Article

LANGUAGE
TEACHING
RESEARCH

Language Teaching Research
1–30
© The Author(s) 2017
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1362168817744389
journals.sagepub.com/home/ltr
SAGE

TBLT implementation and evaluation: A meta-analysis

Lara Bryfonski
Georgetown University, USA

Todd H. McKay
Georgetown University, USA

Abstract

Task-based language teaching (TBLT) is an empirically investigated pedagogy that has garnered attention from language programs across the globe. TBLT provides an alternative to traditional grammar translation or present-practice-produce pedagogies by emphasizing interaction during authentic tasks. Despite several previous meta-analyses investigating the effect of individual tasks or short-term task-based treatments on second language (L2) development, no studies to date have synthesized the effects of long-term implementation of TBLT in authentic language classrooms. The present study uses meta-analytic techniques to investigate the effectiveness of TBLT programs on L2 learning. Findings based on a sample of 52 studies revealed an overall positive and strong effect ($d = 0.93$) for TBLT implementation on a variety of learning outcomes. The study further examined a range of programmatic and methodological features that moderated these main-effects (program region, institution type, needs analysis, and cycles of implementation). Additionally, synthesizing across both quantitative and qualitative data, results also showed positive stakeholder perceptions towards TBLT programs. The study concludes with implications for the domain of TBLT implementation, language program evaluation, and future research in this domain.

Keywords

task-based language teaching, task-based learning, TBLT, language program evaluation, meta-analysis, second language pedagogy

Corresponding author:
Lara Bryfonski, Department of Linguistics, Georgetown University, Box 571051, Poulton Hall 240, 1421 37th Street NW, Washington, DC 20057, USA
Email: leb110@georgetown.edu

I Introduction

Task-based language teaching (TBLT) has attracted attention in recent decades due to its solid theoretical grounding in second language acquisition (SLA) research and theory. Tasks are authentic, communicative uses language learners have for the target language (Long, 2005). Pedagogical tasks are sequenced by complexity until they approximate real-world use. In this way, TBLT strives to provide an alternative to traditional present-practice-produce (PPP) or grammar-translation pedagogies, which have been shown to be inauthentic and inconsistent with SLA research findings (Long, 2016).

Syllabi based on TBLT have been implemented and evaluated program-wide in a variety of contexts around the world (e.g. Burwell et al., 2009; Shintani, 2011; Van den Branden, 2006). Research on tasks and task-based programs have been discussed at international conferences (The International Conference on Task-Based Language Teaching), and have been the subject of recent review publications (e.g. *Annual Review of Applied Linguistics* volume 36, 2016, on Tasks) but not without criticism (e.g. Bruton, 2005; Klapper, 2003; Swan, 2011). Tasks have also been the subject of many empirical investigations with most researchers experimentally manipulating tasks in short-term, classroom based interventions (e.g. Li, Ellis & Zhu, 2016).

The effects of the kinds of interaction-driven learning that occurs during task-based interactions have been meta-analysed several times (Cobb, 2010; Keck et al., 2006; Mackey & Goo, 2007) with overall positive effects for task-based interaction bolstering claims about the potential effectiveness of TBLT as a whole. However, this previous work has focused on short-term treatments of features of TBLT, or task-based interactions, rather than program-level features. Despite interest and enthusiasm about TBLT, less has been published reporting the quantitative effects of the implementation of TBLT programs *in situ*. Moreover, a meta-analytic investigation of the effects of TBLT programs in-context has not yet been undertaken. The current article, therefore, aims to provide quantitative meta-analytic findings on the effectiveness of TBLT programs for language development.

II Background

1 Task-based language teaching

Task-based language teaching (TBLT) utilizes task, as opposed to language, as the unit of instruction in language classrooms (Long, 1985, 2015). While traditional synthetic syllabuses present language in discrete grammatical units, calling for learners to synthesize forms to create meaning when called upon to do so, TBLT emphasizes authentic, communication-driven tasks that provide task-related focus on form purported to be congruent with a learner's own internal syllabus. The aim of TBLT is to prepare students to use their linguistic skills in meaningful interactions outside the classroom (Long, 2015).

According to Long and Norris (2000), the development and implementation of a TBLT program should follow a prescribed set of steps that are task-oriented throughout, beginning with a task-based needs analysis that identifies the authentic language needs of the learners and the target tasks associated with those needs. These needs might range from the academic (e.g. finding a journal article) to the everyday (e.g. making a doctor's appointment). The next steps in TBLT program development all involve preparing the results of the needs analysis for implementation in the language classroom. This occurs through grouping the previously identified needs into superordinate 'target task-types', preparing the pedagogic equivalents of those tasks, and sequencing the tasks according to relative 'task complexity', thereby forming the task-based syllabus. Finally, the syllabus is implemented, integrating focus on form as needed throughout task performances and ending with an evaluation where data on student performance is gathered via task-based assessment. The process is intended to be iterative, with cycles of needs analyses and evaluations working to improve the overall implementation of the task-based program.

The implementation of TBLT programs has been investigated in contexts around the world (e.g. Van den Branden, 2006), reporting various degrees of success after comparing TBLT to other forms of language teaching. But despite growing interest in and empirical evidence to support task-based programs, TBLT has not been without criticism (for a recent review of criticisms and responses, see Long, 2016). Some have argued that the incidental focus on form that is at the heart of TBLT neglects grammar and vocabulary and therefore impacts language development (Swan, 2005; Widdowson, 2003). Critiques have also been leveled at the implementation of TBLT in specific contexts, such as in secondary schools (Bruton, 2005). Furthermore, the compatibility of TBLT amidst the sociocultural realities and educational cultures of certain foreign-language contexts has been questioned by some scholars (e.g. Carless, 2003; Ellis, 2016a, 2016b). Given these criticisms, evidence that points towards the effects of long-term implementation of TBLT in authentic language classrooms seems both timely and necessary.

2 Meta-analyses related to TBLT

A number of meta-analyses have been conducted that examine the overall effect of other, non-program-level features of TBLT on various outcome measures, such as on the effects of task interaction (Keck et al., 2006) and task complexity (Jackson & Suethanapornkul, 2013; Sasayama, Malicka, & Norris, 2015). Additionally, several scholars have compiled useful literature reviews (e.g. Long, 2015, pp. 343–365) or reviews of task-based issues as they pertain to specific regions or settings (e.g. Butler, 2011).

Keck et al. (2006) and Mackey and Goo (2007) both examined the overall effect of task-based interaction on the acquisition of certain grammatical and lexical features. The researchers were also interested in how long the effects of task-based interaction endured over time and how task types and certain design features mediated the acquisition process. In her dissertation, Cobb (2010) builds on this previous work by looking at the effect of task-based interaction on the acquisition of grammatical structures. Cobb meta-analysed 15 primary studies and found that learners who participated in oral-communication tasks (the task-based treatment) performed better on measures of grammatical acquisition than control or comparison groups. Other meta-analyses have looked at the effects of task features on second language (L2) development. Jackson and Suethanapornkul (2013) conducted a review of studies that addressed Robinson's Cognition Hypothesis (see Robinson, 2001). Nine primary studies were meta-analysed to investigate the extent to which tasks of increasing complexity along resource-directing dimensions affected L2 production. The authors found that increasing task complexity along resource-direction dimensions had a small positive effect on accuracy but a small negative effect for fluency. In another study, Plonsky and Kim (2016) conducted a synthesis of both substantive and methodological task-based features associated with L2 learner production. Substantive features for which studies were coded included target features in L2 production, such as grammar, vocabulary, and CAF (complexity, accuracy and fluency) measures, as well as a range of interactional features, including language-related episodes, repairs, and recasts (among others). The authors also coded for several methodological features, such as study design, statistical procedures used, and features of researcher transparency. The authors found that, using tasks to elicit production, task-based researchers tend to analyse grammar, vocabulary, L2 interaction, and accuracy, while little work has been done in terms of pragmatics, pronunciation, and indicators of the quality of task performance. While the above meta-analyses contribute to the field's understanding of task-based interaction and task complexity, and witness the contributions of meta-analysis in the task-based domain, none have looked at program-level components in task-based programs in the course of implementation and evaluation work.

Quantitative meta-analyses aside, several reviews have contextualized the contributions of task-based research within specific regions and settings. In a recent review, Ziegler (2016) discussed the advantages of TBLT within computer-mediated settings and the potential for teaching and learning within them to support the development of TBLT. Butler's (2011) review examines the implementation of communicative language teaching (CLT) and TBLT in a number of East Asian countries; she notes that there are substantial challenges to promoting CLT and TBLT in Asian classrooms, such as cultural educational ethos and national examination systems, that often lead to the rocky incorporation of either approach at the local level (for additional support along these lines, see also Carless, 2012; but for evidence to the contrary, Iwashita & Li, 2012). Butler concludes by calling for more flexible adaptations of TBLT in authentic contexts.

III TBLT program features

Evidence from previous work investigating the implementation and evaluation of task-based programs suggests that a variety of factors are at play when it comes to the success of the program's implementation and subsequent learner development. Given these findings, the present study will also examine the following potential moderator variables at the program level: program region, institution type, the presence of needs analysis, cycles of implementation, and stakeholder perceptions.

1 Program region and institution type

A potential factor moderating outcomes for TBLT programs include the region in which the study took place and the type of institution where the program was based. To date, much of the scholarship contributing to our understanding of the evaluation and implementation of task-based programs in overseas foreign-language settings has originated in East Asian countries (Carless, 2012). However, several notable examples of task-based, foreign-language work in English-dominant countries include those by Markee (1996), Towell and

Tomlinson (1999), and González-Lloret and Nielson (2015), among others. While this work invariably serves to enhance our understanding of TBLT, some have noted that task-based implementation and evaluation in overseas foreign-language contexts is of an altogether different variety, evolving within a 'set of conditions and social practices that do not necessarily coincide with those in [second language] contexts' (Shehadeh, 2012, pp. 3–4). Furthermore, there is some evidence to indicate that TBLT implementation success may be affected by the type of institution where the curriculum is implemented. Some studies, such as Park (2012) have investigated the use of TBLT principles in secondary schools and found overwhelmingly positive effects. However, a study by Tinker-Sachs (2007), which examined school-aged children (grades 4–6) learning English in Hong Kong, found mixed reactions to TBLT principles. Studies in the university setting have also uncovered mixed results, such as in De Ridder et al. (2007), which found students in the TBLT group outperformed a control group on grammar, vocabulary, and fluency measures but not on pronunciation or intonation measures. For these reasons, the current meta-analysis will consider program context, including institution type as potential moderating factors.

