly coincident with the identity line, whereas this is not the case for the PICT.

Figure 7 Scatterplots of performance and confidence data for the VLT and PGT, with identity line, regression line and equation, and coefficient of determination
Correlation coefficients calculated in SPSS 18.0 showed a much stronger confidence-performance relationship on the VLT, $r(61) = .764, p < .001$, than on the PICT, $r(62) = .327, p = .008$, with respective effect sizes that would be characterized as large, $R^2 = .584$, and small to medium, $R^2 = .107$, according to the scale in J. Cohen (1992). Using regression equations generated for each data set, a “low” score on the VLT of one standard deviation below the mean predicts perfect calibration (performance and confidence scores both = 68), whereas on the PICT, a “high” score of one standard deviation above the mean predicts overestimation of 10% (performance = 11, confidence = 13). Thus, the sample’s VLT performance bears hallmarks of mastery or expertise, while the performance on the PICT is generally consistent with a lack of domain knowledge.

To test for a hard-easy effect among the levels of the VLT, descriptive statistics were calculated (Table 6). The difference scores indicated similar patterns of slight underestimation at each level. These scores were submitted to a one-way ANOVA in SPSS, with difference score as the dependent variable and level as the independent variable. No statistical differences were found in the omnibus test, $F(3,248) = .310, p = .818$, partial $\eta^2 = .004$, which means the sample exhibited the same level of calibration in self-assessing their size-related knowledge of lower-frequency, higher-frequency, and genre-specific words, despite varying levels of actual knowledge of these items.

**Table 6** Performance, confidence and calibration scores for VLT by level ($N = 63$)

<table>
<thead>
<tr>
<th></th>
<th>2000 level</th>
<th>3000 level</th>
<th>5000 level</th>
<th>AWL level</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean performance</td>
<td>25.57 (4.43)</td>
<td>22.11 (5.94)</td>
<td>18.75 (6.25)</td>
<td>22.84 (6.11)</td>
</tr>
<tr>
<td>Performance %</td>
<td>85.24%</td>
<td>73.70%</td>
<td>62.49%</td>
<td>76.14%</td>
</tr>
<tr>
<td>Mean confidence</td>
<td>24.33 (4.52)</td>
<td>20.25 (5.47)</td>
<td>17.52 (5.59)</td>
<td>21.27 (6.14)</td>
</tr>
<tr>
<td>$\Delta$ Confidence - performance</td>
<td>-1.24</td>
<td>-1.86</td>
<td>-1.22</td>
<td>-1.57</td>
</tr>
<tr>
<td>Miscalibration %</td>
<td>-4.13%</td>
<td>-6.19%</td>
<td>-4.07%</td>
<td>-5.24%</td>
</tr>
</tbody>
</table>

*Note. Standard deviations in parentheses. Max. possible score = 30 at each level. AWL = Academic Word List*

**Treatment effects on calibration and actual performance on the depth-related measure**

Descriptive statistics for the pre-test and post-test administrations of the PICT are reported by group in Table 7. The figures show similarly low achievement on the pre-test, with a slightly higher mean performance score for the instruction group. For both groups, miscalibration was
positively biased and considerable, at roughly five times the level reported for the vocabulary data in Phakiti (2005). On the post-test, actual performance for both the discovery and instruction groups increased, by more than 100% in the case of the instruction group but only by about 20% for the discovery group. Miscalibration is still positively biased for both groups and, while it has decreased considerably for the instruction group, the discovery group figure actually shows a slight increase.

Table 7 Descriptive statistics for both administrations of the PICT

<table>
<thead>
<tr>
<th></th>
<th>Pre-test</th>
<th>Post-test</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discovery</td>
<td>Instruction</td>
</tr>
<tr>
<td>Mean performance</td>
<td>6.03 (3.8)</td>
<td>6.84 (5.03)</td>
</tr>
<tr>
<td>Performance %</td>
<td>30.2%</td>
<td>34.2%</td>
</tr>
<tr>
<td>Mean confidence</td>
<td>12.22 (3.89)</td>
<td>12.69 (3.45)</td>
</tr>
<tr>
<td>Δ Confidence - performance</td>
<td>6.19</td>
<td>5.84</td>
</tr>
<tr>
<td>Miscalibration %</td>
<td>30.9%</td>
<td>29.2%</td>
</tr>
</tbody>
</table>

*Note.* Standard deviations in parentheses. Max. possible score = 20.

To test for significance, the scores were submitted to a multivariate analysis of covariance (MANCOVA) using the GLM procedure in SPSS, with post-test performance score and post-test calibration score (i.e. confidence minus performance) as dependent variables and group as the fixed factor. Adjustment was made for two covariates: matched pair and pre-test performance score, with the latter included to increase the power of the analysis (Maxwell & Delaney, 2004) and to determine if initial performance on the PICT affected performance and calibration on the post-test. Tests of the assumptions of equality of covariance matrices and error variances were satisfactory but analysis of histograms and Q-Q plots indicated non-normality in both the pre- and post-test data. A base-10 logarithmic transformation, which is recommended in cases of substantial positive or negative skew (Tabachnick & Fidell, 2001), failed to resolve these issues, so it was decided to continue with the analysis while acknowledging the loss in statistical power. In addition, Mahalanobis distance revealed the presence of a multivariate outlier among the discovery group participants at \( p = .00109 \). The analysis reported below was repeated with this participant’s data removed, and the pattern of overall findings remained the same.

The results of the MANCOVA showed that group had a significant effect on post-test performance, \( F(1,60) = 34.529, \ p < .001 \), partial \( \eta^2 = .365 \), as well as post-test calibration,
\[ F(1,60) = 24.206, \ p < .001, \ \text{partial } \eta^2 = .287. \] No effect of matched pair was found on either dependent variable. Interestingly, while adjustment for pre-test performance had a significant effect on post-test performance, \[ F(1,60) = 9.181, \ p = .004, \ \text{partial } \eta^2 = .133, \] there was no effect on post-test calibration, \[ F(1,60) = .557, \ p = .458, \ \text{partial } \eta^2 = .009, \] which means improvements in calibration were not related to initial level of performance.

To address the hypothesis regarding lack of significant change in performance and confidence in the absence of explicit instruction, the performance and calibration scores for the discovery group \((n = 32)\) were isolated and submitted to paired-samples t-tests. The first test, comparing pre-test performance \((M = 6.03, \ SD = 3.80)\) to post-test performance \((M = 7.22, \ SD = 4.94)\), returned a non-statistical result, \[ t(31) = -1.536, \ p = .135, \ \text{two-tailed, } d = .27. \] The second test, comparing the calibration (confidence minus performance) score on the pre-test \((M = 6.19, \ SD = 4.53)\) to calibration score on the post-test \((M = 6.84, \ SD = 4.78)\), also showed no statistical difference, \[ t(31) = -.799, \ p = .430, \ \text{two-tailed, } d = .139. \]

The scatterplot in Figure 8 provides more information about group differences on the post-test. Whereas the instruction group data points cluster above 50% on the performance scale and closer to the identity line, including several on the line itself or below it – suggesting some movement in the direction of expertise – the majority of discovery group data points occur below the 50% mark and relatively high above the identity line. Correlational analysis showed a weaker performance-confidence relationship among the discovery group, \[ r(30) = .412, \ p = .019, \] compared to the instruction group \[ r(30) = .605, \ p < .001, \] although it was an improvement on pre-test correlation observed for the whole sample (.327).
Figure 8 Scatterplots of performance and confidence data for both groups on the post-test administration of the PICT, with identity line, regression line and equation, and coefficient of determination

Treatment effects on dictionary use

Dictionary lookup counts. Descriptive statistics are reported in Table 8. Analysis of the lookup counts indicated similar usage patterns for both groups on the pre-test; on average, participants performed less than three consultations in learner dictionaries and less than four in non-learner dictionaries. The figures diverge markedly on the post-test, however. Among the discovery group there is a slight increase in learner dictionary lookups and a slight decrease in non-learner dictionary lookups, whereas among the instruction group, learner dictionary lookups have jumped to an average of 9.53 (almost one lookup per item on the PICT) and the average for non-learner dictionary lookups has almost dropped to zero.
Table 8 Descriptive statistics for dictionary lookup counts

<table>
<thead>
<tr>
<th></th>
<th>Learner dictionary lookups</th>
<th>Non-learner dictionary lookups</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discovery $n = 32$</td>
<td>Instruction $n = 32$</td>
</tr>
<tr>
<td>Pre-test</td>
<td>2.75 (3.94)</td>
<td>2.63 (3.91)</td>
</tr>
<tr>
<td>Post-test</td>
<td>3.34 (4.71)</td>
<td>9.53 (4.6)</td>
</tr>
</tbody>
</table>

Note. Standard deviations in parentheses

To test for significance, the data were fitted to a generalized linear model using the logarithm as the link function and the Poisson distribution as the probability distribution, which is appropriate in cases where data represent counts and large count outcomes are rare (Kutner, Nachtsheim, Neter, & Li, 2005). Using the GENLIN procedure in SPSS, a Poisson regression was modeled to represent the mean rate at which participants performed lookups in learner dictionaries, with learner dictionary lookups as the response variable, group and time as predictors, and VLT scores as a covariate. Non-learner dictionary lookups were not included in the model as this would have violated the assumption of independence and because use of learner dictionaries was the variable of primary interest. A test of the full model against an intercept-only model was statistically significant, $\chi^2 = 45.049$, $df = 4$, $p < .001$. Pairwise comparisons showed the mean difference in learner dictionary lookups for the discovery group from pre-test to post-test was not significantly different ($M = -0.58$, 95% CI: -1.92, 3.09), whereas for the instruction group, the mean difference in their own learner dictionary lookups from pre-test to post-test was significantly different ($M = -6.79$, $p \leq .05$, 95% CI: 3.24, 10.34), as was the mean difference in post-test learner dictionary lookups between the instruction and discovery groups ($M = 6.08$, $p \leq .05$, 95% CI: 2.43, 9.73).

Dictionary lookup points. Descriptive statistics are reported in Table 9. Analysis of the lookup counts again showed pre-test averages that are similar while post-test figures contrast sharply, especially those representing points obtained through the use of learner dictionaries. Both groups achieved the majority of points on the pre-test without the aid of a reference, with learner dictionary lookups accounting for an average of less than two points and non-learner dictionary lookups accounting for an average of about half a point. The discovery group’s averages change slightly on the post-test, with points from learner dictionary lookups increasing and points from non-learner dictionary and null lookups decreasing. Among the instruction group, by contrast, a
sixfold increase in points from learner dictionary lookups was observed, with no points attributed to non-learner dictionary lookups, and null lookup points reduced nearly by half.

Table 9 Descriptive statistics for dictionary lookup points (max. possible score = 20)

<table>
<thead>
<tr>
<th></th>
<th>Learner dictionary lookup points</th>
<th>Non-learner dictionary lookup points</th>
<th>Null lookup points</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Discovery</td>
<td>Instruction</td>
<td>Discovery</td>
</tr>
<tr>
<td></td>
<td>$n = 32$</td>
<td>$n = 32$</td>
<td>$n = 32$</td>
</tr>
<tr>
<td>Pre-test</td>
<td>1.16 (2.69)</td>
<td>1.78 (3.63)</td>
<td>0.44 (1.04)</td>
</tr>
<tr>
<td>Post-test</td>
<td>2.94 (5.3)</td>
<td>11.06 (6.02)</td>
<td>0.41 (1.01)</td>
</tr>
</tbody>
</table>

Note. Standard deviations in parentheses

To test for significance, the pre-test averages were subtracted from the post-test averages to obtain the difference scores in Table 10. These were then submitted to a two-sample Hotelling’s $T^2$ test using the R statistical software package. Hotelling’s $T^2$ is a generalization of Student’s $t$ for multivariate data that is used when there are multiple dependent variables likely to be correlated with each other, and when the independent variable comprises only two levels (Tabachnick & Fidell, 2007). The test showed a significant difference between discovery and instruction score vectors, $T^2(3,60) = 12.98$, $p < .001$. Post-hoc measures consisted of Welch two-sample t-tests, which returned a statistical result in the case of learner dictionary lookup points, $t(52.43) = 6.22$, 95% CI 5.08,9.92, $p < .001$.

