Article

LANGUAGE
TEACHING
RESEARCH

Task-based language teaching: An empirical study of task transfer

Language Teaching Research
2016, Vol. 20(3) 341–365

DOI: 10.1177/1362168815569829
ltr.sagepub.com
SAGE

Susan D Benson
University of Maryland, USA

Abstract

Since the 1980s, task-based language teaching (TBLT) has enjoyed considerable interest from researchers of second language acquisition (SLA), resulting in a growing body of empirical evidence to support how and to what extent this approach can promote language learning. Although transferability and generalizability are critical assumptions for TBLT, there is little empirical evidence that task-related language abilities are indeed transferable. The current study was conducted to address this need for empirical research on generalizability and transfer critical for the planning of teaching and assessment of learning by specifically investigating whether or not transfer occurs between two similar pedagogic tasks. Fifty-three randomly assigned low-level adult second language learners were trained in a computer lab to complete one of two pedagogic tasks or no task, after which all participants were tested on two transfer tasks. Although the results of a MANCOVA did not provide statistically significant evidence of transfer, a post hoc analysis on a subset of the lowest proficiency learners suggests that task-related language abilities are transferable. Directions for future research and implications for teaching and assessment are discussed in light of the present findings.

Keywords

Generalizability, task-based, task selection, TBLT, transfer

I Introduction

Since the 1980s, task-based language teaching (TBLT) has enjoyed considerable interest from researchers of second language acquisition (SLA), resulting in a growing body of empirical evidence to support how and to what extent this approach can promote language learning. Yet, as noted by Long (2007) and others (e.g. Norris, Brown, Hudson, & Yoshioka, 1998; Robinson, 2009), there are still challenges and areas in need of further

Corresponding author:
Susan D Benson, University of Maryland, 3241 13th Street North, St. Petersburg, FL 33704, USA.
Email: sdb@umd.edu

research when task is the unit of analysis. Two key issues in TBLT are task selection and sequencing for both teaching and assessment. Most of the research concerning task selection and sequencing has focused on task complexity (e.g. Robinson, 2001, 2007, 2009; Skehan, 1996, 1998); however, another issue concerning selection is transfer of learning, or to what extent performance on an assessment task or pedagogic task can be predicted to transfer or generalize to other tasks. Transfer of learning is difficult to research and document, yet determining how much learning will transfer is critical for the planning of teaching and the assessment of learning. Course designers, for instance, need to know which tasks to teach and to test, and whether or not assessment results can be generalized without testing every individual task.

TBLT promotes learning by doing. 'It aims to equip learners to meet their present or future real-world communicative needs, as identified through a task-based learner needs analysis' (Long, 2007, p. 129). In a task-based approach, students learn language by doing relevant, engaging, and hands-on pedagogic tasks that stimulate their interest and keep their attention. Language itself is not studied as object.

Tasks can be further defined as either target or pedagogic. 'Target tasks' are identified via a needs analysis and are specific to a group of learners or an individual. They are real-world activities that students should be able to perform outside of the classroom, such as filling out a form or making a reservation, whereas 'pedagogic tasks' are the less complex versions practiced in the classroom that lead up to the target task (Long and Crookes, 1992).

Long and Crookes (1993) describe how target tasks can be combined and classified into more general task types. For example, they suggest that the target tasks 'serving breakfast, serving lunch, serving dinner, and serving snacks and refreshments', might be classified into 'serving food and beverages'. After tasks are categorized, they can be broken down and sequenced into pedagogic tasks that progress from simple to more complex to create a task-based syllabus (Long & Crookes, 1993). The question of how many tasks from each task type need to be taught or assessed to predict future task performance remains unanswered, and the boundaries for predicting to what extent one task or task type will transfer to a novel task or task type are elusive. The present study was designed to advance the TBLT research agenda by investigating whether or not transfer occurs between two similar pedagogic tasks in a controlled computer-based environment.

II Literature review

1 Task selection and sequencing

Careful consideration of task complexity and similarity to novel tasks may inform predictions of task transfer and generalizability; this in turn may inform the selection and sequencing of both target and pedagogic tasks. Sequencing is a problem for both synthetic and analytic syllabi. As noted by Long (e.g. 2007), for synthetic, grammar-oriented syllabi, materials are generally organized using the designer's intuition about morphosyntactic and phonological linguistic complexity or by lexical frequency for corpus-based materials. In comparison, task-based approaches generally integrate both a content rationale, for example, groupings based on thematic relationships, and a complexity

rationale, where tasks are organized from least to most complex (Norris, 2009). Sequencing decisions for TBLT are made using first-hand knowledge about the learner population, as well as an understanding of the communication needs and acquisition opportunities encompassed within diverse language use tasks (Norris, 2009). The steps in the learning process represented by pedagogic tasks can be organized from simpler to more complex by examining the complexity of a task and the task difficulty for individual learners (Long, 2007). Yet defining complexity and difficulty is a complex and difficult task in and of itself. Although several noted researchers have placed considerable attention on task sequencing and complexity, most agree that the issue is far from resolved (e.g. Brindley, 1994; Long, 1985; Robinson, 2001a, 2001b, 2009; Skehan, 1998). As Robinson (2001a, p. 292) states, 'Task complexity is represented as a series of options which can be manipulated to progressively increase the cognitive demands of pedagogic tasks, so they approach the full complexity of the target task.' Cognitive and linguistic demands must not be considered one and the same. Skehan (1998) and Robinson (2001a, 2001b) both reject sequencing based on linguistic criteria. Skehan proposes sequencing from less cognitively demanding to progressively more cognitively demanding tasks that require more attentional resources to promote 'balanced language development' in the areas of 'accuracy, fluency, and complexity during language production' (Skehan, 1998, p. 97). According to Skehan (1998), the scarcity of attentional resources suggests that tasks can lead to either greater complexity or greater accuracy of language production, but not both.

Alternatively, in a study exploring task complexity and difficulty, Robinson (2001b) showed that increasing the cognitive complexity of a direction-giving map task significantly affected language production, in that there was more lexical variety on a complex version, but greater fluency on a simpler version of the task. Robinson (2009) therefore proposes a framework for grading and sequencing by characteristics of task conditions and complexity that would lead to both accurate and complex language production. Task conditions (i.e. the interactional demands) include participation and participant variables, such as whether a task is open or closed, if the information exchange is 'one-way' or 'two-way', and whether or not agreement is required, along with other participant variables. Task complexity involves the cognitive demands of the tasks, including factors such as whether the task occurs 'here-and-now' or concerns an event in the past, how many elements are referenced (+/– a few), spatial location, and so on (for further discussion, see Robinson, 2009). For Robinson, the interactional demands of the tasks should not be graded and sequenced. Task conditions are replicated each time pedagogic task versions are performed to promote transfer of training to real-world tasks. Only the cognitive demands are graded and sequenced based on the complexity of the task (Robinson, 2009).

2 *Generalizability*

Since the initial selection of target and pedagogic tasks for TBLT is based on a needs analysis, one can assume some confidence about the relevance and the possibility of direct transfer of the abilities developed in classrooms to analogous real-world contexts. But the issue is complex. How many representative tasks should be selected for teaching or assessment?

As noted by Long (2007), transfer of learning is an important unresolved problem related to task selection and assessment and is an issue for all syllabus types, not just task-based ones.