2 Needs analysis

A handful of researchers and practitioners have proposed that a weak version of TBLT is more appropriate than a strong version (for more on this distinction, see Skehan, 1996) in some educational contexts. A 'strong' task-based program would be one in which the tasks that make up a program are derived from a needs analysis and thus directly align with students' needs (Long, 2016). In a 'weak', task-supported approach, vocabulary or grammatical structures are frequently the main units of analyses, with tasks consisting of pedagogic activities that provide learners with a means of practicing target vocabulary or grammar. Also in a task-supported approach, the teacher is more often at the helm of their creation and in-class use (Klapper, 2003). Carless (2004, 2007) is now well known for endorsing a more contextualized, task-supported teaching and learning enterprise for primary and secondary education in Hong Kong. However, Norris (2016) notes that the identification of learners' needs – and their realization and articulation within various program components, such as assessment and materials development – is what really holds task-based programs together and helps them reach their full potential.

For the purposes of the current study, any study addressing the evaluation or implementation of TBLT components (i.e. a task-based and not task-supported study) is open to inclusion in the meta-analysis. However, though some argue that needs analysis (or needs assessment) is a type of evaluation (Stufflebeam, 1983), others note that needs analysis is often tied more narrowly to early phases of program planning (see Altschuld & Watkins, 2014). Therefore, studies that were strictly needs analyses were not included in the meta-analysis.

3 Cycles

No program is ever perfect. Materials are continuously updated, changes are made to the assessment program, faculty leave and new faculty are hired, accountability requirements change, and profiles of students are constantly shifting. One way to make sure a program continues to meet students' needs is through regular cycles of needs analysis and evaluation. Brown (1995) aptly notes that 'the process of curriculum development is never finished' and 'ongoing program evaluation … is the glue that connects and holds all the elements [needs, goals and objectives, teaching, materials, and testing] together' (p. 217). Evaluation as an ongoing, cyclical process is also a key feature of task-based evaluation (see Norris, 2015, 2016). Studies by Hill and Tschudi (2008), Prabhu (1987), McDonough and Chaikitmongkol (2007), Towell and Tomlinson (1999), Lai, Zhao, and Wang (2011), and González-Lloret and Nielson (2015) each employed evaluation cycles wherein findings were obtained and feedback was used to inform programmatic changes. Given the emphasis previous researchers have placed on cyclical evaluation and implementation (Norris, 2015), there is a need to investigate whether or not the inclusion of evaluation cycles has any benefit to the goals of a TBLT program.

4 Stakeholder perceptions

Research into language program evaluation has emphasized the need to triangulate findings beyond linguistic outcomes and include the reactions of key stakeholders, such as teachers, students, and administrators (Beretta, 1992; Norris, 2016; Patton, 2008). Patton (2008), in particular, advocates for the integration of stakeholders in all stages of the evaluation process in order to ensure the results from evaluations are actually used. The needs of key stakeholders are often gathered through questionnaires, interviews, focus groups, and other methods (e.g. González-Lloret & Nielson, 2015; McDonough & Chaikitmongkol, 2007) in the form of stakeholder satisfaction ratings, interest, beliefs, perceptions, attitudes, or opinions.

Such efforts align with suggestions found elsewhere in the field regarding the need to take the pedagogical ramifications of instructed SLA research into account by examining what actually makes a difference for language instructors and learners in practice (Leow, 2016). Given the range of reactions to TBLT programs described above, data from L2 outcomes alone may not be enough to adequately understand the effectiveness of TBLT pedagogy.

IV The present study

Given this gap in the literature, the current project set out to accomplish three goals. First, the current analysis describes the methodological and programmatic features of TBLT implementation research in published and unpublished literature to date. Second, the existing quantitative findings of studies that documented the implementation or evaluation of task-based programs are synthesized. As Beretta (1992) states, 'evaluation is typically concerned with real-world issues rather than with laboratory effects'; therefore, 'studies that show learning or achievement over the long term are going to be more relevant than short-term experiments' (p. 9). This meta-analysis therefore focuses on those studies that either implemented or evaluated (or implemented and then evaluated) program-level components (e.g. task-selection and sequencing, materials and instructional development, assessment, etc.). Finally, this meta-analysis synthesizes findings from implementation and/or evaluation studies of TBLT programs that additionally included data on the effectiveness of the TBLT program according to stakeholder perceptions.

V Research questions

1. What are the methodological and programmatic features of TBLT implementation research?
2. How effective are TBLT programs for L2 outcomes? To what extent do the following factors impact the effectiveness of TBLT programs: program region, institution type, needs analysis, and cycles of implementation.
3. How effective are TBLT programs according to stakeholder perceptions?

VI Methodology

1 Study identification and retrieval

Studies were identified via a comprehensive search through a variety of online sources. Published literature was retrieved from the online research databases ProQuest (including Linguistics and Language Behavior Abstracts), Project Muse, PsycInfo, JSTOR, and EbscoHost (including Education Resources Information Clearing house [ERIC], Academic Search Premier, and Education Full Text). Additional published work was found using backward citations of prominent books and review articles (e.g. Long, 2015) as well as literature reviews and syntheses on the topic of TBLT implementation (e.g. Long, 2016) and previous implementation studies (e.g. McDonough & Chaikitmongkol, 2007). In an effort to include unpublished work (as suggested by, among others, Norris & Ortega, 2006; Oswald & Plonsky, 2010) and to avoid the potential for publication bias, Google Scholar, Google, and the biennial TBLT conference PowerPoint repository were also thoroughly searched. Since the domain of TBLT implementation includes a range of unpublished work, including action research, work from lesser-known journals, international universities and private reports, the current study included data from these unpublished sources.

The keyword search utilized to locate relevant work in the listed sources included the following words in various combinations: *task(-)based language teaching, TBLT, TBL, task(-)based learning, task(-)based instruction*, or *task(-)based**, along with, *assessment, evaluation, implementation* or *innovation.*

2 Inclusion and exclusion criteria

After completion of the search described above, 194 studies were retrieved and were subsequently reviewed. In order to obtain only the studies that could be analysed in the final meta-analysis, the following inclusion and exclusion criteria were applied (for a more detailed list of inclusion and exclusion criteria, see Appendix 1):

1. studies accessible in English;
2. studies that reported on the implementation, evaluation, or assessment of a task-based program;

3. studies that reported on the implementation or evaluation of an intact task-based program; studies that reported on a single experimental treatment, studies that did not last for at least one cycle of a program, or at least a semester, were excluded;
4. to address research question 2: studies that compared the implementation of a TBLT curriculum with some other form of language teaching;
5. to address research question 3; studies that measured changes in student interest, cognition, attitudes or evaluations of the TBLT program on a questionnaire were separately included;
6. studies that examined the effects of the TBLT program's implementation on some outcome measure, such as L2 learning or development, including production or performance were included. For perception-based studies, studies that evaluated the implementation with student perceptions were additionally included;
7. studies that did not include sufficient information (*n*-size, standard deviations) to calculate effect sizes were excluded if authors did not respond to email requests for missing data.

Due to the large amount of high-inference inclusion and exclusion criteria, when deciding to include or exclude a study, the authors first categorized 100 studies together by considering the study as a whole (i.e. as opposed to just the abstract). Ultimately, 52 studies were included for the final analysis.

3 Coding procedure

The 52 studies were coded for the features listed in Table 1. The coding scheme was based on previous syntheses and meta-analyses (e.g. Ziegler, 2016), guides to meta-analysis (Cooper, 2016), previous meta-analyses of task-based interaction (Cobb, 2010; Mackey & Goo, 2007), as well as through the examination of common program evaluation features. After both authors piloted the instrument, two experts in meta-analysis also coded a sub-sample for reliability and provided their feedback, after which the scheme was revised again. The two authors then coded a sub-sample of the studies (k = 10) together to ensure inter-rater reliability and discussed and resolved any disagreements on all coded studies.

4 Contrasts, corrections, and formulas

Effect sizes (*d*) were calculated in line with several identified contrasts. First, if a study reported on the differences in means between a task-based group and a non-task-based group, whether a control or comparison group, on a pre-test and post-test, then the effect size was computed by comparing means of the task-based group with those of the non-task-based group. Second, if a study compared the differences in means between a task-based group and two or more non-task-based groups (e.g. a true control group and a comparison group), then the effect size was computed by comparing means of the task-based group with means of the control group. Third, if a study reported the difference in means for a single task-based group at two points in time (a within-groups, paired-samples design), then the effect size was computed by comparing means from the first

Table 1. Coding scheme.

Variable	Definition/Operationalization
Report characteristics:	
Report type	Journal article, book chapter, dissertation, master thesis, private report, government document, PowerPoint presentation
Peer review	Peer review, funding and funding source
Program characteristics:	
Setting	Foreign or second language setting
Institution type	K–12, university, language institute, multiple
Country	The country where the TBLT program took place
TBLT type	The presence or absence of a task-based needs-analysis used to inform the program's design
Cycles	Whether the program employed cycles of implementation
Length (weeks)	The length of implementation in weeks
Length (class hours)	The length of treatment in hours (total class time)
Modality	Face-to-face, computer-mediated, or multiple
Participant characteristics:	
Participants L1 (first language)	First language(s) of the participants
Participant L2 (second language)	Target language of the TBLT program
Proficiency level	Beginner, intermediate, advanced, heritage, multiple, not specified
Proficiency measure	Impressionistic, norm-referenced, institutional (final exam etc.), achievement, self-assessment, not reported
Selection criteria	Availability, intact classes, random assignment, not reported, multiple
Methodology:	
Mixed methods	Presence of a mixed-methodology design
Pre-test	Inclusion of a pre-test
Delayed post-test	Inclusion of delayed post-tests
Pre-post correlation	Inclusion of pre-post test correlation statistic
Reliability: instrument	Whether or not reliability was reported for the outcome measure or questionnaire
Qualitative methods	Use of interviews, focus groups, observations, journals/reflections/blogs, recordings of classroom discourse/oral performances, course evaluations, field notes, questionnaires.
Questionnaire methods	*n* respondents reported, reporting of response rate
Overall question	Presence of an overall question on a questionnaire, such as 'What is your overall level of satisfaction?' or 'What is your overall opinion?' 'with/of the course?'
Questionnaire indicators	Interest, cognition (beliefs, perceptions), attitudes, satisfaction, opinions, other

observation with means of the second. To address research question 3, if a study reported student or teachers' changes in overall views, perceptions, attitudes, or satisfaction with a program (referring here to those perception studies that were included in the analysis) over time, then the percentage reported (or that could be calculated from means and respondent sample sizes from questionnaire items or scales) at the latest point in time was recorded.