Table 10 Mean differences in pre-test and post-test lookup points

<table>
<thead>
<tr>
<th></th>
<th>Learner dictionary lookup points</th>
<th>Non-learner dictionary lookup points</th>
<th>Null lookup points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Discovery</td>
<td>1.78</td>
<td>-0.03</td>
<td>-0.56</td>
</tr>
<tr>
<td>Instruction</td>
<td>9.28</td>
<td>-0.59</td>
<td>-1.69</td>
</tr>
</tbody>
</table>

To test for differences between the discovery group’s pre- and post-test performances as determined by dictionary use, their lookup points were isolated from the instruction group’s lookup points and then submitted to another two-sample Hotelling’s $T^2$ test, with learner dictionary lookup points, non-learner dictionary lookup points and null lookup points as the dependent variables and time as the independent variable. The test found no nonstatistical differences, $T^2(3,60) = 1.03$, $p = .385$. 
These findings depict sharply contrasting patterns of monitoring and feedback among the sample with respect to size-related versus depth-related vocabulary knowledge. Monitoring accuracy on the VLT exhibited patterns of expertise or mastery, while patterns on the pre-test administration of the PICT were consistent with a lack of domain knowledge. That is, on a test of pure vocabulary size, learners had a good sense of what they did and did not know, whereas on the depth-related measure this was far from the case. This supports the original hypothesis, as well as the popular perception of a considerable gap between knowing what a word means and knowing how to use it.

On the chance that the participants were not familiar with pattern grammar because they had not previously been confronted with a need to use or learn about it, the discovery condition provided them with such an opportunity. It was hypothesized that learners, confronted with a need for pattern-grammar knowledge and repeated opportunities to discover it, would experience difficulty doing so and would not exhibit significant differences in actual performance, monitoring accuracy or reference use following a treatment incorporating those conditions. The data showed this to be the case. The discovery group exhibited a lack of significant improvement on the post-test in terms of both performance and monitoring accuracy – this despite ready access to pedagogical references designed specifically for such purposes and for learners like themselves. This constitutes evidence that the primary internal and external sources of feedback that L2 learners need to guide their self-directed vocabulary learning may be unreliable when it comes to certain features of vocabulary depth-of-knowledge.

To the degree that the learners and context in this study are representative, it thus seems clear that bootstrapping one’s way into greater knowledge and skills vis-à-vis features of word combinability will be difficult for many tertiary-level ESL learners. However, such bootstrapping evidently does occur. The parallel coordinate plots in Figure 9 show changes in direction and magnitude of performance scores across administrations of the PICT for both groups. Among the discovery group, a standout is the participant who scored 2 on the pre-test and 17 on the post-test, and whose calibration scores were +10 and -1, respectively. A post-experiment interview conducted as part of a separate but related study revealed that this participant had substantial previous experience with pedagogical dictionaries as well as a habit of
reading in English for pleasure. Also remarkable was the discovery group participant who scored 18 on the pre-test and 20 on the post-test (calibration scores -2 and 0, respectively). Like her high-scoring counterpart, this participant was very familiar with pedagogical dictionaries, often read in English for pleasure, and moreover had a large receptive vocabulary (115 out of 120 on the VLT) as well as a felt need to learn to use words more correctly. In an interview, she claimed to have discovered the role of lexical patterns, and the value of learner dictionaries for referencing them, while taking the pre-test. Yet these were exceptional cases distinct from the overall poor discovery-group performance depicted in Figure 9, in which post-test scores seemed as likely to decrease as increase.

![Figure 9 Direction and magnitude of changes in individual performances on the PICT](image)

Examination of individual performances yields additional insights not conveyed by the quantitative results. The screen-capture footage showed a number of participants using the LDOCE on the pre-test, perhaps having noticed its prominence in the video instructions, and even managing to find entries for the appropriate node words, yet seemed not to notice bolded dictionary information representing the correct answers. In addition, while many discovery participants followed instructions to use learner dictionaries in their vocabulary database assignments – with some including word-pattern information in their entries – they chose not to avail themselves of the same dictionaries on the post-test. For example, one discovery participant
used pattern information from the LDOCE in seven of his nine submissions, but nevertheless completed the post-test without consulting any reference, with performance and calibration scores of 2 and +12. With respect to errors on the PICT, untrained participants often corrected those involving the patterns *stop doing*/stop to do sth* and *pay attention to sth/sby* without the aid of a reference. These were probably familiar from their previous learning, but not as instances of a broader category of lexical knowledge (as noted by Hunston et al., 1997), so when confronted with other sentences for which existing item knowledge failed them, most appeared to reinterpret the task as “find any error.” Few items went unanswered on either the pre-test or post-test; a correction was nearly always attempted, and since many such attempts did not result in even partial credit, this means participants were changing well-formed parts of the sentences and thus often creating additional errors.

This suggests a role for concepts or schemata in addressing L2 vocabulary depth-of-knowledge. In arguing against the default hypothesis, Laufer (2005, 2010) and others have proposed form-focused instruction and “focus on forms”; that is, explicit teaching of the various meanings, forms and uses of particular words, presumably in hopes that learners can extrapolate from such instances a general understanding of phenomena such as collocation. However, the success of the explicit instruction condition in this study, which focused on concepts as opposed to specific examples, indicates the value of a top-down approach. Conceptual structures help knowledge and skills transfer to new situations (Salomon & Perkins, 1989). They may also help narrow the number of possible outcomes in grammaticality judgments, as well as assist in identification of unfamiliar patterns in text and in categorization and interpretation of dictionary information. Importantly, they may also facilitate learners’ definitions of tasks, which represents one of several theoretical implications that will now be discussed.

**Implications**

These findings support the need for an explicit mechanism connecting monitoring and feedback to strategic behavior in the model of L2 vocabulary learning proposed by Tseng and Schmitt. Without such a mechanism, the model cannot adequately explain or predict why learners may have more trouble with depth-related versus size-related learning, nor how strategic mastery can develop from inexpert strategy use. Miscalibrations of the type documented here may introduce dysfunctions into the feedback loop and reduce incentives to modify one’s strategic repertoire.
For example, in the post-learning appraisal stage, an individual may possess a skilled or satisfied feeling of having used strategies well, but if those feelings are based on erroneous self-assessments, there will be no incentive to modify subsequent use of strategies so as to bring future performance closer in line with desired goals or standards, nor any information about how to do so. Although research has yet to determine specific miscalibration thresholds beyond which self-regulation becomes impossible, the levels observed in this study do not inspire optimism.

The data and the research discussed here suggest elaborations may also be needed regarding the processes that precede strategy selection. One researcher recently reviewed calibration studies to illustrate a “‘frame of reference’ problem which occurs because students have idiosyncratic perceptions of task demands” (Pieschl, 2009, p. 6; emphasis in original). Before setting goals and choosing strategies to fulfill them, learners internally generate a perception of the task requirements and the resources and constraints available to meet these requirements, and these internal models are susceptible to bias, particularly if the task is complex or poorly defined (Winne & Hadwin, 1998). For example, Glenberg and Epstein (1987) found that undergraduates who classified themselves as experts with respect to a particular subject area were more miscalibrated on a related comprehension test than students majoring in another field; self-classification was seen as leading students to confuse familiarity with knowledge. This calls to mind those participants in the dictionary-use studies by Laufer (2011), and Frankenberg-Garcia (2011), whose unfounded confidence in their knowledge of word usage may have been influenced by their knowledge of word meanings. L2 metacognitive strategies are still often discussed in terms of “planning, monitoring and evaluation” (Tseng & Schmitt, 2008, p. 384), which seems to assume appropriate task definition on the part of the learner, despite research showing the potential for idiosyncratic perceptions of tasks to influence learning goals and strategic behavior (e.g. Cotterall, 1995; Elbaum, Berg, & Dodd, 1993; Mori, 1999; Peck, 1990; Wenden, 1995, 1998).

This argues the need for a more variegated, flexible conceptualization of appraisal in Tseng and Schmitt’s model. Appraisal could be expanded to include modeling or definition of the task by the learner, thus potentially taking account of gaps between perceptions and objective task demands. Its jurisdiction could also be extended throughout the learning cycle, as is the case in Boekaerts, who describes appraisals as “non-stop comparison processes [that] alter upcoming and ongoing cognitions and feelings” (Boekaerts, 1991, p. 8), and who sees monitoring and
evaluation as a form of appraisal that occurs during strategy use. Rather than conceptualizing appraisal as a metacognitive strategy that requires prompting or training, as has often been the case in the L2 strategy literature, it might be seen as a basic cognitive operation that occurs throughout learning, that operates over task definitions, goals and strategies, as well as learning conditions, motivation and affect, and that will be more or less effective depending on its inputs.

A model incorporating the elaborations proposed here would not be as parsimonious as Tseng and Schmitt’s, but it would render learning more amenable to internal and external influences throughout the process, thus better reflecting a complexity that accords with our actual experience of L2 vocabulary learning. Such a model may in the end have more in common with contemporary models of SRL (e.g. Winne & Hadwin, 1998; Zimmerman, 2000) than with the motivation-oriented framework proposed by Tseng and Schmitt. SRL models consider motivation and volition to be key factors among others such as strategic behavior and metacognition. They make the process of self-regulation the organizing principle and overarching construct, rather than giving primacy to one variable and subsuming others beneath it. Thus, this paper may ultimately be calling not for a reformulation but a new start toward development of a model of self-regulated learning of L2 vocabulary.

There are other, more practical implications in addition to those already mentioned. This study, and another related investigation (Ranalli, in preparation) provide evidence of the efficacy of an automated, online form of strategy instruction that can address the need for combined language awareness and dictionary reference skills among tertiary-level ESL learners. This type of instruction potentially addresses long-standing questions about the cost-effectiveness of L2 strategy instruction (see Plonsky, 2011) and provides an example of a hybridized research design combining the advantages of the laboratory (random assignment, control over instructional treatments) with those of the classroom (greater ecological validity, especially vis-à-vis motivation). There is also cause to believe that raising learners’ awareness of monitoring accuracy can benefit self-regulation. The present study has shown how simple, global measures of confidence are easy to collect as part of performance assessments. Calibration data could easily be made available to learners as a complement to strategy instruction. Performance and calibration feedback could be part of a wider feedback strategy that also includes process feedback, which research has shown to be in many ways the most valuable form. Finally, the most important and perhaps obvious pedagogical implication is that L2 learners need more
support for learning aspects of word combinability as well as other features of vocabulary depth-
of-knowledge of the types included, but often not consulted, in pedagogical dictionaries. Many learners who are left to their own devices with respect to depth will be unable to recognize what it is they do not know, which precludes the possibility of doing anything about it.