> Little is known about how far, if at all, learners' ability to perform one task predicts their ability to perform another … How close do the training and transfer tasks need to be? Is it possible to test the underlying construct and assume that success with one or more tasks based on it indicates capacity to perform (any?) other such tasks, or must learners' performance on each target task be assessed separately? (Long, 2007, p. 130)

If it can be determined that certain target tasks will generalize, then fewer target tasks of the same type can be included in instructional materials. If research reveals that transfer does not occur, pedagogic tasks will need to be developed for each target task. So, for example, if requesting product information in a store is identified as a target task type, research must first determine if one target task, for instance, where learners request product information for buying a television, will enable a learner to perform a similar task, such as requesting information for buying a refrigerator. The Task-Based Language Assessment Literature (TBLA) outlines the problem from a testing perspective but still offers little by way of solutions to these persistent questions about generalizability.

Assessment of student learning should be centered on 'task-based criterion-referenced tests, whose focus is whether or not students can perform some task to criterion, as established by experts in the field, not their ability to complete discrete-point grammar items' (Long and Crookes, 1992, p. 45). Robinson and Ross (1996) have suggested that tests that tap 'procedural knowledge' appear to be more accurate placement tools in the academic context. For the analytic TBLT syllabus, it is the learner's role to analyse the language used in the activities or tasks, not the syllabus designers' (Robinson & Ross, 1996).

Many agree (Brown, Hudson, Norris, & Bonk, 2002; Long & Norris, 2000; Mislevy, Steinberg, & Almond, 2002; Norris, 2002; Robinson and Ross, 1996) that TBLA simply needs to evaluate how well learners perform a given target task. Performance on the assessment task itself is the construct of interest and 'an appropriate starting point for thinking about the features of language-use situations that reveal the language-use competences that are of interest' (Mislevy et al., 2002). The task itself should be used as the 'fundamental unit of analysis' for selecting test items, creating the test instrument, and rating task performance (Long & Norris, 2000). The fundamental problem, nonetheless, remains: how to define and sample tasks in order to make generalizations across tasks and make interpretations about broad ability and language use domains. A clear threat to validity for TBLA regards domain representativeness or 'the degree to which a task or a few tasks in a testing situation represent the many real world tasks that will be required later' (Norris et al., 1998, p. 25).

TBLA must have representative tasks that reflect the language use in the target domains. The boundaries of multifaceted and dynamic tasks and domains, however, need to be defined in order to predict transfer and to determine if learners can accomplish target tasks or task types that are similar to the language knowledge and abilities they have acquired by learning related or similar tasks. These issues, however, await empirical evidence.

Mislevy et al. (2002) have introduced a framework that could be used to systematize the design of TBLA. They argue that there is a lack of 'systematic means for designing performance assessments that will directly and adequately inform the particular kinds and qualities of inferences that need to be made for various assessment purposes' (Mislevy et al., 2002, p. 478). They suggest that models that specify a domain of tasks combined with the features of the task could be used to 'shed some light' on the low generalizability problem associated with TBLA.

Clearly, research is needed to test generalizability. Domain-referenced sampling and inferencing were not explicitly investigated by Norris, Brown, Hudson, and Bonk (2002), but in a study to investigate the development and use of a prototype English language task-based performance test, they reported initial evidence that holistic ability estimates were valid and could be used to make generalizations about an individual's ability within a domain of related tasks. They suggest further validation inquiry by means of a mechanism for predicting and relating holistic performance estimates to particular tasks or task types in the domain. This mechanism could be used to systematically sample from the domain, after which additional predictions could be confirmed on the basis of a second set of performance data obtained from related tasks from within the domain (Norris et al., 2002).

3 Psychology's take on transfer

The idea of 'transfer' concerns prior learning affecting new learning and can be considered in holistic terms. Transfer of learning is the connection or link between what happens in the classroom and the real world. It is the application of skills and knowledge that were learned in one situation to another similar or novel situation or setting.

Transfer of learning and training has been explored extensively in the field of psychology; nonetheless, after 100 years of research, there is little empirical evidence that transfer in fact occurs. The idea of 'transfer of practice' was originally put forth by Edward Thorndike and Robert S. Woodworth in 1901 (as cited in Blume, Ford, Baldwin, & Huang, 2010).

They explored how individuals could transfer learning in one context to another that shared similar characteristics. Their theory, the identical element theory, implied that transfer of learning depends on the proportion to which the learning task and the transfer task are similar. They predicted that transfer would occur if the goals, method, and approaches of the learning task were similar to those of the transfer task (Thorndike, 1901, as cited in Blume et al., 2010). Thorndike's theory, however, was tied to the physical world and preceded the cognitive revolution. Singley and Anderson (1989) have resurrected Thorndike's theory yet propose a cognitive or information-processing approach to the transfer of skill in which productions, condition-action rules as defined in Anderson's ACT* theory of cognitive skill, serve as the elements or basis for transfer. Productions represent the cognitive processes necessary for skilled behavior, like planning and problem decomposition. The amount of overlap between production sets determines the amount of positive transfer from one task to another (Singley & Anderson, 1989). They further

specify that both procedural and declarative knowledge can serve as either sources or targets of transfer (for further discussion, see Singley & Anderson, 1989, p. 33).

Evidence of transfer of learning, however, seems to be an enduring problem in the psychology literature (e.g. Barnett & Ceci, 2002; Grose & Briney, 1963; Haskell, 2002). Haskell (2002) claims that psychology research has unfortunately failed to capture much evidence of transfer even though transfer is the goal of all education and learning. Yet some research over the past decade has supported the generalization that transfer is more likely to occur with near transfer tasks, which are highly similar, as compared to far transfer tasks, for which the task situations and settings are quite different (Barnett and Ceci, 2002). The problem comes in defining what is 'near' and what is 'far' along the continuum.

There are many different models and taxonomies of transfer prevalent in the psychology literature, yet very little agreement. In a meta-analytic review of transfer of training, Blume et al. (2010) remark that 'scholars have operationalized transfer in various ways, and it is important to quantitatively examine how these different operationalizations influence predictor-transfer relationships' (p. 1066). Since a real-world task never repeats itself in exactly the same way or in exactly the same context, the essential problem in transfer is deciding when and how something is the same as or equivalent to something else.

The content of a task, which would include memory demands or the processes necessary for task performance, must also be considered when predicting transfer. The concept of Transfer Appropriate Processing (Morris, Bransford, & Franks, 1977) suggests that memory will be best when the processes employed during encoding match those used during retrieval. For example, if the learning task is semantic in focus, then the test or assessment task should also have a semantic focus. Schmidt and Bjork (1992) contend that in order to facilitate learning processing activities need to be considered for both the training task and the transfer task. Reviewing experiments both with motor and verbal tasks, Schmidt and Bjork (1992) further suggest that although constant practice may result in more effective performance in the acquisition phase, it may produce less effective capabilities to generalize knowledge to novel situations than does variable practice since variable practice includes a wider range of processing activities.

4 Two useful transfer taxonomies for TBLT

Gagne's (1965, as cited in Blume et al., 2010) taxonomy distinguishes between two types of processes for generalization: lateral and vertical transfer. Lateral has to do with skills that would spread over a wide variety of situations with the same level of complexity or difficulty, whereas vertical transfer has to do with an acquired skill and how it affects a more complex skill.

Barnett and Ceci (2002) have also proposed a useful framework for categorizing studies of far transfer. In their discussion they also note problems with determining how exactly to define near or far, as researchers are not consistent in the use of these terms. Therefore, Barnett and Ceci (2002) suggest breaking down the continuum of near and far into two overall factors: (1) the content, what is transferred, and (2) the context, where and when it is transferred from and to. The content includes the learned skill or procedure, the nature of the performance, and the memory demands of task transfer. They indicate that

vertical and horizontal transfer could also be considered a content dimension. The context dimension, as defined by Barnett and Ceci, includes the knowledge domain, physical context, temporal context, functional context, and social context, and the modality.