Corrections were made while computing effect sizes for the various contrasts outlined above. Although attempts were made to sample randomly from a student population at some level, such as by randomly assigning an available pool of 60 students to task- and non-task-based groups (n = 30 in each), the fact remains that random assignment in schools and higher-learning institutions is rarely ever random in the strict sense, nor does it ensure group equivalence prior to instruction. Therefore, regardless of the size of the effect, pre-test effects were subtracted from the post-test to control for pre-existing group differences, as recommended in Plonsky and Oswald (2014) and as seen recently at the primary level in McManus and Marsden (2017), though not commonly found in primary or meta-analytic effect-size calculations.

VII Analysis

1 Sample-size inflation and multiple effect sizes

The number of effect sizes generated by any one study included in this meta-analysis ranged from one to nine, based on the number of samples and/or outcome measures. Consequently, several decisions were made regarding the handling of such data dependencies. First, if a study reported mean differences between groups on multiple outcome measures, an average of effect sizes was calculated, and a single effect size was added to the analysis. Second, if a study reported on overall and component mean differences (e.g. those for an achievement test on the whole as well as across individual skill areas on the test), the overall effect was calculated and included in the analysis. Third, if a study reported mean differences for a between-groups comparison as well as a within-groups comparison, an effect was calculated for both designs, and the study contributed two effect sizes. The same procedure was followed if a study reported differences in either a within- or between-groups design as well as an overall percentage, say, from a course evaluation or end-of-term program questionnaire. Effect sizes contributed from studies of different designs (e.g. between participants vs. within participants) were treated separately. Likewise, percentages-based results were kept separate as well.

Analyses were performed using Cumming's (2001) Exploratory Software for Confidence Intervals (ESCI), Becker's (1999) online effect-size calculator, and the trial version of the Comprehensive Meta-Analysis (CMA; Borenstein, Hedges, Higgins, & Rothstein, 2005) software.

2 Analysis procedure

Effect sizes were weighted according to sample size and aggregated to calculate descriptive statistics (M, SD, etc.). This analysis used a random-effects model; given the variety of language-learning settings, target languages, countries and regions, outcome measures, treatment and program lengths, and study designs, the researchers concluded that the main effect computed in this study is taken from the distribution of observed effects rather than a fixed effect. Mean differences were all calculated as Cohen's *d* effect sizes, which seemed most suitable given that only three studies reported total sample sizes less than 20. Effect sizes are interpreted with respect to Plonsky and Oswald's (2014) field-specific benchmarks: small ($d = .40$), medium ($d = .70$), and large ($d = 1.00$) for between-groups differences, and small ($d = .60$), medium ($d = 1.00$), and large ($d = 1.40$), for pre-post or within-group contrasts. Outliers in this study were identified in two ways, by analysing *z*-scores (for *d* values) in SPSS and by performing a sensitivity analysis in CMA. Publication bias was assessed by looking at a funnel plot, plotting effect sizes against study sample sizes. Any effect size with a *z*-score exceeding ±3.0 was flagged as an outlier.

VIII Results

1 Research characteristics

The literature search uncovered 47 studies that described either the effect of TBLT on student learning outcomes or stakeholders' appraisal of TBLT on some indicator (e.g. perceptions, views, opinions, satisfaction, attitudes) and contributed a total of 52 samples. 29 studies compared TBLT with either a comparison or control group (between-group studies), 10 studies looked at the effect of TBLT on a single group over time (within-group studies), and 13 studies reported stakeholder appraisal on indicators (perception studies). The 29 between-groups studies included a total of 66 effect sizes ($M = 2$), and the within-groups studies included 13 effect sizes ($M = 1$). One study (González-Lloret & Nielson, 2015) reported effects for both between- and within-group designs. The total sample size across studies was $n = 5{,}965$, and mean and median samples sizes were 110 and 61. Mean and median samples sizes for both treatment and control groups were 45 and 46. One study (Murakami, Valvona, & Broudy, 2012) did not report sample size. All studies were published between the years 1998–2016 (except: Birjandi and Malmir, n.d.). Figure 1 below shows the number of included studies by publication year. This figure clearly demonstrates an increase in publications of TBLT implementation over time, indicating increasing interest in this domain in the field of applied linguistics.

2 Studies by coded feature

Tables 2–4 summarize the number of studies included in the meta-analysis by study, program, and methods features. Information in these three tables covers between-groups, within-groups, and perception studies. The results across study features (Table 2) show that, by and large, journal articles (both peer reviewed and non-peer reviewed) were the most frequent report type in the meta-analysis (62%). However, others included book chapters, dissertations, theses, one pdf of a PowerPoint presentation, and one government document. Studies were split in their peer-review status. Dissertations, theses, and the one PowerPoint pdf, having been given at a conference, were coded as peer reviewed.

Table 3 summarizes the number of studies coded according to certain program features. The target language for most studies was English ($k = 40$; 85%), and most studies were

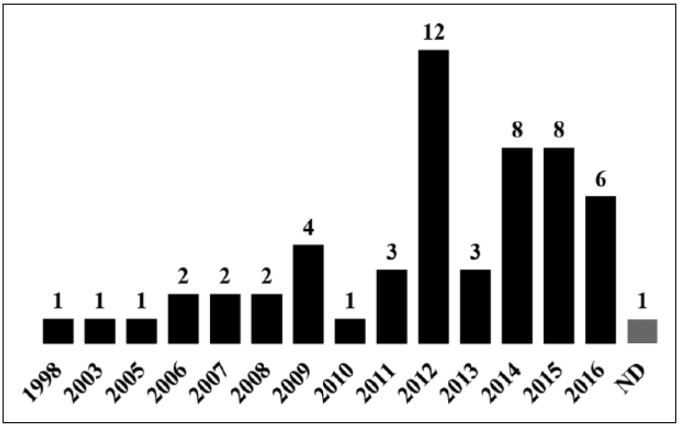


Figure 1. Number of studies by publication year.

Table 2. Included studies by coded study features.

Study type	*k*	Peer review	*k*	Funding	*k*
Journal article	29	Not peer reviewed/Unclear	23	No funding/Unclear	43
Book chapter	4	Peer reviewed	24	Funding	4
Dissertation	9				
Master's thesis	4				
PowerPoint	1				

Note. Dissertations, theses, and the one PowerPoint (having been discussed at a conference) were coded as peer reviewed.

conducted face-to-face (k = 44; 94%) in foreign-language settings (k = 44; 94%). Only a handful of included studies were in the computer-mediated mode (i.e. computer-mediated communication or CMS) or in second-language settings. Studies were conducted predominantly in K–12 (k = 18; 38%) or university settings (k = 25; 53%), with only 4 documenting task-based endeavors in language institutes. As for region, most studies were carried out in the Middle East (k = 17; 36%) or East Asia (k = 9; 19%), followed by studies in North America (k = 7; 15%) and Southeast Asia (k = 7; 15%). Studies report 11 different stakeholder languages with 9 noting a mix of other languages. Roughly a fourth of the studies in the sample mentioned having conducted (or programs based on) needs analyses or performing cycles of needs analysis or evaluation; studies that mentioned needs analyses were also coded as having performed at least one cycle of evaluation.

Lastly, Table 4 summarizes studies coded by four different methods features. 29 studies reported the proficiency level of participants, whereas 18 did not. Roughly half of included studies (k = 23) reported participants' proficiency levels assessed on one of five coded measures, while the other half (k = 24) did not report participants' proficiency levels. Not surprisingly given the domain, 19 studies included participants enrolled in

Table 3. Contextual and programmatic features.

Target language	*k*	Setting	*k*	Institution	*k*	Mode	*k*
English	40	Foreign language	44	K–12	18	Face-to-face (FTF)	44
Spanish	5	Second language	2	University	25	Computer-mediated communication (CMC) / Online	3
Mandarin	2	Other	1	Language institute	4	Multiple	1
Region	*k*	First language	*k*	Cycles	*k*	Needs analysis	*k*
Middle East	17	English	6	No cycles/ Unclear	38	None reported	38
North America	7	Spanish	1	Cycles	9	Conducted	9
Europe	3	Mandarin	7				
East Asia	9	Dutch	1				
South America	2	Korean	1				
Southeast Asia	7	Vietnamese	2				
Africa	1	Thai	4				
South Asia	1	Farsi	10				
		Arabic	6				
		Mixed	4				
		Other	5				

intact classes, and in 13 studies participants were assigned to particular courses based on availability. Twelve studies used some level of random sampling (though not random in the strict sense) to assign participants to courses or conditions. Studies reported effects on a wide range of outcome measures, including traditional CAF measures as well as on domain-specific assessments, such as a writing exam or institutional proficiency measure. The wide range of outcome measures used in the analysed studies precludes this feature from further analysis.

3 Outliers and publication bias

One between-groups study, Tale and Goodarzi (2015), had a *z*-score of 3.53, meaning it was a rather substantial outlier among other studies. A sensitivity analysis in CMA confirmed the *z*-score finding. The overall main effect for between-groups studies, using a random-effects model, was $d = 1.09$. After removing Tale and Goodarzi (2015), the effect was $d = 0.93$; any difference of 0.5 after study removal indicates that the effect in question is outlying. For main-effect calculations and moderator analyses, Tale and Goodarzi (2015) was removed. It is worth noting that Tale and Goodarzi (2015) reported an effect size of $d = 4.36$, which, after only 8.75 hours of treatment (over 20 weeks), is a large effect on a general proficiency outcome measure and one nearly double the effect reported for any other study. No within-groups studies had *z*-scores exceeding ±3.0.

Table 4. Included studies by coded methods features.