LIMITATIONS AND SUGGESTIONS FOR FURTHER RESEARCH

Several factors limit the generalizability of these findings and suggest directions for future research. First, the length of treatment of the discovery condition may have kept more discovery group participants from developing knowledge of pattern grammar and related dictionary skills. That said, the line between “true” discovery and guided discovery will always be indistinct, and a longitudinal study under more naturalistic conditions could contribute to knowledge in this area. Also, the instruments used to measure size- and depth-related knowledge are open to criticism. The VLT, while supported by validation research, is a measure of “pure” size that may not accurately represent the nature of the size-related decisions learners usually have to make in their self-directed learning in the way that a test of lexical inferencing might. Future studies in this vein could usefully incorporate such measures as a middle ground on the size-depth spectrum. In addition, the findings were probably influenced by the predominance of Chinese-speaking participants, whom some research has found to be more confident, and overconfident, compared to other ethnic groups (Yates, Lee, & Shinotsuka, 1996). Research should be conducted in different instructional contexts, and with different proficiency levels and age groups, to see if the findings hold true. It is also clear that finer-grained research is needed to reveal why the exceptional cases in the discovery group were able to bootstrap themselves into strategic mastery when the rest were not. Case studies and online methods such as thinkalouds will prove helpful in this regard. Finally, research is needed to delve further into the reasons for the differential challenges of size- versus depth-related learning, including, as noted, the potential role of conceptual mediation.
NOTES

1. Tseng and Schmitt describe using Dörnyei’s dynamic model of L2 motivation as a theoretical foundation, mapping their vocabulary-related variables on to Dörnyei’s three stages of choice motivation, executive motivation and retrospective motivation. IAVLE corresponds to the first or preactional stage, in which motivation is initially generated; SRCvoc, SVLI and PAVLT occur in the middle or actional stage, where the motivation generated in the first stage is protected and sustained; and VOCKno and PAVLT are positioned in the post-actional stage, where learners evaluate the outcomes of learning and reflect on motivational processes.

2. Some overlap may be seen between monitoring in this sense and that embodied in Krashen’s Monitor Hypothesis. While there are functional similarities, Krashen was concerned with the acquisition of syntax, not vocabulary, and the basis of monitoring in his model was consciously learned grammatical “rules” (Krashen, 1979). Also, Krashen’s work is not based in an information-processing view of human cognition.

3. Calibration can be contrasted with self-assessment, which has been the focus of a number of second language studies (Blanche, 1990; LeBlanc & Painchaud, 1985; Oscarson, 1989, 1997; Oskarsson, 1984; Roever & Powers, 2005; S. Ross, 1998). According to Phakiti (2005), calibration refers to measures of confidence made with respect to a particular performance measure, while self-assessment refers to more general evaluations of one’s abilities. As noted, calibration is also referred to as monitoring accuracy, which was the focus of an ESL pronunciation study by Ellis and Zimmerman (2001).

4. As with similar programs at other universities, vocabulary is assumed in the present context to be addressed in other courses, especially reading courses, although Folse (2010) raised questions about the extent to which this actually occurs.

5. This increase in reliability is predictable if one assumes uniformly poor performance on the pre-test compared to more variation in ability on the post-test (following treatment) among which the instrument could then differentiate.

6. In 12 cases, it was unclear to the researcher whether information from Chinese-English dictionaries had been used to achieve correct or partially correct responses. For this reason, a
PhD student in applied linguistics whose first language is Mandarin Chinese was asked to review the relevant video excerpts to make a determination.

REFERENCES


Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*(2), 159-183. doi: 10.1016/0030-5073(77)90001-0


CHAPTER 5: TASK DEFINITION IN L2 STRATEGY USE: INVESTIGATING DISCREPANT CASES WITHIN A MODEL OF SELF-REGULATED LEARNING

A paper to be submitted to TESOL Quarterly

Jim Ranalli

ABSTRACT

Among the metacognitive processes involved in L2 learning, task definition (also called task analysis and task appraisal) is among the least understood. Research has documented discrepancies between unsuccessful learners’ strategic choices and task demands, but the reasons for these discrepancies remain largely unexplored. The present study builds on the work of Abraham and Vann (Abraham & Vann, 1987, 1996; Vann & Abraham, 1990) by drawing lessons from contrasts of successful and unsuccessful learners, and by triangulating task products along with process data and self-reports; as well as the work of Wenden (1995, 1996, 1998), who surveyed the cognitive and educational psychology literatures to investigate the role of metacognition in L2 strategy use. The present study was positioned within an integrative, multicomponential model of self-regulated learning (Winne & Hadwin, 1998) that views task definition as a process separate from strategic planning, thus allowing a more fine-grained analysis of discrepant cases; and which takes account of the potential role of cognitive processing limitations. Four case studies were drawn from a larger research project evaluating the efficacy of a technology-mediated form of strategy instruction for L2 vocabulary learning; this allowed the strategy to be held constant, treatment effects to be isolated and examined over time, and a variety of data types to be collected. The findings show accurate task definition can be thwarted by epistemological beliefs and goal orientation, as found in previous research, but also by lack of domain knowledge interacting with limitations in processing capacity and monitoring problems. The implications for strategy instruction and L2 vocabulary learning, as well as the value of models of self-regulation in L2 studies, are discussed.
INTRODUCTION

Most classroom teachers have experienced setting a task, using instructions they assume to be sufficiently clear and complete, only to be confronted with agitated whispers as students try to clarify among themselves what is expected. This familiar scenario shows the ever-present potential for a gap between the provider’s original conceptualization of a task and its representation in the minds of learners. Teachers can elaborate or model, but to the extent that learning is self-directed, as aspects of second language learning often are, students may be on their own in interpreting task demands.

Tasks usually allow different approaches, and a particular approach to a particular task selected by a particular learner is the definition of a strategy that will be adopted here. Much L2 research has discussed learners adopting more or less successful strategies vis-à-vis particular tasks, and task-strategy mismatches have been discussed with reference to a number of different constructs in L2 learning, such as metacognition (e.g. Cotterall & Murray, 2009; Wenden, 1998); learner beliefs (e.g. Horwitz, 1987; Mori, 1999), and motivation (e.g. Dörnyei, 2003; Vandergrift, 2005). To date, however, little L2 research has directly investigated the process whereby learners form the internal task representations upon which strategy selection is based (Rubin, 2005; Wenden, 2002).

This process has been referred to as task analysis, but the same term is also used in the fields of task design (Foster & Skehan, 1996) and assessment (Cohen, 1998) to describe efforts undertaken by providers of tasks rather than intended recipients, which seems a more suitable usage. Alternative terms used to discuss the learner’s side of things include task appraisal, task classification, and task interpretation, but these might implicate attention and considered reflection in what can sometimes appear implicit and instantaneous, so this paper adopts the arguably more neutral term task definition.

This paper undertakes to add to the literature on task definition, particularly as it concerns self-directed learning of L2 vocabulary, and to demonstrate the value of one particular theoretical framework from the field of educational psychology (Winne & Hadwin, 1998) for use in studies of L2 task definition and related issues. In doing so, it draws on the empirical work of Abraham and Vann, and the theoretical work of Wenden, which will first be briefly reviewed.
L2 learning strategies and tasks

Abraham and Vann’s case studies (Abraham & Vann, 1987, 1996; Vann & Abraham, 1990) were an important development in L2 strategy research. In the 1980s, after initial reports of success in teaching learners helpful new strategies (Cohen & Aphek, 1980; Hosenfield, 1984), subsequent intervention research had documented only moderate gains (e.g. O'Malley, Chamot, Stewner-Manzares, Kupper, & Russo, 1985). Abraham and Vann argued that such interventions might rest on flawed assumptions, based as they were in observations of the behavior of successful learners and self-reports of their strategy use. Abraham and Vann’s own research showed that, contrary to assumptions, unsuccessful learners were not necessarily inactive or unfamiliar with the strategies in question, but unable to appropriately align their strategic choices to task demands.

These case studies, based in the same university as the present investigation, involved learners making their way through an intensive English program with varying degrees of success. The researchers used a variety of triangulated data sources, notably the products of different instructional tasks, including cloze exercises or guided writing prompts, in combination with thinkaloud protocols. Following Bialystok (1990), they argued that task products could yield important insights into a learner’s strategy use if analyzed with reference to an appropriate theoretical framework (Abraham & Vann, 1996).

In their analyses, the researchers speculated as to causes of task-strategy mismatches, including a lack of domain knowledge (e.g. grammar rules needed for completing a cloze), beliefs about language learning and previous experience of studying the language, as well as lack of attentional capacity when faced with a particularly complex task, leading to “cognitive overload” (Vann & Abraham, 1990, p. 191). While some role was seen for motivation, this was not measured as part of the research design, perhaps because their theoretical frameworks addressed primarily cognitive constructs such as analyzed knowledge and control (Bialystok & Ryan, 1985) and depth of processing (Craik & Lockhart, 1972).

Abraham and Vann called for more task-focused research into the relationships between strategy use and individual differences such as personality, age, and formal schooling, and drew implications for strategy instruction, saying it may be appropriate in cases where learners lack a strategy or fail to use it in contextually appropriate ways. While their selected theoretical
frameworks did not facilitate inclusion of learner beliefs and motivational influences, more recent developments in the field of educational psychology feature better integration of these constructs, as described in the work of Wenden, to which we now turn.

**Task definition as a metacognitive process**

In the field of L2 learning strategies, Wenden played a major role in shifting attention from the outward behavior of learners to the processes underlying this behavior (Gao, 2007) through a research synthesis incorporating work on metacognition from the cognitive and educational psychology literatures. In a series of publications (Wenden, 1995, 1996, 1998, 2002), she laid out a key distinction between, on the one hand, metacognitive knowledge as consisting of learners’ knowledge and beliefs about learning, including domain-specific beliefs about language learning, as well as motivational beliefs; and on the other, the metacognitive strategies that learners use to self-direct and regulate their learning. As part of this synthesis, she highlighted the fact that, in the common characterization of metacognitive strategies as consisting of planning, monitoring and evaluating, appropriate task definition on the part of the learner is often simply assumed.

Following Flavell (1979), Wenden divided metacognitive knowledge into person knowledge (regarding beliefs, experiences, self-efficacy, etc.), strategy knowledge (in general and about particular strategies) and task knowledge. In Wenden’s terms, task knowledge has three facets: task purpose, task classification, and task demands. Citing Boekaerts (1992), Wenden describes how perceptions of a task’s demands trigger an appraisal process whereby learners decide whether or not they possess the capability of completing the task and its relevance for their learning. If this appraisal is positive, they may adopt learning goals, but if not these may be coping goals – cf. mastery- versus performance-goal orientations in Ames (1992) and Pintrich (2000) – with obvious ramifications for the standards they will use in monitoring and evaluating use of strategies.

Moreover, Wenden cites Butler and Winne (1995) in noting that metacognitive knowledge indirectly influences the planning process by determining a learner’s perception of task cues. A learner is likely to classify a new and unfamiliar task with reference to other, similar tasks he has completed before. In other words, acquired knowledge and beliefs will influence the internal representations of a task, which in turn influences goals, strategies and evaluative standards. When learners’ metacognitive knowledge for the purpose is incomplete or flawed, they may not
recognize the need for using a strategy, select an unsuitable strategy, or adopt inappropriate
criteria for judging performance (Wenden, 1996). Thus, the dynamics of task definition, why it
sometimes may fail, and whether it can be taught as part of strategy instruction, are identified as
research priorities.

While Wenden’s work was pivotal, it can be criticized for insufficient attention to the role of
linguistic knowledge, in either declarative or procedural form, as a source of standards for
evaluating learning outcomes. Although she sees domain knowledge as crucial for task
definition, it is usually represented as background or content knowledge activated during
reading, or knowledge of discourse structures. Further, Wenden often presents the discussion
without reference to models, which makes it difficult to see how the separate components
interrelate, and leads to some confusing descriptions. For instance, the function of executive
control is assigned to monitoring, but monitoring would seem to logically require a higher-order
control function to enact it. Finally, no mention is made of processing limitations or the
constraints of working memory, although L2 research in areas such as task processing (Foster &
Skehan, 1996) and syntactic processing as part of metalinguistic judgments (Hu, 2002) attest to
their importance.