These two taxonomies are useful for considering task transfer for language teaching, as they address both task complexity and similarity of task features. For the purposes of the study reported here, complexity will remain constant, in order to explore transfer that crosses knowledge domains. In order to begin to make predictions about generalization and transfer from one pedagogic task to another and from pedagogic to real world tasks, nearness or farness of the task context concerning the knowledge domain needs to be defined. All other contextual factors being equal, it is expected that transfer from one science context to another science context is more likely than from, for example, science to humanities. In the case of TBLT, and second language acquisition, the knowledge domain would account for lexical variation.

There are thus many factors that could affect transfer, including social, individual, and cognitive variables. For TBLT, the question that has yet to be empirically tested is whether or not the ability to perform one target task will transfer to performance of another target task of the same type or characteristics within similar or different knowledge domains. From the psychology literature there is very little evidence that far transfer between knowledge domains occurs, yet two studies provide some promise. Gick and Holyoak (1980) found evidence of far transfer in a prototypical analogical transfer experiment between a military context and medical context, and Chen (1996) demonstrated that the distance between knowledge domains only impedes successful transfer in some cases where there is an interaction effect from the learned skill.

III A study of task transfer for TBLT

The purpose of the current study was to investigate whether or not transfer occurs between two similar pedagogic tasks. The tasks themselves served as the constructs, with a focus on successful task completion rather than linguistic features. This analogical transfer study involved training on one task followed by assessment of a novel task that was considered an analogue of the first. The central question addressed by the current study is: Do similar tasks of the same type and complexity transfer?

All of the tasks chosen are practical and useful for second language learners and were identified from a needs analysis conducted for a community program of English as a second language (ESL). The target tasks were classified into two target task types: *following directions to a destination* and *evaluating product information in a store*. All of the tasks for the study were receptive and designed to be as equal as possible concerning cognitive complexity, following the framework proposed by Robinson (2009), and operationalized in this study as the number of steps and elements in each task. The study aimed to test empirically whether or not students who could successfully perform one direction task could then perform another direction task in a different knowledge domain (i.e. street vs. hospital), and whether or not students who were trained to evaluate product information to buy the best television could then evaluate product information to buy the best refrigerator.

1 Research questions

1 Does learning from one pedagogic task transfer to another task of the same type and complexity?
2 Does more transfer occur between tasks that are similar in knowledge domain (near) than tasks that are dissimilar in domain (far)?

Hypotheses:

Hypothesis 1: Participants who learn the *street directions task* will outperform participants who learn the *television shopping task* on the *hospital directions assessment task*.
Hypothesis 2: Participants who learn the *television shopping task* will outperform participants who learn the *street directions task* on the *refrigerator shopping assessment task*.
Hypothesis 3: All participants who were assigned to a task treatment group will outperform the control participants.

2 Method

a Participants. Participants in the study were second language (L2) English learners of low proficiency, as determined by the Accuplacer Levels of English Proficiency Test (LOEP), and were recruited from nine intact classes at a community college in the USA that offers an intensive English for academic purposes (EAP) program for students who wish to improve their English in order to attend regular college-level courses. Participants came from 15 different first language (L1) backgrounds, with the largest numbers from Spanish (15) and Arabic (11); for the complete list of L1s and the number of participants from each, see Table 1. The students ranged in age from 18 to 64 years, with a mean age of 30.69 (sd = 10.937). Fifty-three who agreed to participate and gave informed consent were randomly assigned either to one of the two treatment groups or the control group. Technical difficulties prevented five participants from completing the protocol, reducing the total number of participants to 48.

b Materials and procedures. Materials for the training and the assessment tasks (following street directions, following directions in a hospital, buying the best television, and buying the best refrigerator) were developed based on the 10 methodological principles of TBLT, as proposed by Doughty and Long (2003). Authentic samples for each task (interactions on the street, in a hospital, buying a television, and buying a refrigerator) were recorded, transcribed, and analysed to determine prototypical characteristics of the target discourse in each case. The identified target language features were used to create authentic pedagogic tasks (for a sample breakdown of a pedagogic task, see Long, 2007, p. 129).

A 45-minute lesson for each task was developed within ANGEL, an online learning management system (LMS) platform, in order to be able to control time on task and ensure that each participant received identical exposure to the task treatment and assessment. Each treatment period began with participants seeing and hearing the task modeled

Table 1. Participants' first language.

First language	Number of participants
Spanish	15
Arabic	11
Albanian	4
Vietnamese	4
Russian	3
Portuguese	2
Chinese	1
Czech	1
French	1
Bisaya	1
Tagalog	1
Bengali	1
Thai	1
Lithuanian	1
Turkish	1
Total	48

multiple times. For example, participants in the 'following directions group' saw a real map and heard directions being given while watching an icon move slowly on the screen to follow the directions. Learners could pause or replay each video as desired. After seeing and hearing the task modeled, participants were given several opportunities to practice simple to progressively more complex versions of the task on the computer. The interactional demands for each practice task remained constant (e.g. information flow was always one-way, the solution was closed, etc.); however, the cognitive demands of the practice tasks were sequenced from simple to complex based on the number of steps needed to complete each task. The first few practice tasks included one or two steps whereas later practice tasks included four or five steps (i.e. directional moves or product information). For each set of street directions, learners were shown a map with an arrow which indicated where to begin. Learners listened to directions (e.g. walk up 14th Street and make a right on Pennsylvania, and it's right there on your right) and were then asked to indicate their location on the map by choosing from a list of destinations (e.g. U.S. Post Office, Botanical Gardens, or National Museum of the American Indian). Learners were given feedback as to whether or not they were in the correct location each time.

Similarly, the participant in the 'product information group' saw the task modeled multiple times by hearing information about products while seeing images of the products and viewing the selection process. During the modeling phase, learners in this group were shown, for example, four different televisions with descriptions underneath each that included the brand, price, size, and type (e.g. LED, LCD). Participants listened while the task of narrowing down to choose the best TV was performed. Based on the desired features, televisions were eliminated from the screen until only one remained. After seeing and hearing the task modeled, participants in this group also had a practice period where

they heard product information and were instructed to choose the best product based on size, features, etc. For each practice task, learners were shown a grouping of four televisions. In the first few practice tasks, participants were given two or three pieces of selection information (e.g. you prefer Panasonic, you'd like a 50 inch television, and you'd like to spend the least amount of money). In later practice tasks, learners were given up to four pieces of selection information. Each piece of information allowed the participant to eliminate one or two of the choices until only one remained. Participants were notified each time as to whether or not they had made the correct or incorrect choice of which television to purchase.

The control group listened to interesting stories, followed by comprehension questions, which were culled from instructional materials used by the college. The stories were presented in a similar format within the LMS so that the control participants would be unaware that they were not in a treatment group. Participants in all three groups were permitted to repeat videos and questions as many times as they liked within the 45-minute treatment period.

After the treatment phase, the students in all groups were prompted to take a 10-minute break before the assessment phase. All three groups then completed two assessment tasks, *following directions in a hospital* and *evaluating product information to buy the best refrigerator*. The format and procedures for the transfer tasks were similar to the treatment tasks; however, participants were only able to view each video and attempt each item once. Participants were randomly assigned to the task assessment order for counterbalancing. Half completed the hospital task first, and half completed the refrigerator task first.