Proficiency level	*k*	Proficiency measure	*k*	Selection criteria	*k*	Pre-test	*k*
Beginner	9	Impressionistic	1	Availability	13	No pre-test	13
Intermediate	11	Norm-referenced	8	Intact	19	Pre-test used	34
Advanced	3	Institutional	13	Random	12		
Heritage	0	Achievement	1	Not reported/ Unclear	2		
Multiple	6	Not reported	24	Multiple	1		
Not specified	18	Self-assessment	0				

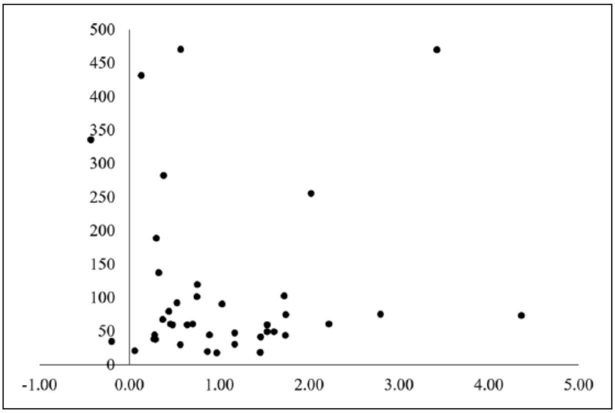
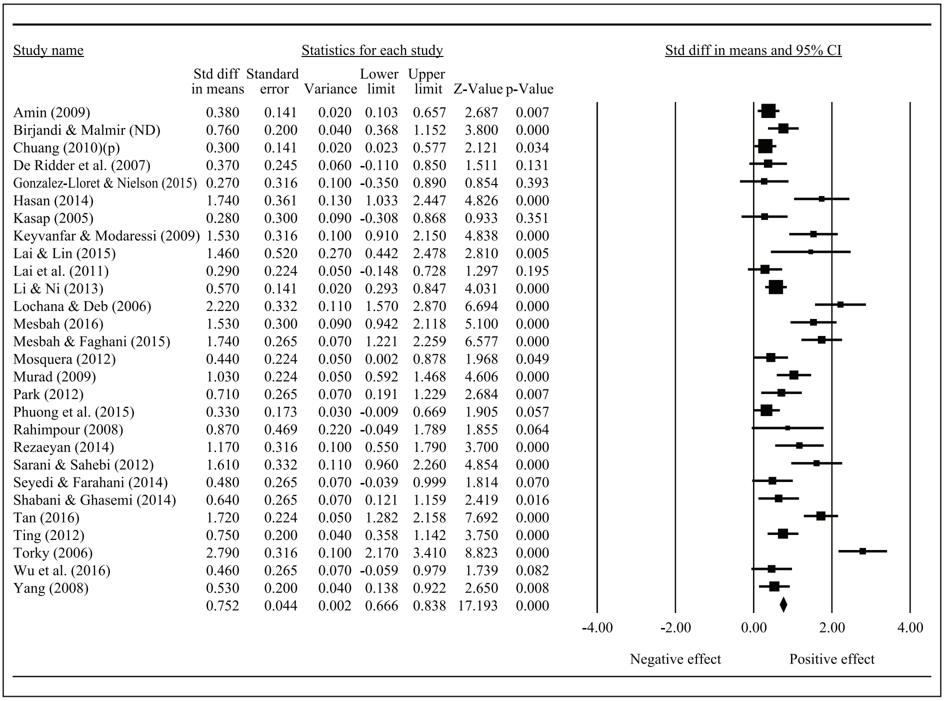


Figure 2. Funnel plot of study sample sizes (*y*-axis) and effect sizes (*x*-axis).

The funnel plot in Figure 2 above shows study sample size plotted against effect sizes based on within- and between-groups contrasts. Figure 2 shows that, despite attempts to include both published and unpublished literature in the meta-analysis, there is still substantial bias across included studies. Two aspects are worth noting. First, apart from a few high-powered studies, most studies include a total sample size less than n = 100. Second, only a handful of studies appear to the left of about d =.30 on the x-axis, indicating a lack of studies reporting small or negative effects of TBLT. Two studies, Al-Olaimat (2012) and Saiyod (2009), report negative effects.

4 Main effects of TBLT

Figures 3 and 4 summarize the main effects for between- and within-group studies. Although findings for the between-groups main effect are more stable, those for the within-groups studies are not and should be interpreted with caution. Figure 3 summarizes the main effect for between group studies. Recall that Tale and Goodarzi (2015), identified as an outlier via *z*-score and sensitivity analysis, was removed from

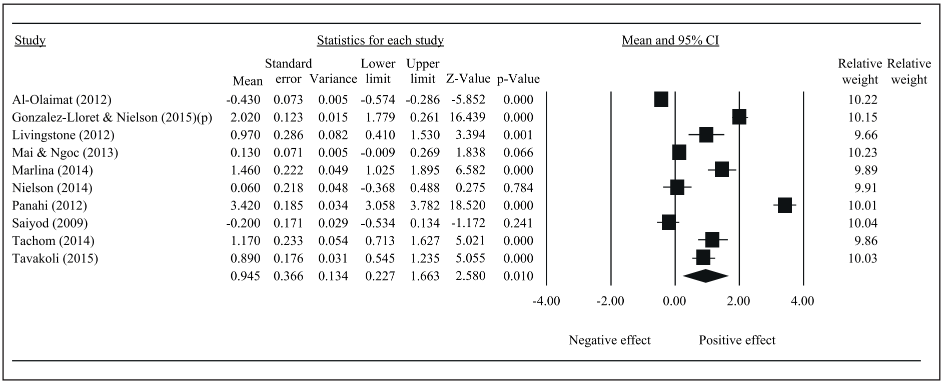


Study name	Std diff in means	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value
Amin (2009)	0.380	0.141	0.020	0.103	0.657	2.687	0.007
Birjandi & Malmir (ND)	0.760	0.200	0.040	0.368	1.152	3.800	0.000
Chuang (2010)(p)	0.300	0.141	0.020	0.023	0.577	2.121	0.034
De Ridder et al. (2007)	0.370	0.245	0.060	-0.110	0.850	1.511	0.131
Gonzalez-Lloret & Nielson (2015)	0.270	0.316	0.100	-0.350	0.890	0.854	0.393
Hasan (2014)	1.740	0.361	0.130	1.033	2.447	4.826	0.000
Kasap (2005)	0.280	0.300	0.090	-0.308	0.868	0.933	0.351
Keyvanfar & Modaressi (2009)	1.530	0.316	0.100	0.910	2.150	4.838	0.000
Lai & Lin (2015)	1.460	0.520	0.270	0.442	2.478	2.810	0.005
Lai et al. (2011)	0.290	0.224	0.050	-0.148	0.728	1.297	0.195
Li & Ni (2013)	0.570	0.141	0.020	0.293	0.847	4.031	0.000
Lochana & Deb (2006)	2.220	0.332	0.110	1.570	2.870	6.694	0.000
Mesbah (2016)	1.530	0.300	0.090	0.942	2.118	5.100	0.000
Mesbah & Faghani (2015)	1.740	0.265	0.070	1.221	2.259	6.577	0.000
Mosquera (2012)	0.440	0.224	0.050	0.002	0.878	1.968	0.049
Murad (2009)	1.030	0.224	0.050	0.592	1.468	4.606	0.000
Park (2012)	0.710	0.265	0.070	0.191	1.229	2.684	0.007
Phuong et al. (2015)	0.330	0.173	0.030	-0.009	0.669	1.905	0.057
Rahimpour (2008)	0.870	0.469	0.220	-0.049	1.789	1.855	0.064
Rezaeyan (2014)	1.170	0.316	0.100	0.550	1.790	3.700	0.000
Sarani & Sahebi (2012)	1.610	0.332	0.110	0.960	2.260	4.854	0.000
Seyedi & Farahani (2014)	0.480	0.265	0.070	-0.039	0.999	1.814	0.070
Shabani & Ghasemi (2014)	0.640	0.265	0.070	0.121	1.159	2.419	0.016
Tan (2016)	1.720	0.224	0.050	1.282	2.158	7.692	0.000
Ting (2012)	0.750	0.200	0.040	0.358	1.142	3.750	0.000
Torky (2006)	2.790	0.316	0.100	2.170	3.410	8.823	0.000
Wu et al. (2016)	0.460	0.265	0.070	-0.059	0.979	1.739	0.082
Yang (2008)	0.530	0.200	0.040	0.138	0.922	2.650	0.008
	0.752	0.044	0.002	0.666	0.838	17.193	0.000

Figure 3. Forest plot for between-groups studies' sample-weighted main effects.
Note. Tale and Goodarzi (2015) was deemed to be an outlier and therefore removed from calculation of the main effect.

calculation of the main effect. The overall effect of task-based programs, shown in the last line of the table, is medium–large $d = 0.93$; the main effect is indicated by the black diamond at the bottom of the forest plot. Each box and whiskers represents one study. The length of the whiskers indicates the precision of the study's observed effect, or the confidence interval (CI), and the size of the box is proportional to study sample size; the larger the box, the larger the study's sample and the more precise (narrow the whiskers) the findings. Of the 27 studies shown here, the CIs of eight pass through the vertical 0 line, indicating that the effect reported may not be statistically significant. The CI of the main effect is indicated by the width of the diamond; in this case, we can be fairly confident that, were we to sample another set of 30 primary task-based studies with similar features (study, method, and program), our effect would again be close to $d = 0.93$.