To build on Wenden’s contribution, a model is needed that separates task definition as a distinct
process and allows it to be analyzed in relation to beliefs and knowledge, motivation, and
strategy use. Ideally, this would be a process model that facilitates dynamic descriptions of the
interplay of the various components, while also taking account of the processing limitations
inherent in the cognitive architecture. Such a model is presented in the next section.

**Task definition as a phase in self-regulated learning**

In the last three decades, self-regulated learning (SRL) has evolved into a central concept in
educational psychology as researchers and theorists tried to combine cognitive, motivational,
behavioral and sociocultural elements in explaining the complex dynamics of self-direction and
self-management of learning. Winne and Hadwin’s (1998) model of SRL is one of many, which
vary in minor and major ways (see Zimmerman, 2001, for an overview). Among its
distinguishing features are the fact that task definition is separated out from planning, giving this
model four phases rather than the more typical three: task definition, goal setting and planning,
enactment of tactics and strategies, and adaptations, as shown in Figure 1. Second, Winne and
Hadwin’s model is based in input-processing theory (IPT), and a set of IPT-related processes is seen to operate in each of the four phases, rendering it capable of finer-grained predictions and analyses (Greene & Azevedo, 2007).

**Figure 8** Winne and Hadwin’s model of self-regulated learning (1998, p. 282)

These processes are captured in the acronym COPES, which stands for conditions, operations, products, evaluations, and standards. With the exception of operations, each of these components represents a type of information generated or used by the learner during a learning event (Greene & Azevedo, 2007).

Conditions are the resources available for, and any constraints that might operate over, completion of a task, and consist of two types. *Task conditions* are external to the learner and might include resources, cues embedded in tasks by a designer, and time, as well as features of
the social context, such as a requirement to work in pairs or small groups. Cognitive conditions are internal to the learner and include beliefs about learning, motivational influences (e.g. goal setting or goal orientation), domain knowledge, task knowledge, and knowledge of tactics or strategies. Conditions influence both operations and the standards learners adopt in goal setting and planning (cf. Wenden on task demands influencing strategy choices and evaluative standards).

Standards consist of a “multivariate profile of attributes” (Winne & Hadwin, 1998, p. 281) generated as a product of task definition and refined during goal setting and planning. Standards define the optimal end state of the current phase in operation and can include both beliefs and metrics. In Figure 1, standards are represented as a bar graph with five different levels, which might incorporate a mental representation of the completed task product, as well as beliefs about how long completion of the task should take, the level of understanding to be achieved, etc.

Operations are cognitive processes enacted in working memory that transform information. They comprise innate processes that are “primitive” in the sense of not allowing further decomposition into subprocesses, such as searching, monitoring, and rehearsing; as well as acquired processes that are more complex, i.e. tactics and strategies. Operations occur in each phase of the model and lead to products, which are then compared to standards through monitoring.

This monitoring, which occurs in each phase of the cycle, is “the pivot on which SRL turns” (Winne, 2001, p. 164). Comparing phase products to standards in turn generates further products in the form of evaluations, which are the basis of self-regulation. If an evaluation indicates discrepancies between products and standards, metacognitive control might be enacted over operations to redefine the task, manipulate conditions, revise goals and standards, refine the products, or possibly abandon the task altogether. In this way, products in any stage can lead to updates in any other, and so the model is a “recursive, weakly sequenced system” (Winne & Hadwin, 1998, p. 281).

Adaptation is the final (and only optional) stage of the model and is characterized by an overall metacognitive evaluation of the products of learning, in which learners may decide whether or not to update beliefs, motivation, standards, or strategies. It is thus where major self-regulatory developments occur.
Task-definition problems in SRL

In Winne and Hadwin’s model, then, task definition is crucial in determining the extent to which learners are able to identify appropriate goals and standards, select suitable strategies, accurately monitor phase products, and make adaptive changes to achieve task outcomes. Problems in task definition arising from task or cognitive conditions can thus have knock-on effects throughout the cycle.

One potential source of such problems is the cognitive architecture itself. The capacity of working memory is limited (Baddeley & Logie, 1999; Sweller, van Merriënboer, & Paas, 1998). Domain knowledge in the form of schemata, i.e. concepts and propositions subsumed into “Chunks”, as well as automatized procedures developed through skill acquisition, can help reduce demands on working memory (Sweller et al., 1998), but when such domain knowledge is inadequate or missing, learner’s processing capacities may be overtaxed, causing them to selectively attend to certain task cues over others, particularly those that are, or appear to be, familiar (Winne, 2001).

Task complexity and difficulty are also factors. If a task unfolds over multiple stages, learners’ working memories may be overloaded by the accumulation of information from each stage, or they may be unable to sustain the motivation needed to monitor and self-regulate across stages (Jamieson-Noel, 2004). Furthermore, learners with a performance goal orientation (as opposed to a mastery or learning goal orientation) have been shown to perform worse, and to express less satisfaction with their performance, on tasks perceived as difficult (Steele-Johnson, Beauregard, Hoover, & Schmidt, 2000). Performance goals are inversely related to the use of deeper cognitive strategies often required by difficult or complex tasks, and learners with such motivational profiles are also less adaptive in their learning (Pintrich, 2000).

As noted, adaptation in SRL depends on monitoring, which ideally should detect task-definition problems. For example, an L2 learner who initially defined a vocabulary matching exercise as easy enough to complete on the basis of existing knowledge might discover through comprehension monitoring that, in fact, too many of the words are unfamiliar, and thus would redefine the task as one requiring use of a dictionary. However, if the standards adopted for evaluation are inappropriate because the original task definition was faulty – for example, because domain knowledge required to help classify the task is missing – monitoring may not
detect problems. Difficulties in task definition, then, may be detected by measuring the accuracy of monitoring (although inaccurate monitoring may be attributable to other causes as well).

The accuracy of monitoring has long been a focus of research in cognitive and educational psychology (Ackerman & Wolman, 2007). One common method is to have research participants self-assess their performance on an ability measure (Nietfeld & Schraw, 2002), for example, through post-diction, i.e. indicating the score they believe they achieved, and then comparing the post-dicted and actual scores. Research has shown that expertise in a particular domain tends to increase monitoring accuracy (Björkman, 1992; Nietfeld & Schraw, 2002), while among subjects who lack knowledge or skills, miscalibrations between actual and expected performance tend to be considerable in the direction of overconfidence (Lichtenstein & Fischhoff, 1977). To date, however, little research has looked into the connections between monitoring accuracy and task definition (Pieschl, 2009).

In SRL research generally, broad understanding of the relationships between the major constructs has been greatly advanced through quantitative studies, but Butler (2011) has advocated a complementary line of inquiry based on case studies to examine particular variations of these relationships, interactions between individuals and specific contexts, and dynamic change over the course of learning episodes. This is the approach taken in the present research.

THE STUDY

This study sought to determine how task and cognitive conditions affect individual L2 learners’ task definitions, and how their task definitions in turn affect their self-regulated learning. It was embedded in a larger project whose main aim was the evaluation of an automated, online course in L2 strategy instruction, such that the case studies were conducted after an experiment involving treatment and comparison groups. The instruction (described below) focused on a particular strategic approach to a single complex task, thus allowing the strategy to be held constant in the design – that is, learners in the treatment group were taught to use a particular set of strategic behaviors in addressing the task, which was itself designed to require use of these behaviors for successful completion – while the repeated measures facilitated comparison of individual performances across time.
The project as a whole exemplified the Embedded Experimental Model in the typology of mixed-methods designs proposed by Creswell and Plano Clark (2007), with the case studies taking the role of a subordinate qualitative component whose purpose is “to explain the results of the intervention or to follow up on the experiences of participants with certain types of outcomes” (2007, p. 69), although in fact the case studies involved both quantitative and qualitative data. The general ontological/epistemological stance of the researcher can best be described as postpositivist (Duff, 2008), insofar as reality is viewed as objective and external, but a variety of triangulated data sources and perspectives are deemed necessary to make warranted conjectures about it. At the same time, elements of a socio-constructivist view are implied in a case study approach presuming learners to be “agents who shape environments and whose perceptions and interpretations mediate between social influences” (Butler, 2011, p. 355).

**Context**

Data were collected over the two semesters of the 2011 academic year at Iowa State University, in a credit-bearing ESL composition course for international students who have achieved a TOEFL iBT score of 71 (or the equivalent) but whom a placement test has identified as needing additional instruction before they can enroll in regular first-year composition courses. Students come from a variety of countries but primarily mainland China, reflecting the makeup of the university’s international student population at present, which according to 2011 figures was 54% Chinese. The composition course in question has a strong technology component, with most assignments distributed and collected via a learning management system (LMS) and one class meeting per week held in a computer lab.

The strategy instruction course was a separate but parallel component of the composition course, and was delivered in a separate LMS. The author was nominally the strategy instructor, although the course was automated; that is, the online materials delivered the instruction and the author’s role was to ensure smooth functioning of the technology, assist with problems, announce new assignments and send reminders about deadlines. Thus, during the strategy instruction, most interaction with participants occurred through email, although the researcher did visit each section of the composition course several times to conduct orientations, recruit participants, and oversee administration of the various instruments involved in the experiment. Once the
investigation moved to the case study phase, however, most interactions with participants were conducted face-to-face.

Participants

As shown in Table 1, two women and two men were selected for the present study. They were part of larger case study groups recruited in each semester of the investigation (Spring 2011, \( N = 6 \); Fall 2011, \( N = 10 \)) according to the principle of maximum variation (Teddlie & Tashakkori, 2008) to capture a range of possibilities in the dimensions of interest and to allow a diversity of comparisons. The four individuals described in this report ranged in age from 19 to 50 and had sharply contrasting amounts of formal English education and time spent in the U.S. Two were Chinese, one Korean and one from an African country (not specified for reasons of confidentiality). They were selected for the present investigation because they represented discrepant cases, having defied expectations in a way the author thought merited further scrutiny.