Each assessment included 15 short variations (one item per video) of the same task (i.e. 15 sets of directions). For each item in the *following directions in a hospital* assessment, participants were shown a map from a hospital and told where to begin. They then followed directions of varying length, which included two to five directional indicators (e.g. walk through these doors, keep going straight down the hall in front of you, on your left, etc.). For each item in the *shopping for a refrigerator* assessment, participants were shown five refrigerators and listened to phrases (e.g. you prefer stainless steel, you would like an in-door ice dispenser) that allowed them to systematically eliminate the choices. For instance, if only three of the five refrigerators shown were stainless steel, knowing that stainless steel was desired allowed two of the five to be eliminated. The learner's ability to complete the tasks successfully was based on whether or not the correct option (destination or product) was chosen from a list of five choices, for a total of 15 items for each assessment. Table 2 provides treatment and assessment order for each group (for examples of each assessment task, see Appendix 1).

Prior to the study, all items for the assessment tasks were successfully piloted (with 20 ESL students in the same program) to evaluate the difficulty and similarity of the two tasks and to verify that the appropriate proficiency level had been selected for the study. Participants in the pilot study did not receive a treatment before completing the assessment tasks. Responses to all 30 items were scored as either correct or incorrect, and a range of descriptive and inferential statistics were computed, which revealed that the two tasks were of comparable difficulty. Additionally, in order to ensure that learning was not occurring during the assessments, item facility was plotted, and the mean score for items

Table 2. Experimental groups and assessment order.

	Group 1a	Group 1b	Group 2a	Group 2b	Group 3a	Group 3b
Treatment	Practice following street directions	Practice following street directions	Practice buying the best television	Practice buying the best television	Listen to stories unrelated to tasks	Listen to stories unrelated to tasks
Assessment 1	Following directions in a hospital	Buying the best refrigerator	Following directions in a hospital	Buying the best refrigerator	Following directions in a hospital	Buying the best refrigerator
Assessment 2	Buying the best refrigerator	Following directions in a hospital	Buying the best refrigerator	Following directions in a hospital	Buying the best refrigerator	Following directions in a hospital

2–8 was compared to the mean score for items 9–15 for both tasks. Qualitative data from a short exit survey (see Appendix 2), which contained background questions and three questions to gauge the students' perceptions concerning task similarity or difference, as well as task difficulty, suggested that the students found the two tasks to be very different. Upon completion of the assessment tasks in the current study, participants likewise completed the exit survey. Reliability estimates for the assessment measures were calculated using the data from the current study. Cronbach's alphas for the 15 items in the directions assessment task and the 15 items in the shopping assessment task were .75 and .74, respectively.

IV Results

The qualitative data from the exit survey provided preliminary indications concerning the similarity and difficulty of the two tasks. The students were first asked to indicate how similar the two tasks were on a scale of 0 (completely the same) to 4 (completely different). The mean for similarity was 2.44, and the median, as well as the mode, was 2 (for the frequency of each response, see Figure 1). The results of the exit survey, therefore, indicate that the students felt the two tasks, shopping and following directions, were somewhat the same, perhaps due to the similar format on the computer; however, the mean of 2.44 and Figure 1 suggest that most participants perceived the tasks as different.

To examine the participants' perception concerning the difficulty of the two tasks, the survey asked the students to rate each task on a scale of 0 (not difficult) to 4 (very difficult). The mode and median for both the *shopping for a refrigerator task* and the *directions task* was a 2 (somewhat difficult); (for the frequency of each response, see Figure 2). The mean for the shopping task was 1.77, and the mean for directions was 2.17. Thus, although the median and mode were the same for both tasks, some participants felt that the directions task was slightly more difficult than the shopping task. A paired-samples *t* test indicated that the perceived difference as measured by the mean rating for the two tasks was significant ($t(47) = 2.289$, $p = .027$).

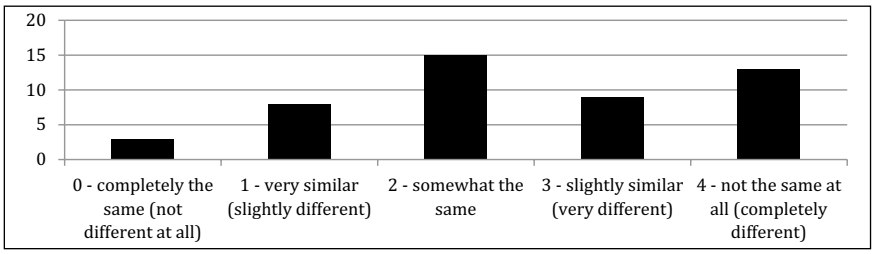


Figure 1. Similarity of tasks.

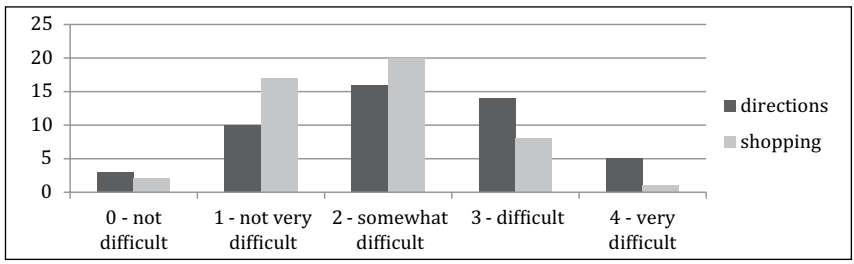


Figure 2. Difficulty of tasks.

Items for each task were scored as either correct or incorrect, and no partial credit was given. Successful completion of the transfer tasks was calculated as a percentage for each task based on the number of accurate responses. Group means and standard deviations were then calculated for each task and for the LOEP listening scores[1] (see Table 3). The mean score for each task indicates that the directions task, which was perceived to be more difficult by participants, was in fact less difficult than the shopping task.

Q–Q plots and the results of a Kolmogorov–Smirnov Test, $p > .05$, showed that the LOEP scores and data for the shopping task were normally distributed; however, the data for the directions task were somewhat positively skewed, and the results of the Kolmogorov–Smirnov Test were significant, $p = .02$.

To test the three hypotheses, which suggested that the independent variable would differentially predict the two dependent variables, a one-way MANCOVA was calculated examining the effect of training (directions, shopping, or none) on directions and shopping assessment task scores, with English proficiency as measured by the LOEP as a covariate. As expected, English proficiency was significantly related to scores on the hospital directions task ($F(1,44) = 8.511$, $p < .05$) and scores on the refrigerator shopping task ($F(1,44) = 16.326$, $p < .05$); however, the main effect of training was not significant (*Lambda* $(4,86) = .866$, $p > .05$). Neither assessment task was significantly influenced by task training. Figure 3 shows that overall, all three groups performed better on the *following directions in a hospital task* than the *shopping for a refrigerator task* regardless of training.

Table 3. Descriptive statistics for Levels of English Proficiency Test (LOEP) scores and both tasks.