Figure 4 summarizes the main effect for within-groups studies included in the meta-analysis. Using a random-effects model, the main effect for within-groups studies is $d = 0.95$. There are several things to note. First, only ten studies are included in the main-effect calculation, the effects of which vary substantially. Second, given the number of studies and their diversity of effects, almost every study is an outlier; removing any one study results in major changes to the overall effect. By removing



Study	Mean	Standard error	Variance	Lower limit	Upper limit	Z-Value	p-Value	Relative weight	Relative weight
Al-Olaimat (2012)	-0.430	0.073	0.005	-0.574	-0.286	-5.852	0.000	10.22	
Gonzalez-Lloret & Nielson (2015)(p)	2.020	0.123	0.015	1.779	0.261	16.439	0.000	10.15	
Livingstone (2012)	0.970	0.286	0.082	0.410	1.530	3.394	0.001	9.66	
Mai & Ngoc (2013)	0.130	0.071	0.005	-0.009	0.269	1.838	0.066	10.23	
Marlina (2014)	1.460	0.222	0.049	1.025	1.895	6.582	0.000	9.89	
Nielson (2014)	0.060	0.218	0.048	-0.368	0.488	0.275	0.784	9.91	
Panahi (2012)	3.420	0.185	0.034	3.058	3.782	18.520	0.000	10.01	
Saiyod (2009)	-0.200	0.171	0.029	-0.534	0.134	-1.172	0.241	10.04	
Tachom (2014)	1.170	0.233	0.054	0.713	1.627	5.021	0.000	9.86	
Tavakoli (2015)	0.890	0.176	0.031	0.545	1.235	5.055	0.000	10.03	
	0.945	0.366	0.134	0.227	1.663	2.580	0.010		

Figure 4. Forest plot for within-groups studies' main effect given sample-size weighting.
Note. Findings should be interpreted with caution; addition/removal of any study results in ±0.05 change in main effect.

Panahi (2012), for instance, the effect is reduced to *d* = 0.68. Third, and crucially, the width of the main-effect diamond at the bottom of the forest plot is extremely wide, indicating a less stable result; given another ten within-groups studies with similar features, the observed effect could be expected to fall 95% of the time anywhere between *d* = 0.10–1.30.

5 Moderator analyses

Tables 5 to 7 below display the results for moderator analyses performed on only the between-groups set of studies. Within-groups and perceptions studies were too few to make any meaningful interpretations in relation to certain coded features. Also, even though additional study, program, and methods features were coded part of the initial coding scheme, not all features were observed in sampled studies. Therefore, the following tables show fewer features than those that were part of the original coding scheme. In discussing the results, discussion is limited only to those areas wherein the number of studies is sufficient to merit some interpretive speculation; where too few studies are present, no such attempts are made.

Table 5 shows the effects of TBLT program implementation as a function of report type and peer-review status. The average effect of studies reported in journal articles is *d* = 1.00, somewhat larger (not surprising given what is known of publication bias) than the other study types. However, a larger effect is observed for studies that are not peer-reviewed (*d* = 1.06) than studies that are peer-reviewed (*d* = 0.80).

Table 6 displays findings for the moderator analysis for several program features: institution type, region, needs analysis, and cycles. Within institution type, studies that were performed in K–12 institutions reported higher effects (*d* = 1.23) than those performed at universities (*d* = 0.80). The 12 Middle East studies observed effects around *d* = 1.31. The data for needs-analysis and cycles features are similar, as those studies that performed needs analyses frequently employed cycles of needs analysis or evaluation

Table 5. Moderator analysis for study features.

Study feature	*k*	Cohen's *d*	SE	95% Confidence intervals	
				Lower	Upper
Study type:					
Journal article	20	1.00	0.13	0.74	1.25
Book chapter	2	0.47	0.18	0.11	0.83
Dissertation	5	0.96	0.32	0.33	1.59
Master's thesis	1	0.28	0.30	–0.31	0.87
Peer review:					
Not peer review/Unclear	14	1.06	0.15	0.77	1.35
Peer reviewed	15	0.80	0.18	0.45	1.13

Table 6. Moderator analysis for program features.

Program feature	*k*	Cohen's *d*	SE	Confidence intervals	
				Lower	Upper
Institution type:					
K–12	11	1.23	0.23	0.79	1.67
University	15	0.80	0.13	0.53	1.05
Language institute	2	0.39	0.20	−0.01	0.79
Region:					
Middle East	13	1.31	0.18	0.95	1.67
North America	4	0.45	0.18	0.11	0.80
Europe	2	0.33	0.19	−0.04	0.71
East Asia	7	0.71	0.17	0.38	1.03
Southeast Asia	1	0.33	0.17	−0.01	0.67
South Asia	1	2.22	0.33	1.57	2.87
Needs analysis:					
None reported	25	1.01	0.13	0.75	1.32
Conducted	4	0.49	0.10	0.30	0.68
Cycles:					
No cycles	25	1.01	0.13	0.75	1.27
Cycles	4	0.49	0.10	0.30	0.68

(see Norris, 2015), and, while coding, performing a needs analysis to begin with counted as one cycle.

Table 7 provides summary results for the moderator analysis across a number of methodological features: data type(s), proficiency measure, pre-test, and selection criteria. Studies that relied exclusively on quantitative data to evaluate a task-based program or component observed a large effect (*d* = 1.25) compared to those that used a qualitative data-collection instrument apart from an assessment (*d* = 0.48). As for proficiency measure, studies that did not report the type of proficiency measure used to gauge stakeholder

Table 7. Moderator analysis for methods features.

Methods feature	*k*	Cohen's *d*	SE	Confidence intervals	
				Lower	Upper
Methods:					
No mixed methods	16	1.25	0.17	0.92	1.59
Mixed methods	12	0.48	0.07	0.34	0.62
Proficiency measure:					
Norm-referenced	4	0.82	0.33	0.18	1.46
Institutional	8	0.60	0.12	0.37	0.82
Not reported	16	1.12	0.18	0.77	1.48
Pre-test:					
No pre-test	5	0.48	0.17	0.15	0.80
Pre-test used	24	0.99	0.13	0.74	1.24
Selection criteria:					
Availability	1	0.27	0.32	−0.35	0.89
Intact classes	15	1.00	0.17	0.67	1.34
Random assignment	10	0.91	0.17	0.58	1.24
Not reported/Unclear	1	0.29	0.17	0.58	1.24
Multiple	1	1.03	0.22	0.60	1.47

ability had an effect of d = 1.12. Effects for non-mixed-methods studies and those for studies that did not report the proficiency measure are higher than the main effect for between-groups studies. In terms of some features of study quality, studies that reported using a pre-test had a large effect of d = 0.99. Finally, effects for random selection and intact classes are both large, at d = 0.91 and d = 1.00; the similarity in effects could be due, in part, to the fact that studies that reported random sampling did not sample randomly in the true sense; rather, participants were sampled from a larger pool of available or intact participants. In general, there is a tendency toward higher effects for methodologies and design features typically considered less preferable for sound research design.

Lastly, Table 8 provides a list of perception studies that reported participants' overall views, attitudes, opinions, and satisfaction towards task-based programs as a percentage. Though other studies included in the analysis used questionnaires as a means of gauging the effect of task-based implementation or evaluation, studies reported here did so using an overall question of some kind (e.g. 'To what extent were students satisfied with the task-based curriculum?') or provided the data necessary for calculation of overall perceptions (e.g. frequency counts for response options, descriptives, etc.). For instance, Lai, Zhao, and Wang (2011) noted that students in their study 'expressed satisfaction with the amount of learning in the class (5.33 on a scale of 7)' (p. 87). When the overall percentage was not reported, instances such as this made it possible to calculate the overall percentage; at all turns, we refrained from reading through items on scales to form our own interpretations of the findings. The average positive perception of TBLT (i.e. in terms of level of satisfaction, views, opinions, perceptions, etc.) after implementation, according to these studies, is a rating of 79%.

Table 8. Studies reporting overall perceptions of task-based programs implementations.

Study	Indicator	Percentage
Buitrago Campo (2016)	Satisfaction	52
Ogilvie & Dunn (2012)	Perceptions	63
Fattash (2013)	Perceptions	72
Lai et al. (2011)	Satisfaction	76
Wu et al. (2016)	Other	78
Mohamad (1998)	Attitudes	79
Ghaouar (2015)	Perceptions	80
Amin (2009)	Satisfaction	83
Lee (2016)	Satisfaction	83
Chuang (2010)	Satisfaction	87
Lin & Wu (2012)	Attitudes	87
Ji (2014)	Satisfaction	94
Iemjinda (2003)	Satisfaction	97

IX Discussion

1 Implications for TBLT implementation and evaluation research

The first research question sought to determine the effectiveness of TBLT program implementation for L2 learning outcomes. The finding of a medium to large between-groups effect of 0.93 supports the notion that program-wide implementation of TBLT is effective for promoting L2 learning above and beyond the learning found in programs with other, traditional or non-task-based pedagogies. This result, when examined alongside the results of research question 3, which found an average of 79% positive rating of TBLT by stakeholders after implementation, indicates that programs were both effective for L2 outcomes and positively received by learners and teachers. Taking the needs of key stakeholders into account during the evaluation of a new program or curriculum is critical for understanding the nuances of successful or unsuccessful implementation. These findings echo positive findings described in previous literature reviews of TBLT program implementations (Long, 2015) and provide robust evidence supporting TBLT pedagogy in contexts worldwide. However, an important caveat is that almost all of the studies included in this meta-analysis represent students learning English (85%) in foreign language contexts (94%), mostly in intact, face-to-face classrooms (94%). Therefore, the results of this analysis cannot purport to generalize to all language learning settings or languages. Furthermore, the presence of publication bias in the sample must be considered as a potential inflator of the overall effect sizes reported here.

The second research question aimed to examine the moderator variables that influenced the effectiveness of TBLT implementation. In terms of region and institution types, effect sizes were highest in programs conducted in the Middle East. Programs in East Asia ($k = 7$) reported overall medium effects. This finding is encouraging in light of previous studies that have questioned the compatibility of TBLT in East Asian countries (e.g. Carless, 2003) due to socio-cultural differences in learning and teaching styles.

Similarly, high effects were found for K–12 institutions, another context where TBLT's applicability has been questioned (Bruton, 2005). Large effects were also found for university settings. Overall, TBLT programs demonstrated positive effects for L2 outcomes in a wide range of contexts throughout the world in several different institutional settings, lending support to TBLT's applicability to language teaching in diverse contexts.

In terms of TBLT components, such as needs analysis and cycles of implementation, the results of the moderator analysis were surprising. The majority of studies analysed with between-groups comparisons did not report conducting a task-based needs analysis when designing and implementing the TBLT program ($k = 25$). The same 25 studies did not report implementing TBLT in cycles. However, these studies reported relatively large effects for TBLT. Only 4 studies reported conducting needs analyses, and those were the same 4 studies that also reported cycles of evaluation and implementation. To Norris (2009) and Long and Norris (2000), needs analyses that inform task selection and sequencing, materials and instruction development, are integral elements for a TBLT program. Given the limited number of studies reporting these elements, the finding that they report lower effect sizes should be considered tentative. This is especially true considering the finding that studies that were not from peer-reviewed sources, or from sources of an unknown quality, reported the highest effect sizes overall. Additionally, those studies that did not triangulate findings with qualitative methods, or utilized other questionable designs, such as non-random sampling or unreported proficiency levels, also reported larger effects. These trends lead us to interpret the findings of some studies with caution. However, the results of this analysis have indicated that even TBLT programs missing these elements can still be considered effective for L2 outcomes.