Eunhee, the Korean woman, had achieved a very high score on the pre-test and a perfect score on the post-test, despite having been assigned to the comparison group. The others were all in the treatment group but their post-test scores had been lower than expected, considering the amount of effort or other behavior they had individually demonstrated. Paul, the African male, and Tsuihua, the Chinese female, had consistently achieved lower scores in the strategy instruction assignments than most of their classmates but seemed engaged and highly motivated, as evidenced by the considerable amount of time they spent working through the tutorials (documented in the LMS log files) and the emails each had sent to the researcher asking for additional attempts at individual exercises. Min, the Chinese male, had also written to the researcher but this was to complain about the amount of work required in the course and to propose changing the format of one exercise to allow higher scores, which seemed to be a major concern for him. A review of screen-capture footage of these three participants’ post-test performances revealed behavior that contradicted important elements of the strategy instruction they had just completed, suggesting problems with task definition.
### Table 6 Case study participant information

<table>
<thead>
<tr>
<th>Participant*</th>
<th>Term</th>
<th>Sex</th>
<th>Age</th>
<th>L1</th>
<th>Years of English instruction</th>
<th>Time in U.S.</th>
<th>Meetings with the researcher</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eunhee</td>
<td>Fall 2011</td>
<td>F</td>
<td>21</td>
<td>Korean</td>
<td>9</td>
<td>2 mths</td>
<td>3</td>
</tr>
<tr>
<td>Min</td>
<td>Spring 2011</td>
<td>M</td>
<td>20</td>
<td>Mandarin Chinese</td>
<td>8</td>
<td>2 mths</td>
<td>1</td>
</tr>
<tr>
<td>Paul</td>
<td>Fall 2011</td>
<td>M</td>
<td>19</td>
<td>an African language†</td>
<td>3</td>
<td>3 years</td>
<td>2</td>
</tr>
<tr>
<td>Tsuihua</td>
<td>Fall 2011</td>
<td>F</td>
<td>50</td>
<td>Mandarin Chinese</td>
<td>2</td>
<td>10 years</td>
<td>2</td>
</tr>
</tbody>
</table>

Note: * Names are pseudonyms. † Language not specified for reasons of confidentiality

### Biodata and motivation questionnaire

This questionnaire collected biodata from the participants and also included 67 motivation-related items taken from a collection of larger instruments developed by Taguchi, Magid, and Papi (2009), based on Dörnyei’s L2 Motivational Self System (2005). The instrument was designed to measure motivation among university-level learners of English in China, Iran and Japan, so a number of questions had to be rewritten to suit an ESL context. In addition, all questions were translated into Chinese, Korean and Arabic to enhance comprehension by the types of student who typically take the course and to minimize survey fatigue. The translations, which were written and reviewed by a team of 2-4 native speakers of each language, appear together in each item, with the possible responses provided in a drop-down menu, as shown in Figure 2. The response choices were either Likert-scale items (ranging from *Strongly disagree* to *Strongly agree*) or question-type items (ranging from *Not at all* to *Very much*). Both types consisted of six-point scales because Asian students, who usually make up the majority of course participants, tend to select the middle option in odd-numbered scales (Chen, Lee, & Stevenson, 1995) possibly to avoid making choices (Dörnyei & Taguchi, 2003). Reliability for each of the eight multi-item scales is reported in Table 2.
Table 7 Reliability for the eight subscales of the motivation questionnaire

<table>
<thead>
<tr>
<th>Scale</th>
<th>Number of items</th>
<th>Reliability (Cronbach’s alpha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attitudes toward L2 community</td>
<td>4</td>
<td>.714</td>
</tr>
<tr>
<td>Attitudes toward learning English</td>
<td>4</td>
<td>.811</td>
</tr>
<tr>
<td>Criterion measures</td>
<td>6</td>
<td>.674</td>
</tr>
<tr>
<td>English anxiety</td>
<td>4</td>
<td>.795</td>
</tr>
<tr>
<td>Ideal L2 self</td>
<td>6</td>
<td>.813</td>
</tr>
<tr>
<td>Instrumentality - prevention</td>
<td>5</td>
<td>.775</td>
</tr>
<tr>
<td>Instrumentality - promotion</td>
<td>7</td>
<td>.740</td>
</tr>
<tr>
<td>Ought-to L2 self</td>
<td>7</td>
<td>.818</td>
</tr>
</tbody>
</table>

Note: Number of respondents: 100

Pattern identification and correction task

The study focused on a vocabulary usage task that was designed to be complex and unfamiliar to the participants. Evidence of these qualities is offered in the following detailed example.

On a web page contained in an LMS (Figure 3), learners are presented with the following sentence:

*In the 1950s, American scientists warned people for the harmful effects of television.*
In both written and video-based instructions, they have been told the sentence contains a content word (i.e. verb, noun, adjective or adverb) used incorrectly; specifically, there is a problem with the pattern of the content (or “key”) word that must be identified and corrected – in this case, the verb *warn*. The learners are to copy and paste the sentence into an editing box, make an appropriate correction, and submit their answer. They are told not to change the key word in any way but to focus on the words around it, adding, deleting or changing them as needed but without altering the meaning of the sentence. A table of links to various online dictionaries is provided, and learners are instructed to use these, or other online resources they are familiar with, to help them.

![Figure 3](image.jpg)

**Figure 3** Item from the pattern identification and correction task showing instructions, reference selection, sentence to be corrected, and editing box

While they could attempt a correction using existing knowledge, these sentences have been specially selected to contain a lexical pattern that will be unfamiliar to most of them, so the appropriate course of action is to select one of the pedagogical dictionaries for learners of English, e.g. the Longman Dictionary of Contemporary English (LDOCE) – since the other dictionary types provided are not reliable sources of this sort of usage information – and find the entry for *warn*, as shown in Figure 4. After identifying the numbered sense of the word that
corresponds to the usage in the sentence, they should select an appropriate syntactic pattern from among those provided in bold (in this case, warn somebody about something or of something) and then return to the task page and make the required changes.

![Figure 4 Entry for warn, with lexical patterns in bold, in the online version of the Longman Dictionary of Contemporary English (www.ldoceonline.com)](image)

The task is complex insofar as successful completion involves the integration and coordination of the following domain-specific knowledge and skills:

- reading/decoding;
- possessing functional and conceptual knowledge of lexical patterns, i.e. patterns of transitivity, complementation, and prepositional collocation;
- syntactic parsing to determine word and phrase boundaries and identify possible lexical patterns;
• identifying the word class of sentence constituents;

• distinguishing among dictionary types and selecting the appropriate type for one’s purpose (e.g. a learner’s dictionary for encoding);

• applying pattern information from learner’s dictionaries to new contexts, with contextually appropriate changes to form.

It was anticipated that the task would be unfamiliar to the target group of learners because lexical patterning is typically taught in an ad hoc way, if at all (Hunston & Francis, 1998; Hunston, Francis, & Manning, 1997), and because dictionary-use studies have shown learners in general to lack many of the skills involved in exploiting pedagogical references for usage information (see Lew, 2011, for an overview).

Measures of domain knowledge and monitoring accuracy

Vocabulary Levels Test (VLT). Vocabulary size was measured using a more recent version (Schmitt, Schmitt, & Clapham, 2001) of the Vocabulary Levels Test (Nation, 1990) to reduce the likelihood that participants had previous experience of particular test items. This version was adapted for online delivery in the project LMS and included four levels based on frequency bands ranging from high to medium to low, with an additional level addressing words that are common in academic discourse. Thirty words are tested at each level, which are in turn totaled for a maximum possible score of 120. Reliability for the instrument is reported in Table 3. Participants’ post-dictated scores on the VLT were collected via survey-type questions immediately following each word-level page in the assessment and then summed across the four frequency levels.

Table 8 Reliability for the online version of the Vocabulary Levels Test

<table>
<thead>
<tr>
<th>Level</th>
<th>Number of items</th>
<th>Cronbach’s alpha</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>30</td>
<td>.857</td>
</tr>
<tr>
<td>3000</td>
<td>30</td>
<td>.878</td>
</tr>
<tr>
<td>5000</td>
<td>30</td>
<td>.866</td>
</tr>
<tr>
<td>AWL</td>
<td>30</td>
<td>.923</td>
</tr>
<tr>
<td>Test as a whole</td>
<td>120</td>
<td>.964</td>
</tr>
</tbody>
</table>

*Note: Number of test-takers: 64*
**Pattern Identification and Correction Test (PICT).** This consisted of 10 instances of the pattern identification and correction task described above, presented in random order, with a 20-minute time limit. The sentences were found in the Longman Learners Corpus by searching high-frequency node words addressed in pattern-grammar references (Francis, Hunston, & Manning, 1996, 1998). The sentences were edited to make the context clear and to isolate the usage error as the only infelicity. A list of acceptable corrections was compiled using information from pedagogical ELT dictionaries and the Corpus of Contemporary American English (Davies, 2008), and by surveying eight native speakers of English who have master’s degrees or higher in TESOL or applied linguistics and three or more years of ESL/EFL teaching experience. The items dealt with verb, noun and adjective patterns. Most corrections involved changes to prepositions either by substitution, addition or removal. Some substitutions involved complement clauses (e.g. an *-ing* form or *that*-clause). In some cases, different types of correction were possible.

For scoring, one point was awarded if any change was made to the grammatical component of the pattern (i.e. any part other than the lexical node word), for example, by substituting a preposition. An additional point was awarded if a change matched an item in a list of acceptable responses. The points achieved on the 10 items were summed to obtain each participant’s score out of a possible 20. To determine reliability, Cronbach’s alpha was calculated for both the pre-test and post-test administrations of the instrument using combined scores from the treatment and comparison groups ($N = 64$). The combined pre-test data returned a value of $\alpha = .744$, while the post-test value was $\alpha = .865$.

Post-dicted scores on the PICT were collected via a single-question online survey instrument, which participants were directed to immediately after completing the assessment. Instructions at the top of the survey form explained the grading scheme and told participants to select a particular score from a range of 0-20 to represent how well they thought they had performed.

**Interventions**

**Strategy instruction condition.** Participants in this condition were provided with strategy instruction aimed at facilitating success in the pattern identification and correction task. The instruction took shape in an automated, online course called VVT (*Virtual Vocabulary Trainer*), which lasted five weeks and consisted of 11 tutorials organized around the topic of learner’s
dictionaries, vocabulary depth of knowledge, and lexical patterning. (The latter two topics were given the titles *Deep Vocabulary Knowledge* and *Word Patterns*, respectively, to render them more pedagogically friendly). The course is based on instructional design principles that aim to 1) reduce the burdens on L2 learners’ cognitive processing capacities while using monolingual learner’s dictionaries, and 2) differentiate appropriately among instruction targeting domain-relevant declarative knowledge (e.g. conceptual understanding of lexical patterns), procedural knowledge (e.g. syntactic processing of sentences to identify possible lexical patterns), or the integration and coordination of the two. It was developed as part of a separate study to investigate the feasibility of an automated, online form of L2 strategy instruction (see Ranalli, in preparation).

**Comparison condition.** Participants in this condition were assigned a recurring activity in the project LMS requiring them to add entries to a personal vocabulary database over the five-week period. The activity was an online version of the paper-based task that normally constitutes the vocabulary component of the composition course in question. Learners identify words in their course readings or other sources that they want to learn, then research and record lexical features and other information they think will help in using and remembering these words. The database entries were graded each week but the researcher gave no feedback regarding usage information of the type addressed in the pattern identification and correction task.

**General procedures**

Parallel procedures were used across semesters and case studies, with minor exceptions (see below). In the experimental phase, students in each section of the composition course were recruited as research participants and registered in the project LMS. In the second week of the term, on a regularly scheduled computer lab day, they completed the pretest administration of the PICT and post-dicted their score, then were assigned the motivation/biodata questionnaire for homework. The following week, they took the VLT, again on a regularly scheduled computer lab day, and post-dicted their scores. On the basis of the VLT scores, they were assigned to treatment and comparison groups using a matching procedure described in Dörnyei (2007, p. 118). Over the next five weeks, the treatment group worked their way through the strategy instruction while the comparison group added words to the vocabulary database. In the ninth week, again on a computer lab day, the participants completed the post-test version of the PICT
and post-dicted their score. This completed the experimental phase of the project. In the weeks that followed, the two groups switched treatments so the comparison group could get a chance to work through the strategy instruction. Within one week of the post-test, participants identified as potential case studies were contacted by email, and a mutually convenient time was arranged within the following two weeks to meet those who agreed to participate. All case study work was conducted in a private studio designed for web usability studies and digital media production. At the end of the initial thinkaloud/interview session, the case study participants were thanked and given a gift card as a sign of appreciation. Follow-on meetings arranged with three of the participants took place in the same studio for recording purposes and so that screen-capture footage could be reviewed as needed.