	N	Minimum	Maximum	Mean	SD
LOEP	48	44	98	63.23	11.771
directions task	48	13.33	100.00	57.9163	22.68820
shopping task	48	13.33	93.33	51.3892	21.47953

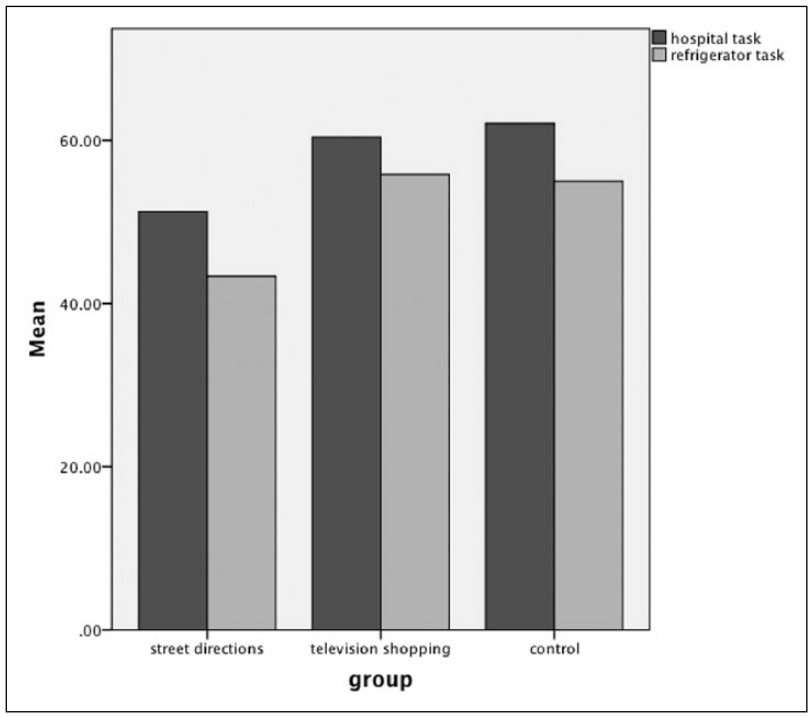


Figure 3. Mean accuracy for all participants.

Although students were randomly assigned to groups, descriptive statistics [(group 1 ($m = 58.56$, $sd = 10.22$), group 2 ($m = 65.25$, $sd = 13.29$), group 3 ($m = 65.88$, $sd = 10.85$)] and box plots revealed that the control group had overall higher proficiency students (see Figure 4). However, when LOEP scores were compared by group using a one-way ANOVA, no significant difference was found ($F(2,45) = 1.976$, $p > .05$). Furthermore, the interaction between the covariate, LOEP, and the independent variable, group, was not significant in the prediction of the dependent variables ($F(2, 47) = .798$, $p = .457$ for the directions task and $F(2, 47) = 1.068$, $p = .353$ for the shopping task).

While all of the participants were recruited from lower-level ESL courses, the range of LOEP scores varied considerably from 44 to 98. This sizeable range, in addition to the fact that the hospital directions task scores were not normally distributed, prompted further analysis. Moreover, treating percentage correct as a continuous linear variable

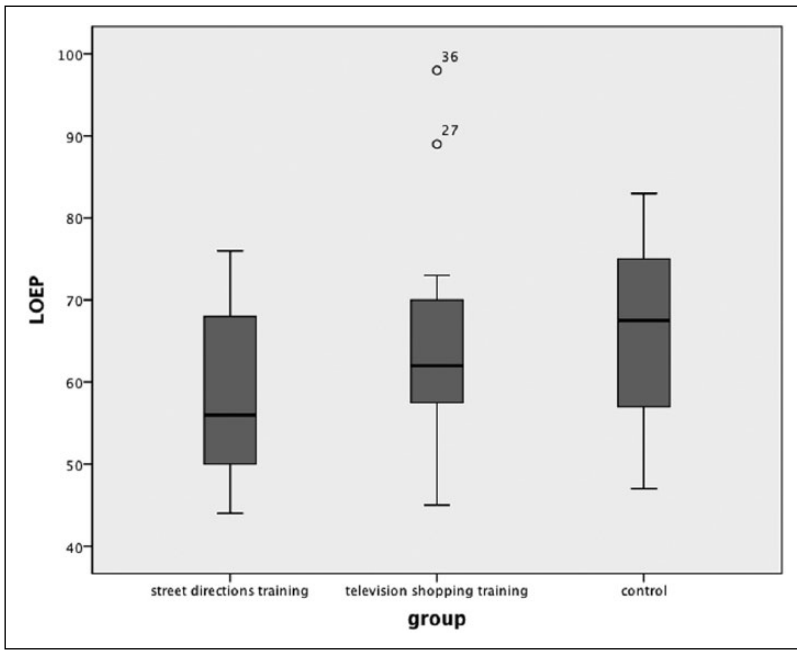


Figure 4. Levels of English Proficiency Test (LOEP) scores by group.

Table 4. Results of logistic regression for entire data set.

	Estimate	SE	z value	Pr (> |z|)
(Intercept)	0.23366	0.20590	1.135	0.256
Treatment	0.03561	0.20762	0.171	0.864
LOEP	0.05380	0.01208	4.456	8.37e–06***
Treatment:LOEP	−0.01428	0.01948	−0.733	0.464

Note. LOEP = Levels of English Proficiency Test

can be problematic (Jaeger, 2008); thus, a mixed-model logistic regression was conducted using binary trial level data.

All data were analysed using R (R Development Core Team, 2010) and the R package *lme4* (Bates, Maechler, & Bolker, 2011). A mixed effects logistic regression model was fitted, with treatment, LOEP, and the interaction of treatment by LOEP as fixed effects. Participant and item were fitted as random effects. More specifically, a random intercept by participants was fitted, as well as a random intercept and random slopes for all fixed effects by item. This random effect structure is analogous to the assumptions of a standard repeated-measures ANOVA. The LOEP predictor was centered for better interpretation. The fixed effects of treatment by participant and treatment adjusted by LOEP score were not significant, $p > .05$ (for results, see Table 4).

Table 5. Results of logistic regression for lower quartile based on Levels of English Proficiency Test (LOEP) score.

	Estimate	SE	*z* value	Pr (> |z|)
(Intercept)	−0.69878	0.33500	−2.086	0.0370*
Treatment	0.72776	0.33639	2.163	0.0305*
LOEP	0.10570	0.07994	1.322	0.1861
Treatment:LOEP	−0.04394	0.08663	−0.507	0.6120

A plot of participants ordered by LOEP scores revealed that the treatment was potentially having an effect on learners of lower proficiency, especially those in the lower quartile. Figure 5 provides a view of the mean accuracy on treated and untreated tasks while ordering participants by LOEP scores. Figure 6 shows the effect of treatment by participant ordered by LOEP score, and an indication that participants of lower proficiency were positively affected by the treatment.

To determine if this visible trend was significant, an exploratory mixed effects logistic regression was performed post hoc on a subset of the data (n = 12, 360 total observations), which included only participants in the lower quartile, based on LOEP scores. The results of this analysis indicate that the treatment, task training, was a significant predictor, $p < .05$, of participants' responses to assessment items (see Table 5). For every one unit change in LOEP, the log odds of choosing the correct response in the near assessment tasks only increases by 0.10570, yet having received task training results in an increase of the log odds or probability of success in choosing the correct response for similar tasks by 0.72776 for a participant with the mean proficiency (m = 49.5) of this subgroup of lower proficiency participants.

Figure 7 provides a view of the mean accuracy based on the percentage scores for the lower quartile. In contrast to Figure 3, here it is clear that those who learned how to follow directions on the street are performing better on the *following directions in a hospital task* than the *shopping for a refrigerator task*. Likewise, participants who learned how to shop for a television are performing better on the *shopping for a refrigerator task* than on the *following directions in a hospital task*. Overall both of the groups who received task training outperformed the control group when assessed on similar or near tasks.

To determine whether or not both of the trained groups were outperforming the control group on dissimilar tasks, the same analysis was conducted on this subgroup of low proficiency participants. No significant effect of task training was found for far assessment tasks, p = .8.

V Discussion

The purpose of this research was to investigate empirically whether or not any transfer occurs between tasks in similar and different domains. As previously mentioned, transfer of learning is an important unresolved issue related to task selection and sequencing for TBLT. Answers to each of the research questions are considered in turn, based on the results presented above.