In general, the findings from the main effect and moderator analyses lend support to TBLT as an effective pedagogy in a variety of contexts for learners at a variety of levels. No large differences in effect sizes were found between institution type, and positive effects were found in diverse regions of the world. Criticisms of TBLT that claim it is unsuitable for East Asian countries or younger learners appear to be unsubstantiated (e.g. Carless, 2002, 2003; Deng & Carless, 2009). This finding lends further support to the positive findings of task-based interaction meta-analyses, which found similar medium to large effect sizes when comparing interactive treatments to control groups (Cobb, 2010, $d = 0.67$; Keck et al., 2006, $d = 0.92$; Mackey & Goo, 2007, $d = 0.75$).

2 Methodological implications

The current meta-analysis also examined methodological features of the studies under consideration, uncovering issues in reporting practices that echo findings in other areas of applied linguistics (Plonsky, 2013, 2014). On the positive side, most studies reported the use of a pre-test to measure L2 development, and those that utilized a pre-test reported the highest effects for L2 outcomes. Additionally, a surprising amount of studies reported using random assignment for treatment groups, a rarity in classroom-based research, and both those studies that utilized intact classes and those with random assignment reported similarly high effects.

On the other hand, 18 of the studies analysed here did not report the proficiency level of the learners in the TBLT program. Of those that did report on proficiency, 24 did not indicate how proficiency was measured. The issue of clarity in proficiency measurements has been raised previously (Norris & Ortega, 2000). For example, in Cobb's (2010) meta-analysis of task-based interaction, proficiency could not be treated as a moderator due to a lack of uniform classification system across studies (also see Thomas, 2006). This was also the case in the current meta-analysis. Issues of reporting were also present in the analysis of perception studies. Surveys of stakeholder perceptions took many forms, from overall satisfaction to attitudes and beliefs about TBLT, making comparisons across studies difficult, and at times under-informative. In his methodological synthesis of study quality in quantitative L2 research, Plonsky (2014) found missing data to be a persistent problem in SLA research and called for reform in reporting practices. The findings of the current study echo this result and extend it to the domains of TBLT implementation and language program evaluation research.

The majority of studies analyzed in the current meta-analysis were published in journals outside the mainstream publications in the field of applied linguistics. While a broad net was cast in order to avoid publication bias and to gain a more comprehensive view of the research in this domain, the finding that journal articles and non-peer reviewed studies contributed the largest effect sizes means that the results should be considered with an eye to study quality. Some studies were not clearly described as implementation and could have possibly been very long experimental treatments without the motivation of improving a pre-existing program. Future studies of TBLT program implementation and evaluation should aim to be more transparent in their reporting of TBLT programs so that more specific inclusion and exclusion criteria can be utilized and any studies involving experimental treatments can be more systematically avoided.

X Limitations and future directions

The current meta-analysis is the first to quantitatively synthesize the findings from TBLT implementation studies. However, the current study is not without limitations. Our selection criteria led to the exclusion of some key studies in the field that are valuable contributions to the body of TBLT implementation studies (see, for example, among others, McDonough & Chaikitmongkol, 2007; Van den Branden, 2006; Wichitwarit, 2014). Additionally, the domain of TBLT implementation is still growing. Therefore, the results of this first meta-analysis remain tentative until a greater, more reliable, body of research can be examined.

Despite the inclusion of unpublished literature from a variety of sources, the likelihood of bias in our sample remains. Much like in the domain of research on cognitive benefits of bilingualism (see de Bruin et al., 2014) favorable attitudes among researchers toward TBLT as an empirically investigated second language pedagogy might contribute to the bias for publishing findings that favor TBLT programs. Our results show that this finding may also carry over to unpublished sources with studies finding nonsignificant or negative results remaining unpublished or possibly unwritten. Future studies should attempt to contact known TBLT programs to investigate if other research or unpublished data sets exist. Furthermore, the current study was limited to work available in English. This could be considered a limitation due to TBLT's worldwide influence. Additionally, the majority of studies meta-analysed here were between-groups designs with unstable findings from the analysis of within-groups designs. The inclusion of a wider pool of within-group designs will bolster the results of future meta-analyses.

Finally, due to the limited sample size, not all features of TBLT or language program evaluation that could contribute to outcomes were able to be analysed as potential moderators. Another feature of frequent concern in the evaluation/implementation literature is the tension between the internal or external. This may refer to the origin of the evaluator, either an internal stakeholder or a jet-in, jet-out expert (Alderson & Scott, 1992) or the impetus for the evaluation, which can be externally mandated (e.g., for accreditation purposes) or driven by a desire on the part of teachers or administrators to know something about their program. However, this critical component that could contribute to the success of a program could not be examined due to limitations in sample size and reporting. Similarly, other factors including the target language of the learner, the modality of the course (computer-mediated communication or CMC vs. face-to-face or FTF) and setting (foreign vs. second language) could not be analysed as moderators due to the limited number of studies that were available. These variables would be exciting areas of investigation for future studies of TBLT programs.

Results from this meta-analysis suggest there are also many other areas for development in the domain of TBLT implementation research. Future studies should aim to compare other TBLT implementation in other contexts, including domains outside of K–12 and university institutions. Other interesting contexts, such as study abroad programs or other immersion contexts, are ripe for investigation as well. Such efforts will shed light on the extent to which the effects observed in this study might generalize to other contexts and demographics. With respect to program implementation, future studies should employ the features of cyclical language program evaluation (as suggested by Norris, 2009) and triangulate data from multiple and methods sources, including both L2 outcomes and stakeholder perceptions. This will enable future researchers to better understand the tension between what is recommended for TBLT implementation and what kinds of program elements are actually integral to program success.

Additionally, results from the methodological features analysis revealed that there is room to grow in both the methodology that is used to investigate the effectiveness of TBLT programs and when reporting those effects in published or unpublished studies. Future studies of TBLT implementation should take care to explicitly describe the ways in which the TBLT curriculum was design and executed. Without this information, it is difficult for readers or future meta-analysts to decide if a program is implementing TBLT appropriately and effectively for a given context. Issues in stakeholder perception measures also should be carefully considered in future studies. The studies uncovered for the current meta-analysis that reported stakeholder perceptions overwhelming did not triangulate findings with L2 outcome measures. Future studies would be improved by including mixed methodologies that would enable researchers to examine the relationship between stakeholder satisfaction and overall program effectiveness in promoting L2 outcomes.

Furthermore, the study highlighted the need for improvement in the reporting of statistical measures that are necessary to understand the outcomes of implementation and calculate effect sizes. Future researchers should take care to report features critical to interpreting findings such as: confidence intervals, means, *n*s, standard deviations, pre-test–post-test correlations, and exact *p* values. Studies without these features often times had to be excluded from the current meta-analysis if they impeded effect-size calculation. Without these features, research consumers should be skeptical of reported results, as they become difficult to interpret (for calls for more transparent reporting in SLA research, see Plonsky, 2013, 2014).

The meta-analysis reported on here provides a first look into the quantitative effectiveness of authentic TBLT programs. Despite the limitations described above, this study provides a foundation upon which future meta-analyses of TBLT programs and evaluation can build. As this study has shown, reports of TBLT implementations are increasing each year. Given the amount of interest TBLT has garnered in language programs across the globe, it seems that this domain of second language pedagogy will surely continue to grow in the coming years.

Acknowledgements

We would like to thank Luke Plonsky for his invaluable feedback and guidance on this project. We would also like to thank the students of Dr Plonsky's course, 'Research synthesis and meta-analysis in applied linguistics', who commented on various stages of this project.

Funding

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

Notes

In the spirit of synthetic ethics (e.g. Larson-Hall & Plonsky, 2015; Norris & Ortega, 2000, 2006), all materials from this study, including the coding scheme and raw data, can be accessed at the following website: https://sites.google.com/site/toddhavilandmckay (accessed November 2017).

References

(For studies that were included in the meta-analysis, see Appendix 2.)

Alderson, J.C., & Scott, M. (1992). Insiders, outsiders and participatory evaluation. In J.C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 25–58). Cambridge: Cambridge University Press.

Altschuld, J.W., & Watkins, R. (2014). A primer on needs assessment: More than 40 years of research and practice. *New Directions for Evaluation*, *144*, 5–18.

Becker, L. (1999). Effect size calculators [web page]. Available at: https://www.uccs.edu/~lbecker (accessed November 2017).

Beretta, A. (1992). Evaluation of language education: An overview. In J.C. Alderson & A. Beretta (Eds.), *Evaluating second language education* (pp. 5–24). Cambridge: Cambridge University Press.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive meta-analysis (Version 2.2.027) [Computer software]. Englewood, NJ: Biostat.

Brown, J.D. (1995). *The elements of language curriculum: A systematic approach to program development*. Boston, MA: Heinle & Heinle.

Bruton, A. (2005). TBLT for the State secondary school classroom? *Language Learning Journal*, *31*, 55–68.

Burwell, G., Nielson, K., & Gonzalez-Lloret, M. (2009). Evaluating a TBLT Spanish immersion program. Unpublished paper presented at the Third Biennial Conference on Task-Based Language Teaching, Lancaster, UK.

Butler, Y.G. (2011). The implementation of communicative and task-based language teaching in the Asia-Pacific region. *Annual Review of Applied Linguistics*, *31*, 36–57.

Carless, D. (2002). Implementing task-based leaning with young learners. *ELT Journal*, *56*, 389–396.

Carless, D. (2003). Factors in the implementation of task-based teaching in primary schools. *System*, *31*, 485–500.

Carless, D. (2004). Issues in teachers' reinterpretation of a task-based innovation in primary schools. *TESOL Quarterly*, *384*, 639–662.

Carless, D. (2007). The suitability of task-based approaches for secondary schools: Perspectives from Hong Kong. *System*, *35*, 595–608.

Carless, D. (2012). TBLT in EFL settings: Looking back and moving forward. In A. Shehadeh & C.A. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 345–358). Amsterdam: John Benjamins.