**Thinkaloud/interview procedure**

The pattern identification and correction task served as the basis for all but one of the thinkalouds. In contrast to the implementation of the task in the PICT, however, only five items were included in the thinkaloud version and participants could take as long as they wanted. The sentences in the thinkaloud version were also slightly longer than those in the pre-test/post-test version because they contained a distractor pattern in addition to the erroneous pattern. While working on the task, the individual participant was seated at a computer equipped with screen-capture software and a microphone. The researcher sat nearby at a comfortable distance, where he could observe and help if necessary. The thinkaloud protocol, which followed the procedure described in Mackey and Gass (2005), was condensed into a set of printed instructions that were given to the participant to read. The researcher then modeled the procedure using a sample item.

While the participant worked through the task, the researcher made observations and wrote questions for the follow-on interview. Immediately after the participant had finished and post-dicted his or her score, the screen capture software was stopped and audio recording was started. The researcher and participant then went back through the thinkaloud video and addressed any points the researcher had noted. The interviews were allowed to naturally segue from the particulars of the task, to the students’ experience with the instructional treatments, to wider issues of language learning, motivation, etc. They were structured with the help of concept maps prepared beforehand for each participant using a software tool.
called MindManager. The maps contained their responses on the motivation/biodata questionnaire, with interesting or puzzling responses highlighted by the researcher in advance, as well as ad hoc questions arising from previous email correspondence or performance records in the strategy instruction. The screen-capture video and audio files were later imported into Transana, a qualitative software tool, where they were transcribed.

**Analysis**

The case study analysis employed Miles and Huberman’s (1984) recommended procedure based on three concurrent activities: data reduction, data display, and conclusion drawing and verifying. The unit of analysis was the individual participant, so each case was bounded by the individual and by the duration of his or her participation in the research. Data reduction involved recursive cycles of thematic analysis with constant reference to Winne and Hadwin’s SRL model, alternating between written and audio forms of the data to allow identification of themes or patterns that might be missed if only one channel were employed. Data displays involved case-oriented and variable-oriented matrices, as well as concept maps, which were used to abstract themes and facilitate cross-case analyses. In addition to induction and deduction, the researcher also used the logical strategy of abduction (Agar, 1999) to generate premises offering plausible explanations for the unexpected results observed in each case. Conclusion drawing and verifying were accomplished using analytic memos (Esterberg, 2002) written for each participant, which also helped in further developing and exploring themes, making connections among the various data sources, and preparing the data for reporting. To warrant the assertions resulting from the analysis and enhance trustworthiness (Merriam, 2009), the study employed triangulation, searches for data that contradicted emerging themes, and verification with participants in follow-on interviews, in cases where these could be arranged.

**RESULTS**

Results from the quantitative measures of motivation, domain knowledge, and monitoring accuracy will be presented first. These will then be referenced in narrative accounts of each participant’s thinkaloud and interview.
Motivation

Motivation for each participant, along with averages and standard errors for the combined Spring and Fall 2011 data (N = 100), are reported in Figure 5. Eunhee emerges as the most highly motivated overall, tying or topping the other case study participants on most scales and ranking above the group average on nearly all. Particularly striking is her high self-reported English anxiety, which relates to nervousness speaking with native speakers and fear of making mistakes. Also notable are her high positive attitudes toward learning English and her high ratings for ideal and ought-to L2 selves. Paul also emerges as highly motivated, though less anxious about his English than Eunhee. His positive attitudes toward learning English and toward the L2 community both top out the scales and surpass the standard deviations for the group, while his ratings for ideal L2 self are slightly lower than the cohort and group average. Tsuihua could be characterized as lower in motivation than Paul and Eunhee, though still high on many subscales compared to the group averages. The marked contrast between her ideal and ought-to L2 selves suggests her motivation is based much more on an internal desire to be proficient in English than any pressure from perceived social norms. Min stands out as the least motivated, with self-ratings falling below the cohort and group averages on most scales, although most are still in the upper halves of each scale. Like Tsuihua, his motivation seems more internally than externally driven, and he reports little anxiety about his English. His relatively low score on the “Instrumentality - prevention” scale is somewhat surprising given the concern he expressed for achieving high enough grades in the VVT course in email to the researcher.
Figure 5 Motivation across eight subscales, including group average ($N = 100$) and standard deviation bars.
Actual and post-dicted performance on domain knowledge measures

Actual and post-dicted scores on the VLT are visually represented in Figure 6, with the identity line representing perfect calibration and the group average representing the combination of Spring and Fall 2011 semester participants (N = 101). Eunhee had the largest vocabulary size of the case study cohort and demonstrated knowledge of considerably more words than the group as a whole (87.1). She was also perfectly calibrated, accurately post-dicting her score of 115 out of a maximum possible of 120. Tsuihua also demonstrated good receptive vocabulary knowledge with an actual score of 109, but showed considerable underestimation in her post-dicted score of 75, a miscalibration of 31%. Min’s actual score (78) was closest in line with the group average and he was relatively accurate in his post-diction (85), although this demonstrated slight overestimation in contrast to the slight average underestimation of the group (80.4). For his part, Paul showed receptive knowledge of the fewest words (74) compared to the group average, as well as a 25% overestimation in his post-diction (99).

Figure 6 Scatterplot of actual and post-dicted scores from the Vocabulary Levels Test
Actual and post-dicted scores on the pre-test administration of the PICT are visually displayed in Figure 7, with the group average representing the combined Spring and Fall 2011 groups ($N = 101$). Compared to the other performances, Eunhee’s score of 18 out of 20 is striking, as is the fact that she slightly underestimates with a post-dicted score of 16, while the other scores fall at considerable distances above the identity line. The other case study participants performed below the group average (6), although Min was the closest with a score of 5. The overestimation Min displayed in his post-diction (10) corresponded to the 50% overestimation in the average post-diction for the group (12). Tsuihua recorded the lowest actual score (1), while her post-diction (15) represented a 93% overestimation. The difference between Paul’s actual (2) and post-dicted (18) scores also represents a considerable overestimation of 88%. In short, Eunhee displayed the hallmarks of expert performance on the pre-test, whereas the contrasts in the others’ actual and post-dicted scores are indicative of a lack of domain and task knowledge.

![Figure 7 Scatterplot of actual and post-dicted scores from the pre-test administration of the Pattern Identification and Correction Test](image-url)
The post-test scores on the PICT are presented in Figure 8, with the Spring and Fall 2011 participants included to represent a combined strategy instruction group ($N = 51$) and a combined comparison group ($N = 50$). On the post-test, Eunhee achieved a perfect score of 20 and was perfectly calibrated in her post-diction. Min’s actual score (16) was above the group average (13.8), and whereas the group’s average post-diction (14.7) represented slight overestimation, Min had reversed the trend from the pre-test and showed slight underestimation in his post-diction (14). While Tsuihua’s actual score (9) represented a considerable improvement over her pre-test performance, it was closer in line with the comparison group average (6.96) than the strategy instruction group average; and yet her post-dicted score of 10 showed close calibration on the measure. Paul’s actual score (8) was also a significant improvement over his pre-test performance, although he too was only slightly above the comparison group mean, and his post-dicted score was wide of the mark (17).

![Figure 8](scatterplot.png)

**Figure 8** Scatterplot of performance and confidence scores from post-test administration of Pattern Identification and Correction Test
Task definition

For cohesion and clarity, the thinkaloud and interview data for each participant are integrated into the form of a narrative, with the following codes used to identify the source of data in cases where it was not obvious: $T$ and $I$ represent thinkaloud and interview, respectively; and a 1, 2 or 3 following $I$ signifies the number of the interview in sequence, in cases where more than one was conducted with a particular participant (e.g. $I2$ refers to the second interview).

Eunhee

Eunhee scored 9 out of 10 on the thinkaloud version of the pattern identification and correction task and accurately post-dicted a score of 9. She even correctly identified the item where she had lost the single point, noting, “I think I had to just delete with, but I didn’t do that” ($I1$).

Her performance was characterized by attempts to coordinate form and meaning on each item; before selecting a pattern, she took time to make sure that the sense of the word given in the dictionary entry matched the usage in the sentence. Her skills in dictionary exploitation actually went beyond those addressed in the strategy instruction when she used example sentences to accurately identify subtle differences in usage between two patterns, cooperate with and cooperate in.

Eunhee used the online version of the Macmillan dictionary, which she knew from Korea and preferred because “it can give us more correct sentence … the sentence that are really used in America” ($I1$). This was in contrast to the other case study participants, whose preference for the Longman dictionary was no doubt based on its prominence in the VVT materials. In addition to the dictionary, she also drew on other sources of knowledge, noting that she had learned one pattern, regret doing something, “when I was high schooler” ($I1$) and recalled seeing another, through the Internet, in common usage, which she asked permission to confirm with a Google search ($T$).

Eunhee demonstrated strong syntactic processing skills, parsing sentences quickly to identify constituents and at one point opting not to look up an item because she knew the adjacent preposition was not part of a phrase but could “be separately used” ($T$). She displayed
knowledge of a range of grammatical terms (adjective, noun, verb, preposition, phrase) which she used appropriately (T, I1).

Eunhee monitored her work carefully, in one case explaining that she would continue to the next item despite feeling unsure about her answer (which was correct): “I don’t find any uses where the secure is used with into … I might have to delete into and just leave it like that.” She also seemed to be monitoring the form of her thinkaloud utterances, many of which could stand as well-formed sentences; for example: “I already know the basic meaning of secure, but I don’t know how to use it with another preposition or another word.”

She said she had been unfamiliar with pattern grammar before the VVT course but that her exposure to it happened to coincide with the discovery of a related gap in her own abilities (I1). She had arrived in the U.S. three months earlier thinking she was “pretty good at English,” she said, but after I came here I had very big problem in speaking, especially using verb. You know, I memorized quite great deal of vocabulary, I think, but whenever I have to speak it, speak with that verb, I don’t know how to use it … and at that time, fortunately you, in the VVT course, you taught something like that. So, oh, I think, “It’s very good idea. So that’s why I can’t speak properly.” (I1)

Eunhee said she spent a considerable amount of her free time studying English, using the dictionary “to prepare what to say” (I2) in upcoming classes or social gatherings. She said she did indeed feel “nervous, still nervous to talk with Americans” (I2) as she had indicated on the motivation questionnaire, but in general, learning English was “not that stressful, cause I like doing that” (I1). She described herself as a fan of English-language films and an avid reader of English-language blogs “about everything. I mean, entertainment or travelling or cosmetic things” (I2).

Min

For the thinkaloud, Min completed a slightly different pattern identification and correction task than the other participants. It featured the same sentences but included point-and-click functionality that guided users to the correct answers when they needed help. Because this feedback provided information about performance, and because the exercise employed a
different scoring procedure, comparisons with the other case-study participants’ actual and postdicted scores are not instructive.

As he had on the post-test, Min displayed speed in identifying potential errors and consulting the Longman dictionary for usage information, as well as a tendency to exploit bolded patterns indiscriminately. In three cases, his lack of concern for determining whether the accompanying meaning was appropriate to the task context led to lost points. His speed and accuracy in syntactic analysis and word-class identification rivaled Eunhee’s, and his navigation of the interfaces and his keyboarding and mouse skills were among the best of all the case study participants. Despite the lack of a time limit, however, he seemed in a hurry to finish and sighed audibly more than once as he progressed through the thinkaloud.

In a few instances, Min had to look up word meanings, noting once as he did so that he would normally use a bilingual dictionary for the purpose since “I can just get the answer quickly and easily” (T). He also mentioned that, while he was using the Longman dictionary because it was recommended in the VVT course, he continued to prefer a particular Chinese-English dictionary for usage information because “it’s very awesome … it also has the iPhone version” (I).