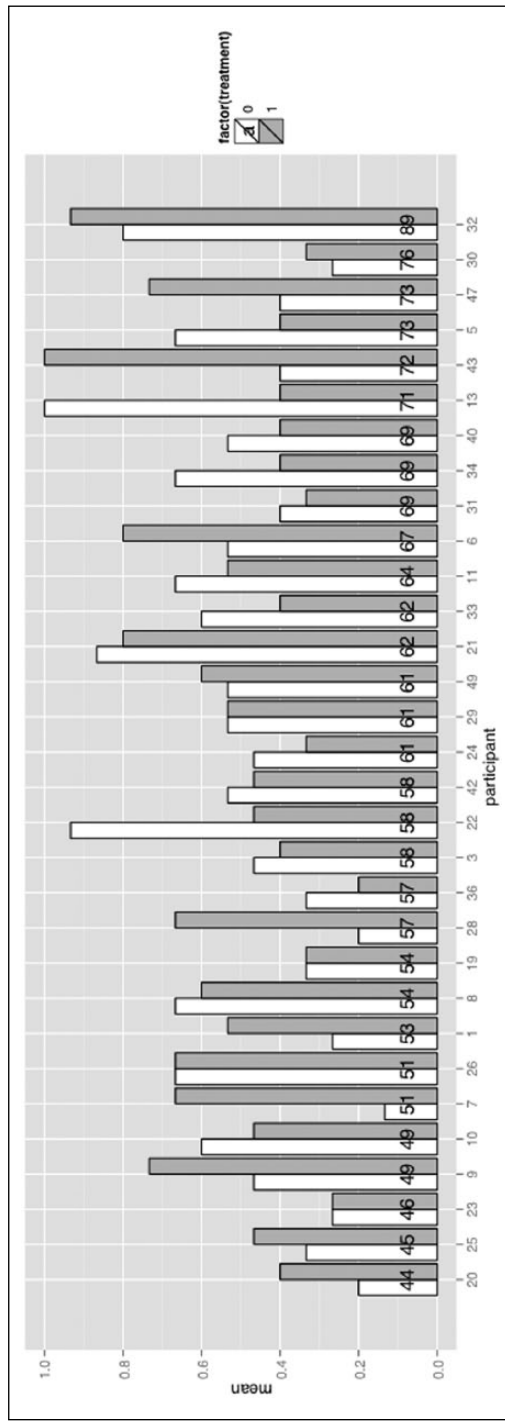


Figure 5. Mean accuracy of each participant ordered by Levels of English Proficiency Test (LOEP) score.

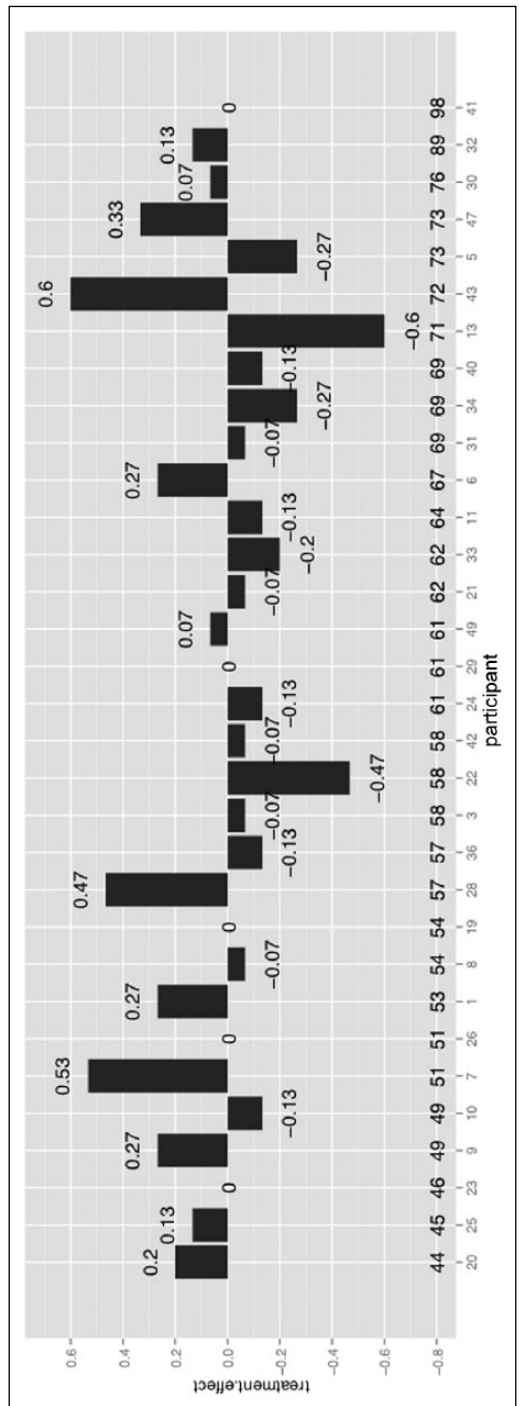


Figure 6. Effect of treatment by participant ordered by Levels of English Proficiency Test (LOEP) score.

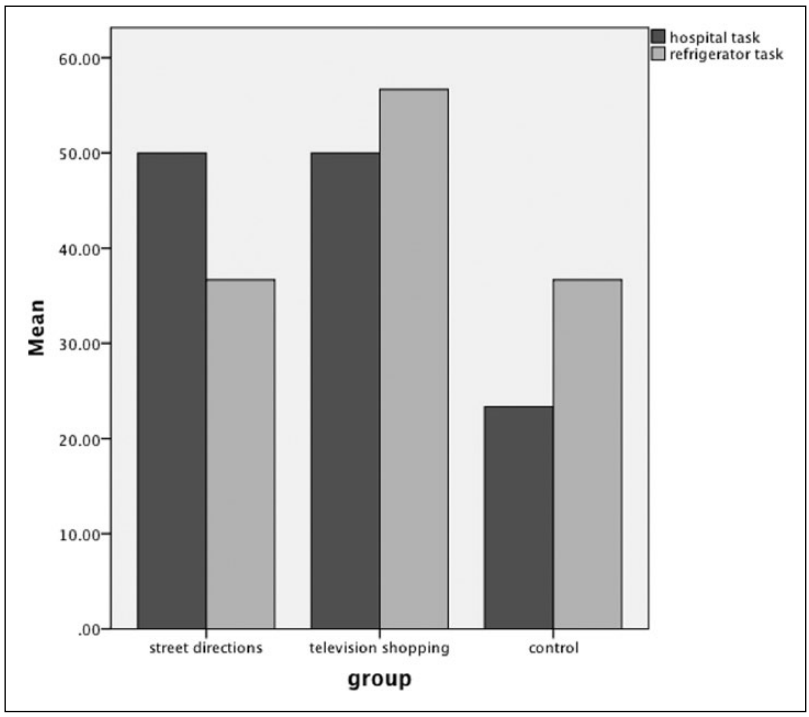


Figure 7. Mean accuracy for lower level participants.

Research question 1: Does learning from one pedagogic task transfer to another task of the same type and complexity?

The results of both a MANCOVA and a mixed effects logistic regression when all participants were included in the data set did not find task training to be a significant predictor of success on the outcome tasks (see Table 4 and Figure 3). In other words, significant transfer was not observed between tasks of the same type and complexity.

In a recent study by Lin (2010), less proficient learners made more progress than higher proficiency learners in comprehension and incidental vocabulary acquisition when they received input enhancement via subtitles within a video-based CALL program. Time on task was not tracked in Lin's study, but the results suggest that treatment may have a differential effect based on proficiency. Likewise, plots ordered by LOEP score in the current study suggested that the treatment was having a greater effect for the lower proficiency students and prompted further exploration. The results of a post hoc exploratory analysis on a subset of the data revealed significant transfer between pedagogic tasks within similar knowledge domains. Low proficiency participants (LOEP score < 55) who learned how to follow directions on a street map task during the treatment significantly outperformed those in the other two groups (those who learned how to shop for a television and those who were in the control group) on the *following directions in the hospital task*. Likewise, low proficiency participants who learned how to shop for a television during the treatment phase significantly outperformed those in the other

two groups (those who learned how to follow directions on the street and those who were in the control group) on the *shopping for a refrigerator task* (see Figure 7). These results are consistent with both the first and second hypothesis.