Cobb, M. (2010). Meta-analysis of the effectiveness of task-based interaction in form-focused instruction of adult learners in foreign and second language teaching. Unpublished doctoral dissertation, University of San Francisco, CA, USA.

Cooper, H. (2016). *Research synthesis and meta-analysis: A step-by-step approach*. 5th edition. Thousand Oaks, CA: Sage.

Cumming, G. (2001). Exploratory software for confidence intervals (ESCI) [Computer software]. Available at: http://thenewstatistics.com/itns/esci (accessed November 2017).

de Bruin, A., Treccani, B., & Sala, D. (2014). Cognitive advantage in bilingualism: An example of publication bias? *Psychological Science*, *26*, 99–107.

Deng, C., & Carless, D. (2009). The communicativeness of activities in a task-based innovation in Guangdong, China. *Asian Journal of English Language Teaching*, *19*, 113–134.

Ellis, R. (2016a). 'Focus on form': Past and present. Plenary at The Second Language Research Forum (SLRF), New York, USA.

Ellis, R. (2016b). Focus on form: A critical review. *Language Teaching Research*, *20*, 405–428.

González-Lloret, M., & Nielson, K.B. (2015). Evaluating TBLT: The case of a task-based Spanish program. *Language Teaching Research*, *19*, 525–549.

Hill, Y.Z., & Tschudi, S. (2008). A utilization-focused approach to the evaluation of a web-based hybrid conversational Mandarin program in a North American university. *CELEA Journal*, *31*, 37–53.

Iwashita, N., & Li, H. (2012). Patterns of corrective feedback in a task-based adult EFL classroom setting in China. In A. Shehadeh & C.A. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 137–162). Amsterdam: John Benjamins.

Jackson, D.O., & Suethanapornkul, S. (2013). The cognition hypothesis: A synthesis and meta-analysis of research on second language task complexity. *Language Learning*, *63*, 330–367.

Keck, C.M., Iberri-Shea, G., Tracy-Ventura, N., & Wa-Mbaleka, S. (2006). Investigating the empirical link between interaction and acquisition: A quantitative meta-analysis. In L. Ortega & J. Norris. (Eds.), *Synthesizing research on language learning and teaching* (pp. 91–131). Amsterdam: John Benjamins.

Klapper, J. (2003). Taking communication to task? A critical review of recent trends in language teaching. *Language Learning Journal*, *27*, 33–42.

Larson-Hall, J., & Plonsky, L. (2015). Reporting and interpreting quantitative research findings: What gets reported and recommendations for the field. *Language Learning*, *65*, 127–159.

Leow, R. (2016). ISLA: How implicit or how explicit should it be? Theoretical, empirical, and curricular/pedagogical issues. Colloquium presented at The Second Language Research Forum (SLRF), New York, USA.

Li, S., Ellis, R., & Zhu, Y. (2016). Task-based versus task-supported language instruction: An experimental study. *Annual Review of Applied Linguistics*, *36*, 205–229.

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 development* (pp. 77–99). Clevedon: Multilingual Matters.

Long, M.H. (2005). *Second language needs analysis*. Cambridge: Cambridge University

Long, M.H. (2015). *Second language acquisition and task-based language teaching*. Oxford: Wiley-Blackwell.

Long, M.H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, *36*, 5–33.

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.

Mackey, A., & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational interaction in SLA: A collection of empirical studies* (pp. 408–452). New York: Oxford University Press.

Markee, N. (1996). Making second language classroom research work. In J. Schachter & S. Gass (Eds.), *Second language classroom research: Issues and opportunities* (pp. 117–156). Taylor & Francis: New York.

McDonough, K., & Chaikitmongkol, W. (2007). Teachers' and learners' reactions to a task-based EFL course in Thailand. *TESOL Quarterly*, *4*, 107–132.

McManus, K., & Marsden, E. (2017). L1 explicit instruction can improve L2 online and offline performance. *Studies in Second Language Acquisition*, *39*, 459–492.

Norris, J.M. (2009). Task-based teaching and testing. In MH Long, & C.J. Doughty (Eds.), *Handbook of language teaching* (pp. 578–594). Oxford: Blackwell.

Norris, J.M. (2015). Thinking and acting programmatically in task-based language teaching: Essential roles for program evaluation. In M. Bygate (Ed.), *Domains and directions in the development of TBLT: A decade of plenaries from the international conference*. Amsterdam: John Benjamins.

Norris, J.M. (2016). Language program evaluation. *Modern Language Journal*, *100*, 169–189.

Norris, J.M., & Ortega, L. (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. *Language Learning*, *50*, 417–528.

Norris, J.M., & Ortega, L. (2006). *Synthesizing research on language learning and teaching*. Philadelphia, PA: John Benjamins.

Oswald, F.L., & Plonsky, L. (2010). Meta-analysis in second language research: Choices and challenges. *Annual Review of Applied Linguistics*, *30*, 85–110.

Patton, M.Q. (2008). *Utilization-focused evaluation*. Thousand Oaks, CA: Sage.

Plonsky, L. (2013). Study quality in SLA. *Studies in Second Language Acquisition*, *35*, 655–687.

Plonsky, L. (2014). Study quality in quantitative L2 research (1990–2010): A methodological synthesis and call for reform. *Modern Language Journal*, *98*, 450–470.

Plonsky, L., & Oswald, F.L. (2014). How big is 'big'? Interpreting effect sizes in L2 research. *Language Learning*, *64*, 878–912.

Plonsky, L., & Oswald, F.L. (2015). Meta-analyzing second language research. In L. Plonsky (Ed.), *Advancing quantitative methods in second language research* (pp. 106–128). New York: Routledge.

Plonsky, L., & Kim, Y. (2016). Task-based Learner Production: A Substantive and Methodological Review. *Annual Review of Applied Linguistics*, *36*, 73–97.

Prabhu, N.S. (1987). *Second language pedagogy*. Oxford: Oxford University Press.

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics*, *22*, 27–57.

Sasayama, S., Malicka, A., & Norris, J.M. (2015). Primary challenges in cognitive task complexity research: Results of a comprehensive research synthesis. Paper presented at the colloquium 'An international collaborative research network (CRN) on task complexity', Sixth International Conference on Task-Based Language Teaching (TBLT), Katolieke Universiteit Leuven, Belgium.

Shehadeh, A. (2012). Broadening the perspective of task-based language teaching scholarship: The contribution of research in foreign language contexts. In A. Shehadeh & C.A. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 1–20). Amsterdam: John Benjamins.

Shintani, N. (2011). A comparative study of the effects of input-based and production-based instruction on vocabulary acquisition by young EFL learners. *Language Teaching Research*, *15*, 137–158.

Skehan, P. (1996). A framework for the implementation of task-based instruction. *Applied Linguistics*, *17*, 38–62.

Stufflebeam, D.L. (1983). The CIPP model for program evaluation. In G.F. Madaus, M. Scriven, & D.L. Stufflebeam (Eds.), *Evaluation models: Viewpoints on educational and human services evaluation* (pp. 117–141). Norwell, MA: Kluwer.

Swan, M. (2005). Legislation by hypothesis: The case of task-based instruction. *Applied Linguistics*, *26*, 376–401.

Swan, M. (2011). Legislation by hypothesis: The case of task-based instruction. In M. Swan, *Thinking about language teaching: Selected articles 1982–2011* (pp. 90–113). Oxford: Oxford University Press.

Thomas, M. (2006). Research synthesis and historiography: The case of assessment of second language proficiency. In J.M. Norris & L. Ortega (Eds.) *Synthesizing research on language learning and teaching* (pp. 279–298). John Benjamins.

Towell, R., & Tomlinson, P. (1999). Language curriculum development research at university level. *Language Teaching Research*, *3*, 1–32.

Van den Branden, K. (2006). *Task-based language education: From theory to practice*. Cambridge: Cambridge University Press.

Wichitwarit, N. (2014). Essential elements contributing to the success of task-based implementation in an English classroom. *Panyapiwat Journal*, *5*, 47–62.

Widdowson, H.G. (2003). *Defining issues in English language teaching*. Oxford: Oxford University Press, 111–133.

Ziegler, N. (2016). Taking technology to task: Technology-mediated TBLT, performance, and production. *Annual Review of Applied Linguistics*, *36*, 136–163.

Appendix 1

Detailed inclusion/exclusion criteria.

1. Studies accessible in English.
2. Studies that reported on the implementation, evaluation, or assessment of a task-based program in either its weak or strong form (presence or absence of a task-based needs-analysis as one program element) based on the description of a program or through the inclusion of citations of task-based sources. Studies exclusively considering task-based needs analyses were excluded.

3. Studies that reported on the implementation or evaluation of an *intact* task-based program. Studies that reported on a single experimental treatment, and studies that did not last for at least one cycle of a program, or at least a semester, were excluded. Studies that did not report the length of treatment or the length of the program were also excluded. Studies that collected data from multiple disparate programs (i.e. different curricula, syllabi, etc.) were excluded.
4. To address research question 2, studies that compared the implementation of a TBLT curriculum with some other form of teaching, such as PPP or grammar translation approach, using a pre-test–post-test between-participants or within-participants design, were included. Studies that compared TBLT with other task-based forms of teaching, such as task-supported teaching were excluded.
5. To address research question 3, studies that measured changes in student interest, cognition, attitudes, or evaluations of the TBLT program on a questionnaire were separately included. Studies that reported ultimate satisfaction at the end of TBLT program were also included. Studies that reported only on student perceptions of certain aspects of a TBLT program, such as particular tasks or features, were excluded (note: it was not necessary for studies to meet the criteria for both research questions 1 and 2 to be considered for inclusion).
6. Studies that examined the effects of the TBLT program's implementation on some outcome measure, such as L2 learning or development, including production or performance, were included. For perception-based studies, studies that evaluated the implementation with student perceptions were additionally included.
7. Studies that did not include sufficient information (*n*-size, standard deviations) to calculate effect sizes were excluded if authors did not respond to email requests for missing data.

Appendix 2

Reference list of studies included in the meta-analysis.

Al-Olaimat, M.A. (2012). The effectiveness of task-based language learning approach in teaching English as a second language to the students at the vocational education development centre in Abu Dhabi, UAE. Unpublished doctoral dissertation, The British University in Dubai, UAE.

