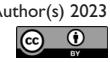
Article

LANGUAGE TEACHING RESEARCH

Meta-analysis to estimate the relative effectiveness of TBLT programs: Are we there yet?

Language Teaching Research
1–19



DOI: 10.1177/13621688231167573
journals.sagepub.com/home/ltr


Frank Boers
University of Western Ontario, Canada

Farahnaz Faez
University of Western Ontario, Canada

Abstract

Bryfonski and McKay published a meta-analytic review of reports on the effectiveness of task-based programs relative to other types of programs. The aggregated effect in support of task-based programs was substantial. However, Boers et al. re-examined the 27 comparative studies from which this effect size was calculated and argued that, with one exception, they did not serve the intended purpose of the meta-analysis well. However, in a recent publication, Xuan et al. argue that a good number of the primary studies used by Bryfonski and McKay are in fact suitable for a meta-analysis if the programs they assess are re-defined as task-supported rather than task-based programs. Xuan et al. then conducted a meta-analysis of 16 of the 27 comparative studies that were originally included in Bryfonski and McKay's analysis, confirming the benefits of programs that use tasks, albeit with a smaller aggregated effect size. The present article revisits these 16 studies and, in addition, examines a handful more recent ones published since Bryfonski and McKay's original meta-analysis. The conclusion remains that the field is not ripe yet for a meaningful meta-analysis of the relative effectiveness of either task-based or task-supported programs. Interpreting the outcomes of the meta-analytic endeavours reported so far is especially difficult because of a lack of clarity in both the primary study reports and the meta-analyses of what constitutes a task-supported program and of what is understood by 'task'.

Keywords

task-based language teaching, task-supported language teaching, meta-analysis, inclusion criteria, study design, reporting quality, teacher beliefs

Corresponding author:
Frank Boers, Faculty of Education, University of Western Ontario, 1137 Western Road, London, ON N6G 1G7, Canada
Email: fboers@uwo.ca

I Introduction

Bryfonski and McKay (2019) published a meta-analytic review of research reports on the effectiveness of task-based programs (not single lessons) relative to other types of programs. The aggregated effect in support of task-based programs was substantial ($d = 0.93$). However, Boers et al. (2021) re-examined the 27 comparative studies from which this effect size was calculated, and argued that, with one exception, they did not serve the intended purpose of the meta-analysis well. Many of the studies were problematic owing to confounding variables, such as using assessment tools that disadvantaged the comparison groups. Another recurring problem noted by Boers et al. is that authors of several primary studies used 'task' to refer to activities which proponents of a task-based approach would consider language-focused exercises, not tasks. Tasks are communicative activities where language is used for purposes that we normally use language for in real life. We use language to exchange information, to reach an agreement, to solve a problem, to give and follow directions, to persuade people, to entertain people, to console someone, to express admiration or love for someone, and so on. We rarely use language just for the sake of using language. That so many of the activities that authors labelled 'tasks' in the primary studies did not have a purpose other than language practice was surprising because these reports did cite influential publications about task-based language teaching (TBLT), such as Willis (1996), Bygate et al. (2001), Ellis (2003), Nunan (2004), and Willis and Willis (2007). Apparently, some of the commonly used descriptors of 'task' can still be misunderstood. For example, when it is stipulated by TBLT proponents that a task should have 'a clear outcome' (that is, a clear communicative purpose), readers may not realize this refers to a non-linguistic outcome, but instead consider correct answers to a language exercise to be a clear outcome of the activity. When it is stipulated that there should be some sort of 'gap', readers may not realize this refers to the use of language for a genuine exchange of information or opinions but instead think of exercises where students are required to complete gapped sentences (for examples, see further below).

Regardless of the sources of misunderstanding, the unorthodox use of 'task' by some of the authors in Bryfonski and McKay's (2019) collection of primary studies makes it hard to evaluate other reports in the collection where authors characterize a program as task-based without providing any examples or descriptions of the activities in the program. Considering these issues with the reports included in Bryfonski and McKay's meta-analysis, Boers et al. (2021) suggested that the field was not yet ready for such a meta-analytic endeavour.