In explaining his approach to the task, he said: “Because this is the ‘Patterns of Words’ exercise so the problem always like the in, with, of, from, things like this, and always them after the adjective or after the verb” (T). Indeed, as on the post-test, Min avoided noun lookups, even when the exercise format clearly indicated they were the node to be consulted. Rather, he looked up an adjacent verb or adjective outside the editing box – key rather than factor, earn rather than reputation – which also resulted in lost points. Asked about this later, he said, “I had never thought the problem was a noun” (I). When the researcher noted that patterning as a feature of not only verbs and adjectives but also nouns had been repeatedly emphasized in the strategy instruction, Min asserted, “But I can just try the verb and adjective. Most time I can get the correct answer” (I).

As the interview ranged more widely, Min indicated that this performance-oriented approach was not limited to the VVT assignments but extended to his other university coursework. He ascribed this attitude to Chinese university students generally, saying, “Of the most time we care about is the strategy to finish the assignment, not to truly understand or to learn something” (I).
While acknowledging learning new things from the VVT course, he felt he could not afford to spend time on the usage of individual words, especially when he could “just send my paper to my, whatever, roommate or American friends, to let them check” (I). In drawing a contrast with his preferred approach to vocabulary study, which was to “just learn the English, especially the new words, just in daily life” (I), he said the VVT course might be more useful in China or Korea, where students “just study English as one subject” and where sentence-correction exercises are a common classroom activity (I).

Min said he was taking six classes so he could “finish the university, like, as soon as possible” while achieving an acceptable grade point average “like 3.3, like B plus” (I). He saw himself living and working in the U.S. because he would need to get a good return on his parents’ investment in an American education: “If I back to China for working, I don’t think I can get my tuition back, you know, after even ten years” (I). While he wished to someday have a native-like command of English, he said, “It’s kind of impossible, and I have more important thing to do which is the course pass, get higher GPA” (I).

Paul

On the thinkaloud version of the task, Paul scored 3 out of 10 and post-dicted a score of 9, showing the same difficulty in monitoring as on the post-test. Rather than pinpointing where the error might have occurred, his approach to estimation seemed impressionistic: “I guess that I’m just, like, nine because I’m not a hundred per cent for what I did” (I).

It became obvious soon into the thinkaloud that much of Paul’s attention was devoted to decoding; all of the lookups he performed began as attempts to find the meanings of unfamiliar words. Two of his corrections resulted from these lookups, and one of these was successful. Being unsure of the meaning of identical, he found it in the LDOCE, noticed the use of identical in an example that resembled the item in the task, and substituted the preposition appropriately.

However, none of his other four corrections were connected to any of his dictionary consultations. Having satisfied himself that he understood a sentence semantically, he proceeded to look for any features that seemed incorrect, including, in two cases, comma placement. In one item, he removed the personal possessive pronoun her from The popular TV chef is noted for her
unusual methods … because “it doesn’t make sense” (T). In a sentence beginning *I’m satisfied with my major, but …* he changed it to *I’m satisfied that my major, but …* And in a third, he changed the aspect of the main verb from simple to continuous.

Paul used a limited repertoire of grammatical terms, referring to clauses as *sentences* and to all word classes as *verbs* (T). In addition, none of his corrections seemed to be based on conscious application of grammar rules but rather implicit knowledge applied in a trial-and-error approach, which he explained this way: “When I make some correction, I just go back to the sentence and read again. See if I’m OK with the sentence. Like now, I’m feel OK with that sentence so I just have to move to the next” (T).

In the final item, he looked up the correct node in the error (*… my wife is in her element*), stating that he “did not understand the meaning” (T). However, he only consulted the first two entries in the LDOCE, not scrolling down to the appropriate phrasal usage near the bottom. Instead, he returned to the task item and again removed the personal possessive pronoun, saying “I don’t know why they using *her* … Just take the *her* out. We don’t need it” (T). When asked later if *my wife is in element* now made sense, given his lookup and correction, Paul again cited his good enough standard: “It just sound, like, good for me. For some other people, maybe it’s not going to be, like, sound OK … but for me just” (I1).

In the post-thinkaloud interview, when asked why he had not used the skills he had learned on the VVT course, Paul said, “Sometimes I do, like, forgot and sometime when I go back, check my notes, and I just, oh yeah, I remember that, I should do this, I should do that” (I1). Asked to further explain why he used notes, he said, “When I’m doing the homework, I don’t want to go back, like, to see the video again, and so I’m using my notes” (I2). He cited the need to review his notes when asked about the term *word pattern*, which he could not recall despite its prominence in the VVT materials.

Paul said he had arrived in the U.S. three years earlier “without any English” (I1). He said his formal instruction had been limited to ESL classes at the public high school he attended, during which a teacher had recommended the Longman dictionary “and so we used to use it.” (I1). Before immigrating, he and his family had been living in a refugee camp for many years, where “everybody was speaking French” (I1). In addition to English and French, Paul also spoke several African languages (other than his own), which he had picked up informally. In accord
with his responses on the motivation questionnaire, he said he enjoyed learning languages because they helped different cultures to “respect each other” (I1). Yet among his friends at university, most of whom were international students, he said, “we talk in English” (I1).

Tsuihua

Tsuihua’s score on the thinkaloud version of the task was 2 out of 10, the lowest among the case study participants, and she post-dicted a score of 6. Asked how she settled on that figure, she laughed and said, “I usually guess half” (I1). She also took longer than any other participant to complete it – 31 minutes – primarily because of the amount of time she spent reading and rereading each sentence quietly to herself before or after making changes.

In the first item, it became apparent that she had forgotten what the task called for when at first she mentioned the possible need to add a definite article, and then to replace one content word with another, at which point the researcher asked if she wanted to watch the video instructions again. Despite doing so, the remainder of her performance showed she was having trouble maintaining a stable, coherent conceptualization of the task, as will be shown.

She made an appropriate type of correction to the first item, adding a preposition to create an adverbial clause, but not addressing the larger problem with the prepositional phrase in which it was embedded (… the two girls were identical with nearly in every way), having looked up every instead of identical. Word class knowledge should have assisted her here, since the strategy instruction had only addressed patterning as a feature of verbs, nouns and adjectives. Although she used these terms to describe what she was doing, Tsuihua demonstrated difficulty in properly identifying the word class of sentence constituents throughout the thinkaloud; for example, by referring to verbs as adjectives and vice versa.

The second item was the only one answered correctly. She found the node word in the error, looked it up in the LDOCE, and used pattern data to correct it, but it was information for a different sense that helped her. In the next item, the task requirements seemed to slip from her grasp as she fixated on the use of satisfied in the clause I’m satisfied with my choice of major … as a verb in the “passive tense,” despite the immediately preceding contracted form signaling its status as a predicative adjective. Then, she changed I regret to choose Harvard to I’m regreted to
choose Harvard, applying an idiosyncratic rule that verbs in different clauses of the same sentence require parallel use of tenses: “After I change it, maybe I need to use both.”

On the fourth item, she was about to exchange the adjective noted with a verb when she recalled the task constraints. “I think maybe noticed … Oh, but that will change the [content] word.” She then looked up note as a verb and, not seeing any pattern with for, deleted the preposition, leaving her answer as The popular TV chef is noted her unusual methods to combine flavors.

In the final item, she looked up the node in the error … on her element and identified the correct entry, but neglected to compare the formal realizations of the phrase, focusing only on meaning: “I think this correct because in here be in your element, just enjoy, something like that.” She then looked up awkwardly and found the adverb form, which she said was required because “feel is verb,” and so submitted her final answer as I feel awkwardly at parties ...

In a subsequent interview, Tsuihua confirmed the researcher’s observation that nearly all of her lookups were for usage purposes. “I don’t know which part wrong, how to change it, but I think I understand the meaning” (I2). Like Paul, she reported using written notes to help her remember the content of the VVT presentations: “I don’t have a good memory about that. This is when I have to stop and take notes. I take a lot of notes” (I1). Unlike Paul, however, she said she understood what patterns were, but the challenge was applying that knowledge in the task.

For me, it is difficulty to identify which part is wrong, and I didn’t have learn English step by step. I just thought, ‘Oh, I know this word meaning and that word meaning.’ Terrible! [laughs] But I don’t know which pattern (I1).

She had studied Japanese in university, and although she knew she had very limited knowledge of English grammar, she felt it was “not very complicated compared to Japanese” (I1).

Tsuihua said she and her family had been living in the same university town for 10 years. They had come so her husband could take a job as a research scientist, and when they arrived, “I just know How are you? That’s all, maybe, and later on I feel I can’t go outside, I can’t speak anything” (I1). After a while, she began taking ESL classes at a community college and then at the university, and within a few years she was able to use her professional degree from China to find work in a laboratory on campus.
Her daily English usage was limited mostly to email communication – which contained “maybe a lot of mistake, but I think they can guess” (I1) – while most of her social interactions were with other Chinese speakers in the university community. Nevertheless, she expressed a strong desire to improve her English skills because “the language limit each opportunity” (I1). She said the VVT course had benefitted her in this regard because “it encourage student to learn more” and “give us the tools how to use the dictionary” (I2).

**DISCUSSION**

In Eunhee’s case, success in accurately defining the task without the benefit of strategy instruction can be attributed to the alignment of a number of factors that primed her to adapt her knowledge and skills. First was abundant domain knowledge in the form of a large receptive vocabulary, facility in reading, familiarity with learner’s dictionaries and, it can be assumed, syntactic parsing abilities honed over years of exam-oriented English instruction in her Korean schooling. Despite the latter, she also had a strong, mastery-focused goal orientation, viewing the acquisition of English as a difficult but achievable, and pleasurable, undertaking, which required not only communicative practice but attention to form. The need for the specific types of form embodied in the task had already been identified in self-monitoring of her output, and this awareness was heightened by her daily interactions with English speakers, so that when she was exposed to the task in the pre-test, she was able to figure it out. Development of this task knowledge was further aided by her highly accurate monitoring.

By contrast, Min’s definition of the task was close to the objective standard but his inability or refusal to incorporate nouns into this internal representation affected his performance and self-regulation. He seemed both helped and hindered by what he already knew; his well-developed skills in computer use and syntactic processing aided in task completion, as did fast referencing of unfamiliar word meanings, but the task seemed to suffer by association with similar form-focused exercises Min had experienced in school and, by extension, with a language learning approach he viewed as incompatible with his current circumstances. This previous experience may also have led him to self-categorize as an expert in the task domain, which can hinder monitoring ability (Glenberg & Epstein, 1987). Without a doubt, Min’s consciously espoused performance goal orientation impeded his ability to monitor accurately in cases where nouns played a central role, despite explicit cues in the instructional materials and feedback. This
suggests views of English undergirded by entrenched domain-specific beliefs, which in other areas of education, such as science learning, have been shown to lead students to ignore contradictory evidence (Chinn & Brewer, 1993). Mastery of the task clearly lost in Min’s cost-benefit analysis when weighed against the standard of time; this aligns with research showing that strategies perceived as taking too long to bear fruit may be abandoned despite gains in learning (Carver & Scheier, 2000; Schommer, 1990).

For Paul, an accurate internal representation of the task was clearly impeded by a lack of domain-specific knowledge and skills, reflecting a very different experience of language learning as compared to his East Asian classmates. Paul’s smaller receptive vocabulary meant much of his processing capacity was given over to decoding, and he seemed to lack knowledge of how to address tasks requiring syntactic processing and the application of formal grammar rules. The fact that he did not appear to have acquired schematic or even conceptual knowledge of pattern grammar, despite taking pains to complete the strategy instruction, suggests an impoverished metalinguistic system that could not facilitate its elaboration. His idiosyncratic, meaning-oriented standards for judging grammaticality are further evidence of this. These characteristics had probably been reinforced by his successes in naturalistic language learning and the fact that his primary interactions in English were with other non-native speakers. The notable tolerance of ambiguity he demonstrated in monitoring may have been a side effect of the need to rely on a monolingual dictionary whose defining vocabulary was beyond his level of proficiency, from early in his learning. Immersion learning experiences have been found to encourage learners to define tasks as primarily communicative in nature (Elbaum, Berg, & Dodd, 1993). Whatever the cause, accurate monitoring of his formal use of lexical items, not to mention an accurate definition of the task, were thwarted and it seemed clear that, left to his own devices, Paul would have trouble learning and using patterns, despite his high motivation and self-confidence.