Amin, A.A. (2009). Task-based and grammar-based English language teaching: An experimental study in Saudi Arabia. Unpublished doctoral dissertation, University of Newcastle Upon Tyne, UK.

Birjandi, P., & Malmir, A. (n.d.). The effect of task-based approach on the Iranian advanced EFL learners' narrative vs. expository writing. *Iranian Journal of Applied Language Studies, 1*, 1–26.

Buitrago, & Campo, A.C. (2016). Improving 10th graders' English communicative competence through the implementation of the task-based learning approach. *Profile, 18*, 95–110.

Carmichael, S., Wu, K., & Lee, J. (2013). Designing and evaluating a genre-based technical communication course incorporating a task-based model of instruction. *Hong Kong Journal of Applied Linguistics, 14*, 20–44.

Chuang, Y. (2010). Task-based language approach to teach EFL speaking. Unpublished article. Available at: http://ir.csu.edu.tw/dspace/bitstream/987654321/1941/1/型教學法於英語口說教學之質性研究_12-16修改_.pdf (accessed November 2017).

Chuang, Y. (2012). Implementing task-based language approach to teach and assess oral proficiency in the college EFL classroom. Unpublished doctoral dissertation, Cheng Shiu University, Kaohsiung, Taiwan.

De Ridder, I., Vangehuchten, L., & Gómez, M.S. (2007). Enhancing automaticity through task-based language learning. *Applied Linguistics, 28*, 309–315.

Fattash, N.A. (2013). The effect of applying task-based approach on learning English in elementary schools from the teachers' perspectives in Tubas Governorate. Unpublished masters thesis, An-Najah National University, Nablus, Palestine.

Ghaouar, N. (2015). Developing first year EFL students' learning skills through adopting task-based learning in the study skills session. *Arab World English Journal, August*, 100–111. Available at: http://www.awej.org/index.php/2013-04-17-12-20-35/59-university-of-bejaia-international-conference-proceedings-2015/737-nesrine-ghaouar (accessed November 2017).

González-Lloret, M., & Nielson, K.B. (2015). Evaluating TBLT: The case of a task-based Spanish program. *Language Teaching Research, 19*, 525–549.

Hasan, A. (2014). The effect of using task based learning in teaching English on the oral performance of the secondary school students. *International Interdisciplinary Journal of Education, 3*, 250–264.

Iemjinda, M. (2003). Task-based learning and curriculum innovation in a Thai EFL context. Unpublished doctoral dissertation, University of Tasmania, Hobart, Australia.

Ji, K. (2014). Effects of a task-based approach to public speaking instruction. *Chinese Journal of Applied Linguistics, 37*, 21–32.

Kasap, B. (2005). The effectiveness of task-based instruction in the improvement of learner's speaking skill. Unpublished masters thesis, Bilkent University, Ankara, Turkey. Available at: http://www.thesis.bilkent.edu.tr/0002848.pdf (accessed November 2017).

Keyvanfar, A., & Modarresi, M. (2009). The impact of task-based activities on the reading skill of Iranian EFL young learners at the beginner level. *Journal of Applied Linguistics, 2*, 81–102.

Lai, C., & Lin, X. (2015). Strategy training in a task-based language classroom. *The Language Learning Journal, 43*, 20–40.

Lai, C., Zhao, Y., & Wang, J. (2011). Task-based language teaching in online ab initio foreign language classrooms. *Modern Language Journal, 95*, 81–103.

Lee, L. (2016). Autonomous learning through task-based instruction. *Language Learning & Technology, 20*, 81–97.

Li, G., & Ni, X. (2013). Effects of a technology-enriched, task-based language teaching curriculum on Chinese elementary students' achievement in English as a foreign language. *International Journal of Computer-Assisted Language Learning and Teaching, 3*, 33–49.

Lin, T. & Wu, C. (2012). Teachers' perceptions of task-based language teaching in English classrooms in Taiwanese junior high schools. *TESOL Journal, 3*, 586–609.

Livingstone, K.A. (2012). Task-based language teaching as a suitable didactic method for the teaching and learning of second and foreign languages. *Baraton Interdisciplinary Research Journal, 2*, 63–75.

Lochana, M., & Deb, G. (2006). Task-based teaching: Learning English without tears. *The Asian EFL Journal Quarterly*. September 2006 Special Conference Proceedings Volume, *8*, 140–164.

Mai, H., & Ngoc, T.B. (2013). Evaluating task-based syllabus for EFL learners. *BELT Journal, 4*, 58–85.

Marlina, N. (2014). The implementation of task-based language teaching to improve students' grammar mastery. Unpublished thesis, Sebelas Maret University, Surakarta, Indonesia.

Mesbah, M. (2016). Task-based Language Teaching and its effect on medical students' reading comprehension. *Theory and Practice in Language Studies, 6*, 431–438.

Mesbah, M., & Faghani, M. (2015). Task-based and grammar translation teaching methods in teaching reading comprehension to nursing students: An action research. *Aula Orientals, 1*, 319–325.

Mohamad, N. (1998). To investigate the effectiveness of a task-based approach to language learning in a university in Malaysia. Unpublished doctoral dissertation, University of Manchester, UK.

Mosquera, L.H. (2012). A research study on task-based language assessment. *Revista de Lenguas Modernas, 16*, 215–227.

Murad, T.M. (2009). The effect of task-based language teaching on developing speaking skills among the Palestinian secondary EFL students in Israel and their attitudes towards English. Unpublished doctoral dissertation, Yarmouk University, Irbid, Jordan.

Murakami, C., Valvona, C., & Broudy, D. (2012). Turning apathy into activeness in oral communication classes: Regular self- and peer-assessment in a TBLT programme. *System, 40*, 407–420.

Nielson, K.B. (2014). Evaluation of an online, task-based Chinese course. In M. González-Lloret & L. Ortega (Eds.), *Technology-mediated TBLT: Researching technology and tasks* (pp. 295–321). Amsterdam: John Benjamins.

Ogilvie, G., & Dunn, W. (2010). Taking teacher education to task: Exploring the role of teacher education in promoting the utilization of task-based language teaching. *Language Teaching Research, 14*, 161–181.

Panahi, A. (2012). Binding task-based language teaching and task-based language testing: A survey into EFL teachers and learners' views of task-based approach. *English Language Teaching, 5*, 148–159.

Park, M. (2012). Implementing computer-assisted task-based language teaching in the Korean secondary EFL context. In A. Shehadeh & C. A. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation*. Philadelphia, PA: John Benjamins.

Phuong, H.Y., Van den Branden, K., Van Steendamn, E., & Sercu, L. (2015). The impact of PPP and TBLT on Vietnamese students' writing performance and self-regulatory writing strategies. *International Journal of Applied Linguistics, 116*, 37–93.

Rahimpour, M. (2008). Implementation of task-based approaches to language teaching. *Pazhuhesh-E Zabanha-Ye Khareji, 41*, 45–61.

Rezaeyan, M. (2014). On the impact of task-based teaching on academic achievement of Iranian EFL learners: Case study: Female high school students in Yasuj). *International Journal of Language Learning and Applied Linguistics World, 7*, 476–493.

Saiyod, P. (2009). Effects of task-based English reading instruction on reading comprehension ability of elementary school students. Unpublished masters thesis, Chulalongkom University, Bangkok, Thailand.

Sarani, A., & Farzaneh Sahebi, L. (2012). The impact of task-based approach on vocabulary learning in ESP courses. *English Language Teaching, 5*, 118–128.

Seyedi, S.H., & Farahani, A. (2014). The application of task-based writing and traditional writing on the development of reading comprehension of EFL advanced Iranian learners. *International Journal of English Language Education, 2*, 225–240.

Shabani, M.B., & Ghasemi, A. (2014). the effect of task-based language teaching (TBLT) and content-based language teaching (CBLT) on the Iranian intermediate ESP learners' reading comprehension. *Procedia: Social and Behavioral Sciences, 98*, 1713–1721.

Tachom, K. (2014). Researching innovation in task-based teaching: Authentic use of professional English by Thai nursing students. Unpublished doctoral dissertation, University of Southampton, UK.

Tale, S.M., & Goodarzi, A. (2015). The impacts of task-based teaching on grammar learning by Iranian first grade high school students. *International Journal of Applied Linguistics & English Literature, 4*, 144–153.

Tan, Z. (2016). an empirical study on the effects of grammar–translation method and task-based language teaching on Chinese college students' reading comprehension. *International Journal of Liberal Arts and Social Science, 4*, 100–109.

Tavakoli, P. (2015). Development of language proficiency through task-based classroom instruction in study abroad context. Paper presented at the biennial conference on Task-Based Language Teaching, Leuven, Belgium.

Thompson, J.L. (2011). An evaluation of a university level English for tourism program. Unpublished masters thesis, Payap University, Chiang Mai, Thailand.

Ting, L. (2012). The implementation of task-based language teaching approach in EFL oral English teaching in art academy. *Overseas English, 8*, 90–92.

Tinker Sachs, G. (2005). The challenges of adopting and adapting task-based cooperative teaching and learning in an EFL context. In K. Van den Branden, K. Van Gorp, & M. Verhelst (Eds.), *Tasks in action: Task-based language education from a classroom-based perspective* (pp. 235–264). Newcastle-upon-Tyne: Cambridge Scholars Publishing.

Torky, S. (2006). The effectiveness of a task-based instruction program in developing the English language speaking skills of secondary stage. Unpublished doctoral dissertation, Ain Shams University, Cairo, Egypt.

Vásquez, J.M., Chaves, M.M., & Morales, C.G. (2016). The roles of the instructors in an ESP-task based language teaching course. *Actualidades Investigativas En Educación, 16*, 1–23.

Wu, X., Liao, L., & DeBacker, T.K. (2016). Implementing task-based instruction in ESP Class: An empirical study in Marine Engineering English. *Journal of Language Teaching Research, 7*, 936–945.

Xiongyong, C., & Samuel, M. (2011). Perceptions and implementation of task-based language teaching among secondary school EFL teachers in China. *International Journal of Business and Social Science, 2*, 292–302.

Yang, J. (2008). The task-based approach and the grammar translation method with computer-assisted instruction on Taiwanese EFL college students' speaking performance. Unpublished doctoral dissertation, Alliant International University, San Diego, CA, USA.