During the 45-minute treatment, participants learned either how to select the best item to buy or how to follow directions to arrive at the correct destination, and this training appears to have facilitated successful task completion on near transfer tasks for learners with the lowest proficiency. The post hoc and selective nature of the analysis on this subset prevents definitive conclusions, yet what limited transfer was observed occurred between tasks that were near or similar in situation and skill as would be predicted by Haskell (2002). These results also indicate lateral (in the terms of Gagne, 1965) or horizontal (in the terms of Singley & Anderson, 1989) transfer of skill between two tasks with the same level of complexity based on the interactional and cognitive demands of the transfer tasks, which were similar to their respective training tasks. Although the content of the two shopping tasks and the content of the two directions tasks would be considered closer on the continuum as concerns skill and memory demands, the contexts of all four tasks (the two training tasks and the two transfer tasks) might be considered further apart on the continuum proposed by Barnett and Ceci (2002) due to divergence in knowledge domain. In sum, more transfer was observed between the two shopping tasks and between the two directions tasks, which were nearer to each other in both content (i.e. skill, procedure, memory demands, and the nature of the performance) and context (i.e. knowledge domain, and physical, temporal, functional, and social context) based on Barnett and Ceci's framework (2002).

For the higher proficiency participants, the results are less straightforward and more difficult to interpret. Some of these more proficient learners did equally well on both tasks, regardless of treatment condition, yet, others only performed well on one task, and not always the one that would be expected based on the treatment. In other words, there was no effect for treatment in some cases, perhaps due in part to preexisting knowledge. These more advanced learners may already have known one task better than the other and thus been less receptive to 45 minutes of training. The lack of clear evidence of transfer may also be the result of differences in context (e.g. lexical domain), which were great enough that individual learner differences were not affected by 45 minutes of training. Although the two shopping tasks were closer to each other in domain than either shopping task to either directions task, lexical differences in all four tasks may have reduced the amount of observable transfer.

Research question 2: Does more transfer occur between tasks that are similar in knowledge domain (near) than tasks that are dissimilar in domain (far)?

No clear pattern was observable concerning transfer between tasks that were dissimilar. As discussed in the previous section, for the lowest proficiency learners the greatest transfer occurred between tasks within similar domains. When considering the entire data set, there is no evidence of transfer between tasks from dissimilar domains. Likewise, for the subset of lower proficiency participants, no significant far transfer was observed. The third hypothesis, therefore, which predicted that participants who were assigned to a

task treatment group would outperform the control participants, was not supported by the post-hoc analysis conducted on the lower quartile of the data set.

3 Limitations and considerations for future research

It is important to keep in mind that the sample size of the low-proficiency subset is very small (n = 12; 360 total observations); thus, these results need to be interpreted with caution. Furthermore, the selection of the lower group was guided by visual inspection of the data, making this a very post hoc and exploratory analysis, rather than a strict test of the initial hypotheses. Although the tasks were piloted before the study to determine the appropriate proficiency level for participants, the range of LOEP scores found in the low level ESL courses was much greater than expected. A replication of this study would need to control for proficiency level more tightly and perhaps include additional pretests to determine prior domain knowledge. And, while the current study provided multiple trials in both the practice task and the transfer task, the design did not include an assessment to ensure that something was learned from the training. Future designs could include assessment checks built into the training tasks to ensure that learners were indeed able to complete simpler versions of the tasks (e.g. with one or two pieces of directional information) before moving on to more complex versions (e.g. with more directional moves).

Variations in treatment length should also be considered in future studies; 45 minutes of training may not be sufficient for consistent transfer. However, lengthening the treatment or administering multiple assessments will likely require multiple sessions with participants, which could result in another set of problems concerning variable control.

Finally, in the current study, the task itself served as the construct of interest, and task transfer was measured based on whether or not participants were able to accurately complete each item (i.e. choose the correct product or arrive at the correct destination). In the absence of clear transfer findings, a more fine-grained analysis of exactly what is transferring might be desirable in future studies. Singley and Anderson (1989) suggest measuring all learning and transfer separately and independently. For instance, since all tasks in the current study were receptive and controlled, it might be possible to more closely analyse lexical and structural similarities from one task to another. Certainly some lexical overlap existed between the *following directions on the street task* and the *following directions in a hospital task*. Phrases such as ‘make a left’, ‘turn right’, ‘on your right’, and ‘to your left’ existed in both of these tasks. Likewise, there was lexical crossover between the two shopping tasks. Both shopping tasks included phrases such as ‘you prefer’, ‘you are looking for’, and ‘you would like’, and some overlap in the terms used to discuss product features (e.g. price, brand, and size). However, differences existed in the landmark related lexis found in the two following directions tasks (e.g. elevators, gift shop, and cashier in the *following directions in the hospital task* compared to U.S. Post Office, Botanical Garden, and Museum in the *following street directions task*) which could have contributed to the lack of transfer between the direction tasks. There were also distinct differences in the technical terms found in the shopping tasks (e.g. LED, HDTV, and so on for the televisions, and French door, Side-by-Side,

and Stainless Steel for the refrigerators) which may have contributed to the lack of transfer between the shopping tasks. Future studies of task transfer should thus consider a more systematic analysis of lexical overlap between tasks.

VI Conclusions

When analysing the data from all 48 participants combined, there was no effect for treatment, and overall listening proficiency was found to be a greater predictor of successful task completion than 45 minutes of training. Those in the lower quartile, however, were significantly affected by the training, as both treatment groups outperformed the control group on their respective tasks, demonstrating some transfer of learning across similar task domains. Despite the limitations due to the small *n*-size of low proficiency participants and the post hoc nature of the second analysis, this study provides at least some evidence that task-related language abilities are transferable. Future research should aim to replicate these results with additional low-level learners and should also explore different task types at varying levels of proficiency, in order to begin to identify characteristics of tasks or skills that are generalizable. Tasks-types in different domains may have varying degrees of transfer. Citing Johns (1988), Long (2015) points out that even in specialized areas such as EAP, some skills and tasks may generalize, whereas others might be highly specialized. Thus, conducting similar studies in a variety of settings is necessary before more robust claims can be made concerning task selection for TBLT.

Acknowledgements

The study reported in this article was conducted as part of my doctoral studies in Second Language Acquisition at the University of Maryland. I would like to thank my advisor, Michael Long, and my committee members for this paper, Robert DeKeyser, Scott Jackson, and Steven Ross for their constructive comments and valuable feedback. I would also like to thank the two anonymous *Language Teaching Research* reviewers whose insightful comments helped me to improve and strengthen my discussion. Last but certainly not least, I would like to thank the participants in the study and the instructors who graciously granted me access to their classes.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Note

1. The LOEP (Levels of English Proficiency) is a standardized placement test used by the college and was administered prior to the study. According to the College Board Research and Development Office, the reported reliability estimate for the LOEP listening test is .88.

References

Barnett, S.M., & Ceci, S.J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin*, *128*, 612–637.

Bates, D., Maechler, M., & Bolker, B. (2011). *lme4: Linear mixed-effects models using S4 classes*. R package version 0.999375-40. Retrieved from: http://CRAN.R-project.org/package=lme4 (January 2015).