However, in a recent publication, Xuan et al. (2022) argued that the inclusion criteria used by Boers et al. (2021) were too stringent, and that a good number of the primary studies used by Bryfonski and McKay (2019) are in fact suitable for a meta-analysis if the programs they put to the test are re-defined as task-supported rather than task-based programs. While tasks are the principal building blocks of a task-based program, they are considered to constitute one essential component of a task-supported program, among a broader set of activities. A task-supported program typically includes a fair amount of planned language-focused work in conjunction with communicative activities, and it is an approach that more professionals around the world may be prepared to

adopt (Chen & Clare, 2017; Shehadeh & Coombe, 2012) than the 'strong' version of TBLT outlined in the 1980s by Michael Long (e.g. Long, 1985). Long (2015, pp. 5–7) in fact considered the previously mentioned influential publications (e.g. Ellis, 2003; Willis & Willis, 2007) to be representative of diluted versions of TBLT, indeed more properly labelled task-supported than task-based approaches.

Xuan et al. (2022) then conducted a meta-analysis of 16 of the 27 comparative studies that were originally included in Bryfonski and McKay's (2019) analysis, after excluding ones in which they detected flaws such as lack of pre-testing. The outcome of Xuan et al.'s (2022) meta-analysis confirmed the comparative effectiveness of task-supported programs, although the aggregated effect size was smaller ($g = 0.61$) than the effect reported by Bryfonski and McKay (2019). The new meta-analysis also found that the comparative advantage of said programs was particularly marked (1) when class sizes were relatively small, and (2) when the researcher was also the teacher of the courses. The former finding suggests that tasks are easier to implement effectively with small classes, which aligns with teachers' beliefs (e.g. Liu & Ren, 2021; Zheng & Borg, 2014). The latter finding suggests that teacher-researchers' enthusiasm regarding a promising approach that they wish to put to the test may influence the outcome of their study.

We fully agree that comparing task-supported programs to programs not featuring tasks at all or featuring very few tasks is a meaningful endeavour. This was in fact also the stance taken in Boers et al. (2021, p. 14). To ensure a valid comparison, however, it remains essential that the so-called task-supported programs include regular use of activities that are tasks, as understood in the literature. It is noteworthy in this regard, that neither Bryfonski & McKay (2019) nor Xuan et al. (2022) list this as an inclusion criterion. While recognizing that definitions vary somewhat, Boers et al. applied the four features of 'task' proposed by Ellis and Shintani (2013, p. 135), summarized and reordered here as follows:

1. There is a clear purpose (e.g. solving a problem; reaching an agreement about a dilemma) other than practicing language (because language use is a means to an end, not the end itself).
2. The focus is primarily on the content of messages[1] rather than on the language code.[2]
3. There is a communication gap between interlocutors, that is, learners exchange information or opinions rather than telling interlocutors what they already know.
4. The task instructions do not stipulate explicitly what language elements or patterns the students should use when performing the communicative activity (because this may turn the communicative activity into a language-focused exercise).

It is worth clarifying that the fourth feature does not preclude language-focused learning in the design of task-supported lessons. Following Willis and Willis (2007), for instance, a task should be preceded by directions about the communicative purpose of the activity and what the students are expected to do. This pre-task phase will typically include model language that students find useful to perform the task. There will also be a post-task phase, where the students are invited to reflect on their task performance and where

language problems are addressed, to equip the students with improved resources for a subsequent task. The students may thus be well aware that certain previously noticed or studied language items or patterns will help them perform the task even if they are not explicitly told to use them (although this likely depends on whether the items or patterns are felt to be essential to get the task done; see Loschky & Bley-Vroman, 1993). The point remains that in the actual task, using those items or patterns is a means to an end rather than an end itself.

When the examples or descriptions of so-called tasks in a primary study corresponded to none or only one of these four features, Boers et al. (2021) decided that the program could hardly be labelled task-supported (let alone task-based). When it was impossible to tell if activities corresponded to the above features because they were not described and no examples were included in the reports, it was felt prudent to exclude such reports as well (Boers et al., 2021, pp. 12–13). We do not think such criteria are too stringent, if the aim is to compare the outcomes of programs that use tasks to ones that do not. Without such criteria, the principal question addressed by the meta-analysis becomes:

> Do studies which evaluate the comparative effectiveness of a program which includes activities *that the authors for one reason or another call 'tasks'* tend to report an advantage for these programs?

We do not consider this particularly informative for either researchers or practitioners.