Finally, in Tsuihua’s case, the task seemed to come into focus at times but would then slip into a haze of poorly differentiated grammaticality concerns. Working in her favor was a relatively large receptive vocabulary, apparently developed through years of naturalistic learning, as well as a more mastery- than performance-oriented goal setting with respect to the task. It was clear, however, that her syntactic processing skills were far behind those of her younger Asian counterparts, and while her study of another language and limited experience of classroom English learning had oriented her toward applying formal rules in the task, her very provisional
knowledge of these rules, combined with her lack of automaticity in parsing, meant her cognitive resources were severely taxed. In particular, she had trouble selectively attending to those sentence constituents that conformed to the task guidelines and classifying other parts as error-free. Her lack of domain knowledge also impeded accurate monitoring, which probably contributed to her “guess half” approach to post-diction. While in her own estimation Tsuihua had benefited from the VVT course, her difficulties in syntactic processing, attentional control and monitoring suggested she would have trouble independently improving her grasp of the strategy and extending it in future applications.

Considered together, these findings provide evidence of the potential for motivational factors, particularly goal orientation and language-learning beliefs and experiences, to influence strategy selection as well as monitoring and evaluation. More importantly, they demonstrate how task definition may be affected by the dynamic interplay of domain knowledge, including linguistic knowledge in both declarative and procedural forms, operating within the restrictions of the cognitive architecture.

With respect to the field of L2 vocabulary learning, this means some learners will face challenges in self-regulated learning of vocabulary depth-of-knowledge features, even given instruction specially designed to support such learning. For students like Tsuihua, finding and exploiting usage information in learner’s dictionaries may be difficult in the absence of metalinguistic knowledge of lexical features and syntactic parsing skills. Those with smaller receptive vocabularies, like Paul, may struggle when faced with the competing demands of decoding, leaving little attentional capacity to devote to aspects of usage. And for learners like Min, long experience of meaning-oriented vocabulary learning may lead ultimately to self-defeating beliefs of having learned all that can be known about how to learn new words.

The results may also provide insights into puzzling questions from previous L2 strategy research. For example, some learners have reported finding certain strategies very useful but not actually using them (e.g. Schmitt, 1997; Zhang & Goh, 2006), which might be explained by time being the key attribute among a multivariate profile of attributes that constitutes a standard for evaluating strategy efficacy. And whereas some research has asserted that strategy use is moderated by “cultural background” (e.g. Politzer & McGroarty, 1985; Rees-Miller, 1993; Schmitt, Bird, Tseng, & Yang, 1997), such a monolithic construct begs unpacking. A task-
definition framework that can take comprehensive account of previous learning, epistemological and language-learning beliefs, individual goal orientations, contextual variables, and the nature of specific tasks, may prove useful in this regard.

There are also implications for L2 strategy instruction. A systematic review of intervention studies (Hassan et al., 2005) raised questions about the relative effectiveness of two common, contrasting methodologies: those based on simple awareness-raising techniques versus those based on more labor- and time-intensive behavior-modeling approaches. The present findings suggest the answer may rest in part on the degree to which domain-specific schemata and procedural knowledge are required in the strategy in question, and the extent to which learners have already acquired this knowledge. Designers of strategy instruction should bear this in mind when conducting task analyses to determine the content of interventions.

In addition, this investigation has demonstrated the value of a multicomponential model of self-regulated learning for studying and theorizing about L2 strategies. With its separate phase for task definition and IPT-influenced cognitive architecture in every phase, Winne and Hadwin’s model is capable of complex, integrative descriptions of self-directed L2 learning, as has been shown. Furthermore, it foregrounds dynamic interrelationships without emphasizing the role of one construct at the expense of others, as is the case with current accounts in the field that privilege metacognition (Wenden, 1998), motivation/volition (Tseng, Dörnyei, & Schmitt, 2006), or strategies themselves (Oxford, 2011). Although an SRL-based account necessarily entails complexity, this accords well with our common experience of L2 strategy use as itself a complex process.

The findings also suggest potential advantages of such a model for other areas of applied linguistics where task definition is a concern, such as task-based learning and teaching (TBLT). In particular, it could help illuminate the stage of TBLT known as *task-in-process*, in which teachers and learners enact the task designer’s original workplan and “redraw [it] in terms of their own frames and their own knowledge and experience” (Breen, 1989, p. 188). Some TBLT research into learners’ idiosyncratic responses to tasks has focused on the role of motivation (e.g. Julkunen, 2001) but, as discussed above, motivation does not operate in a vacuum. Winne and Hadwin’s model might support analysis of motivation in dynamic relation to other key variables in TBLT, such as attention and planning time.
Learners’ divergent framing of tasks also suggests potential applications of the model to L2 assessment, in which some case-study research (Spence-Brown, 2001) has shown how a purportedly authentic task, when perceived by learners as a form of assessment, caused them to adopt strategies rendering it inauthentic. The model’s advantages in terms of understanding learners’ responses to external evaluation and feedback might be used to provide theoretical support to the burgeoning areas of formative and classroom-based assessment, where a lack of theorization and reliance on traditional psychometric testing theories has made it difficult to develop a research agenda and synthesize studies (Huang, 2010).

Finally, because they provide an important complement to generalized research based on self-report data or experimental findings, more microanalytic case studies of L2 task definition are warranted. Such studies would benefit by use of validated measures of additional variables implicated in the present research, such as syntactic processing ability and working memory capacity. Also, the task employed here was unfamiliar and complex, which does not describe many other tasks typical of L2 learning, so future research might profitably incorporate task familiarity and task complexity as variables.

CONCLUSION

It seems fair to say that, only at the point where task-strategy mismatches crop up do we realize how much of our work in L2 instruction, as well as L2 assessment and research, is based on the assumption that learners will see the tasks we design the same way we do. It would seem important, then, to be more mindful of the potential for wayward task definitions, particularly in those aspects of L2 learning, such as vocabulary acquisition, where learners are largely left to fend for themselves.

NOTES

1. While task is typically used in second language acquisition to mean a learning activity involving a primary focus on meaning and a lack of restrictions on learners’ linguistic choices (see Ellis, 2003), the conceptualization adopted here is Breen’s much more inclusive notion of “a structured plan for the provision of opportunities for the refinement of knowledge and capabilities entailed in a new language and its use” (1989, p. 187).
2. Interestingly, Elbaum et al. (1993) had empirically documented links between language learning beliefs and learners’ task definitions, but this study is not cited by Wenden, perhaps because it does not position beliefs within a metacognitive framework.

3. According to Dörnyei (2005), the *ideal L2 self* refers to an idealized version of one’s self who possesses the ability to speak the L2, and can thus represent a positive motivating influence to the degree that the individual seeks to reduce discrepancies between the actual and the ideal selves. The *ought-to L2 self*, by contrast, refers to L2-related attributes one feels one should possess “in order to avoid possible negative outcomes” (2005, p. 106).

REFERENCES


Lichtenstein, S., & Fischhoff, B. (1977). Do those who know more also know more about how much they know? *Organizational Behavior and Human Performance, 20*(2), 159-183. doi: 10.1016/0030-5073(77)90001-0


CHAPTER 6: CONCLUSION

This dissertation has presented four separate papers involving an online instructional resource called VVT (Virtual Vocabulary Trainer) designed to teach learning strategies that support second language (L2) vocabulary learning, and to provide a platform for research into broader, related issues. The papers, targeted at peer-reviewed journals in the fields of computer-assisted language learning (CALL), applied linguistics, and teaching English as a second or foreign language, make contributions that can be divided into the practical, pedagogical, theoretical, and methodological.

On a practical level, the project has provided empirical evidence to support the popular conception that learning to use a word is more difficult than learning what a word means. The findings show that while tertiary-level ESL learners generally have the metacognitive tools and referencing skills needed to self-direct their acquisition of simple form-meaning pairings, this does not appear to be the case with respect to learning and use of certain vocabulary depth-of-knowledge features, namely lexical patterns. As has been shown, many learners simply do not know they lack such knowledge, which impedes the chances of self-directed remedial action. And while some vocabulary experts have argued that the way forward lies in intensive teaching of individual words, the success of the instructional approach described here illustrates the potential of a complementary, or alternative, methodology based on metalinguistic concepts, strategies, and referencing skills.

The dissertation has also demonstrated the potential of automated, online materials to deliver some aspects of strategy instruction (SI), which could shift the cost-benefit balance toward greater inclusion of SI in syllabuses and language programs. It has suggested principles for the design of SI interventions, whether online or teacher-led, based on Cognitive Load Theory, Multimedia Learning Theory, and the notion of complex cognitive skills, in recognition of the considerable processing demands involved in some aspects of L2 learning. At the same time, the project has met calls in recent lexicographical research for dictionary training that integrates consultation skills with language awareness, and has shown how such training can be accomplished cost-effectively.
In terms of theory, the study reviewed an important new model of L2 vocabulary learning, using empirical data to show the type of elaborations that would be needed for that model to function adaptively, as envisioned by its authors; namely, explicit feedback mechanisms connecting evaluation variables to those representing strategic behavior. Without feedback supported by accurate internal monitoring or use of L2 lexical referencing tools, the model in Tseng and Schmitt (2008) cannot adequately explain why learners may face greater challenges from depth-related versus size-related learning, nor how they might be able to bootstrap their way into strategic mastery. In addition, the study has demonstrated the value of a multicomponential model of self-regulated learning (SRL) for researching and theorizing about L2 vocabulary acquisition. The SRL model proposed by Winne and Hadwin (1998) was shown to facilitate complex, integrative descriptions of the effects of strategy instruction by virtue of its separate phase for task definition, by taking cognitive processing limitations into account, and by foregrounding dynamic interrelationships without overemphasizing the role of any single construct.

Finally, the study has shown how technology can be used to combine some of the primary advantages of laboratory- and classroom-based research, while also further demonstrating the value of microanalytic case studies for investigating strategic behavior through the triangulation of a variety of data types, including task products.

Because the study did not include tests of transfer or maintenance, these are now research priorities, and work in this vein should include efforts to identify complementary combinations of online versus teacher-led forms of SI that take best advantage of both, particularly the latter’s capacity to address individual needs.

Future work should be also conducted with other types of technology-mediated strategies that lend themselves to this form of SI, especially projects integrating metalinguistic knowledge and computer-based reference tools, such as corpora. In addition, learning strategy research, and possibly other areas of L2 studies, will benefit from greater focus on task definition and the potential for learners to develop idiosyncratic task representations. Future research in this direction should investigate the role of schematic or conceptual knowledge in determining the accuracy of both task definition and monitoring, as well as the effects of task complexity and familiarity on these cognitive processes.
The VVT course itself may provide a platform for addressing many of these objectives, and thus it represents the potential for software technologies to simultaneously serve the needs of research and instruction, as discussed in Winne (2006). As noted, there is scope for expanding the project to incorporate other aspects of vocabulary, both depth- and size-focused, within an SRL framework, and in the process to facilitate multiple, simultaneous lines of inquiry into the complex ecology of self-directed L2 vocabulary learning.

REFERENCES