Blume, B.D., Ford, K.J., Baldwin, T.T., & Huang, J.L. (2010). Transfer of training: A meta-analytic review. *Journal of Management*, *36*, 1065–1105.

Brindley, G. (1994). Task-centered language assessment in language learning. The promise and the challenge. In: N. Bird, P. Falvey, A. Tsui, D. Allison, & A. McNeill (Eds.), *Language and learning: Papers presented at the Annual International Language in Education Conference, Hong Kong* (pp. 73–94). Hong Kong: Hong Kong Education Department.

Brown, J.D., Hudson, T., Norris, J.M., & Bonk, W. (2002) *An investigation of second language task-based performance assessments*. Technical Report #24. Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Center.

Chen, Z. (1996). Children's analogical problem solving: The effects of superficial, structural, and procedural similarity. *Journal of Experimental Child Psychology*, *62*, 410–431.

Doughty, C.J., & Long, M.H. (2003). Optimal psycholinguistic environments for distance foreign language learning. *Forum of International Development Studies*, *23*, 35–73.

Gagne, R.M. (1965). *The conditions of learning*. New York: Holt, Rinehart & Winston.

Gick, M.L., & Holyoak, K.J. (1980). Analogical problem solving. *Cognitive Psychology*, *12*, 306–355.

Grose, R.F., & Briney, R.C. (1963). *Transfer of learning: An enduring problem in psychology*. Princeton, NJ: Van Nostrand.

Haskell, R.E. (2001). *Transfer of learning: Cognition, instruction, and reasoning*. San Diego, CA: Academic Press.

Jaeger, F. (2008) Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language*, *59*, 434–446.

Johns, A.M. (1988). The discourse communities dilemma: Identifying transferable skills for the academic milieu. *English for Specific Purposes*, *7*, 55–59.

Lin, L. (2010). A video-based CALL program for proficient and less-proficient L2 learners' comprehension ability, incidental vocabulary acquisition. *Educational Media International*, *47*, 199–216.

Long, M.H. (1985). A role for instruction in second language acquisition: Task-based language teaching. In: K. Hyltenstam & M. Pienemann (Eds.), *Modeling and assessing second language acquisition*. (pp. 77–99). Clevedon: Multilingual Matters.

Long, M.H. (2007). Texts, tasks, and the advanced learner. In: M.H. Long (Ed.), *Problems in SLA* (pp. 119–138). Mahwah, NJ: Lawrence Erlbaum.

Long, M.H. (2015). Task-based needs and means analysis. In: M.H. Long (Ed.), *Second language acquisition and task-based language teaching*. Oxford: Wiley-Blackwell.

Long, M., & Crookes, G. (1992). Three approaches to task-based syllabus design. *TESOL Quarterly*, *26*, 27–56.

Long, M., & Crookes, G. (1993). Units of analysis in syllabus design: The case for task. In: G. Crookes & S. Gass (Eds.), *Tasks in a pedagogical context: Integrating theory and practice* (pp. 9–54). Clevedon: Multilingual Matters.

Long, M.H., & Norris, J.M. (2000). Task-based teaching and assessment. In: M. Byram (Ed.), *Encyclopedia of language teaching* (pp. 597–603). London: Routledge.

Mislevy, R.L., Steinberg, L.S., & Almond, R.G. (2002). Design and analysis in task-based language assessment. *Language Testing*, *19*, 477–496.

Morris, D.D., Bransford, J.D., & Franks, J.J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*, 519–533.

Norris, J.M. (2002). Interpretations, intended uses and designs in task-based language assessment. *Language Testing*, *19*, 337–346.
Norris, J.M. (2009). Task-based teaching and testing. In: M.H. Long & C.J. Doughty (Eds.), *Handbook of language teaching* (pp. 578–594). Oxford: Blackwell.
Norris, J.M., Brown, J.D., Hudson, T.D., & Bonk, W. (2002). Examinee abilities and task difficulty. *Language Testing*, *19*, 425–443.
Norris, J.M., Brown, J.D., Hudson, T.D., & Yoshioka, J.K. (1998). *Designing second language performance assessments*. Technical Report #18. Honolulu: University of Hawaii Press.
R Development Core Team (2010). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: http://www.R-project.org (January 2015).
Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In: P. Robinson (Ed.), *Cognition and second language instruction* (pp. 287–318). Cambridge: Cambridge University Press.
Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57.
Robinson, P. (2007). Criteria for classifying and sequencing pedagogic tasks. In: M. del P. Garcia-Mayo (Ed.), *Investigating tasks is formal language learning* (pp. 7–26). Clevedon: Multilingual Matters.
Robinson, P. (2009). Syllabus design. In: M.H. Long & C.J. Doughty (Eds.), *Handbook of language teaching* (pp. 294–310). Oxford: Blackwell.
Robinson, P., & Ross, S. (1996). The development of task-based assessment in English for academic purposes programs. *Applied Linguistics*, *17*, 455–476.
Schmidt, R.A., & Bjork, R.A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.
Singley, M.K., & Anderson, J.R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press.
Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*, 38–62.
Skehan, P. (1998). *A cognitive approach to language learning*. Oxford: Oxford University Press.
Thorndike, E.L., & Woodworth, R.S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. *Psychological Review*, *8*, 247–261.

Appendix 1

Example of hospital directions assessment task (via video)

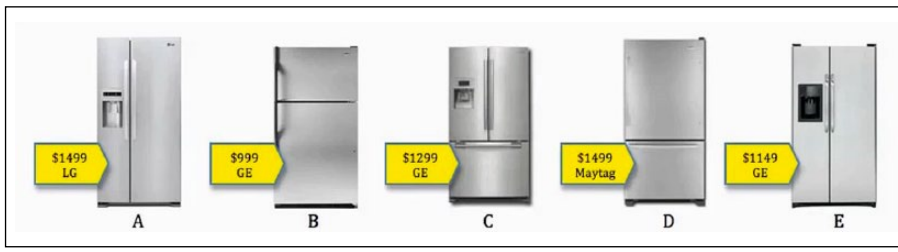
Participant sees a map of the first floor of the hospital. An arrow indicates where to start. The participant hears:

Start here. Go straight down this hall to the double door. When you get to the end, make a right. Keep walking until you pass the pharmacy and then make the next right. Where are you now?

a. Medical records
b. Laboratory
c. Customer relations
d. Chapel
e. Main lobby

Example of refrigerator shopping assessment task (via video):

Participant sees an image with several different refrigerators (e.g. different colors, features and brands).



The participant hears:

You said that you do not like the traditional top freezer. And, you prefer GE, and you would like an in-door ice dispenser. You like the side-by-side style.

Choose the best refrigerator to buy.

a. Refrigerator A
b. Refrigerator B
c. Refrigerator C
d. Refrigerator D
e. Refrigerator E

Appendix 2

Exit survey

1. What is your native (first) language?
2. Where were you born? (in what country?)
3. How long have you been living in the United States?
4. Please indicate whether you are male or female.
5. How old are you?
6. How similar were the two tasks you were asked to do? (following directions in a hospital and buying the best refrigerator)

   0 = completely the same (not different at all)
   1 = very similar (slightly different)
   2 = somewhat the same (different)
   3 = slightly similar (very different)
   4 = not the same at all (completely different)

7. How difficult was the *following directions in a hospital task*?

   0 = not difficult
   1 = not very difficult
   2 = somewhat difficult
   3 = difficult
   4 = very difficult

8. How difficult was the *buying the best refrigerator task*?

   0 = not difficult
   1 = not very difficult
   2 = somewhat difficult
   3 = difficult
   4 = very difficult