We still thought it was worth re-visiting the 16 primary studies that were retained in Xuan et al.'s (2022) new meta-analysis in case we had misread them before and had as a result excluded studies which were in fact eligible. Below, we provide these re-assessments. To be clear, we are not judging the quality of these studies, but just whether they serve the purpose of a meta-analysis which aims to compare the benefits of programs characterized by the regular implementation of tasks to the benefits afforded by other programs. Neither do we question the usefulness of meta-analytic reviews. We do call for (1) explicit mention among inclusion criteria of how the construct under examination is defined, and (2) meticulous reading of candidate studies to determine if they meet the inclusion criteria.

II Eligible?

We will review the reports that were included in Xuan et al.'s (2022) meta-analysis in the order in which they are referenced in Appendix 1 of their article, alphabetically by (leading) authors' names.

Amin (2009)

This is a dissertation reporting learning outcomes from a semester-long course which used an ESP textbook in both the 'task-based' condition and in the comparison condition, which the author labelled grammar-based learning (GBL). The difference was that in the latter condition, the textbook activities were tackled in a teacher-led manner, whereas in the so-called task-based condition, the students were asked to tackle the activities as groupwork:

> During each lesson, they were given copies of the unit (reading passage and exercise) to be covered in that lesson . . . Small work groups were then formed and asked to read and discuss the topic and its vocabulary and to complete the accompanying exercise. (p. 109)

> [T]he two methods were similar in focusing on form because of curriculum control . . . In the GBL class, rules were explained and then students completed fill-in-the-blanks exercises individually whereas in the TBL [task-based language] class, students worked in small groups to discuss and complete the exercises and then to create their own illustrative examples. (p. 147)

So, this author appears to consider groupwork to be the defining feature of tasks (see also Zheng & Borg, 2014). It is doubtful, however, whether many proponents of TBLT would agree that doing a language-focused exercise as groupwork suffices to turn the exercise into a task. Amin (p. 185) acknowledges this:

> [I]t was not clear whether they were in fact 'tasks' in the sense this term is used in the TBL literature. Consider for example the following activity used in the TBL class: In the following exercise, you will need to put the right word in blanks (students work in groups to deal with the exercise).

Amin (2009) thus furnishes evidence of the benefits of collaborative learning, but we need to wonder if this alone makes the study eligible for inclusion in a meta-analysis that aims to compare programs featuring tasks to programs not featuring tasks, because it appears neither program in Amin's comparative study did.

Chuang (2010)

This is a report (not published in a journal or edited volume, as far as we can tell) about learners' perceptions of a task-supported course. This was a study involving one group of learners who expressed their opinions of a task-supported course in comparison to the 'traditional' course they were used to. Perception studies are of course valuable, but it was surprising to see this one included in Bryfonski and McKay's (2019) meta-analysis of between-group contrasts, because there was neither a control nor a comparison group. It is a within-participant study. This matters because effect size evaluations differ between within-participant and between-participant studies (Plonsky & Oswald, 2014).

De Ridder et al. (2007)

This is a brief report about a one-year L2 Spanish-for-business course which was identical for two groups of learners for the most part of the school year but differed toward the end. One group were asked to individually read texts about Spanish companies and to collect more information about them with a view to presenting the information in an oral examination. The other group engaged in 'communicative practice' (but no details are provided), and their culminating assignment was to create an advertisement. The latter condition is labelled TBLT by the authors. Arguably, collecting information and sharing this with an interlocutor (albeit in an oral exam situation) is a bit task-like as well, at least

when the interlocutor does not already know the information. The difference in the two groups' culminating work is rather that the interaction in an oral exam can only to some extent be planned whereas a finished script for an advertisement is the outcome of planned work. An important methodological issue with this study is that the students' linguistic performance was compared regarding these different culminating assignments: The oral examination for one group and the advertisement for the other. Given that these involve different content and different task conditions, the outcome of this comparison should be interpreted very cautiously.

Kasap (2005)

This is an unpublished dissertation reporting the benefits of adding communicative activities to an existing course focusing on speaking skills. The existing course book, followed by the comparison group, contained scripted role plays. Extra unscripted role plays and other activities such as information gap tasks (e.g. using a map to give directions) were added for the 'task-based' group. Before and after the course, the students' speaking skills were assessed in an oral exam requiring them to perform unscripted role plays. No significant differences in learning gains were found between the two groups, but the 'task-based' group's gains were slightly larger in real terms, yielding a small effect in favour of this condition ($g = 0.285$, according to Xuan et al., 2022). It is doubtful if many of the additional activities in the 'task-based' condition would be considered genuine tasks by all TBLT proponents (for descriptions of the activities, many of which are language-focused exercises, see Kasap, 2005, pp. 44–45), but it is safe to say that they increased the amount of student–student interaction in class. A potential methodological issue may be that (according to the examples given in the dissertation) the comparison group only did the scripted role plays included in the course book, and so these students were less well prepared for a test where they needed to create their own dialogues.

Keyvanfar and Modarresi (2009)

These authors evaluated a task-supported approach to L2 reading with young learners of English as a foreign language (EFL). Two groups of students spent the first half of each class on the language-focused work stipulated by their textbook. In the second half, one group continued doing the exercises in the textbook, while the 'task-based' group engaged in alternative reading-based activities such as following textual instructions to create something and matching pictures with the place in the text where their referents were mentioned. It therefore seems likely that the 'task-based' group used reading materials to supplement the course book ('task papers were brought to the class'; p. 91), while the comparison group did very little reading. If the task-based group outperformed the comparison group in a reading test after completing the program, this may then be attributed to the greater amount of reading practice and not necessarily to the task-like nature of this practice. Besides, the post-test (p. 90) included response formats that closely resembled the activities used with the 'task-based' group (p. 91).

Lai and Lin (2015)

These authors examined the merits of meta-cognitive strategy training for optimizing the learning outcomes from task-supported online classes (labelled task-based in the article). The task-supported classes were identical for two groups of learners, but one group in addition received strategy training while the other did not. The independent variable of interest in this study is therefore not whether a program features tasks, and therefore this study does not serve the purpose of a meta-analysis which aims to compare learning outcomes from programs with vs. without tasks.

Lai et al. (2011)

These authors evaluated the benefits for L2 learners' speaking skills of adding communicative activities ('tasks') to an existing language course, by adding such activities for one student group but not for another. The test used to compare the two groups' speaking skills after their respective courses required students to describe a picture of someone's bedroom. One of the recurring activities in the task-supported but not the comparison treatment was to describe pictures, including a description of someone's bedroom. A difference in test performance may therefore be attributed in part to practice-test congruency again.

Li and Ni (2013, reprinted 2014)

These authors report how adding a multimedia-based component to an existing EFL program brought about greater learning gains than only following the regular program. Nothing is said about that regular program, and it is unclear what activities – if any – the students did while others tackled the multimedia-based lessons. The latter lessons are presented by the authors as task based. They revolve around two fictitious characters, one Chinese and one American, who befriend each other. The lessons start with a dialogue (e.g. the two characters introducing themselves or telling each other about their families) as a model for the students to perform the same activity. The tests designed by the authors to assess learning gains concerned the same theme (Li & Ni, 2014, p. 1380): 'The tests were also task-based in that students were asked to learn about an American peer and his family (via listening, reading, vocabulary and grammar) and then introduce themselves and their own families.' The 'task-based' group thus practiced what they were going to be tested on, while it is impossible to tell if the comparison group (in their regular program) engaged with similar themes and practice.

Li (2012)

This is a very brief report (just 2.5 pages) that mentions Willis' proposal for a three-stage lesson but does not give a single example or description of the activities used in the 'task-based' course. Since we do not know what sort of activities were considered tasks in this study, it feels prudent not to include its outcome in a meta-analysis. Surprisingly, the comparison group's scores (on a speaking test) were poorer on the

post-test than on the pre-test, but no information is provided about the scoring system used by the teacher-researcher.

Mosquera (2012)

This author examined the usefulness of adding speaking tasks (which he called assessment tasks) to a language-focused course for beginning learners of L2 Spanish. It is not clear what these tasks entailed, because all that is said about them is:

> The content for the task-based tests was dictated by the textbook syllabus and adapted according to real language demands. Students received a complete description of the task (workplan) and started preparing with classmates in the allotted preparation time. The task handouts contained the task description, performance guidelines, and assessment rubric (if necessary). (pp. 218–219)

We do learn from the report that the speaking post-test, at the end of the semester, was an oral exam called 'Job Interview 2', which suggests that a similar simulation activity (Job Interview 1?) had been done before in the 'task-based' class, and so there is an issue with practice-test congruency again. More generally, the 'task-based' group's better performance on the speaking post-test can be attributed to their greater amount of speaking practice, as acknowledged by the author (p. 222): 'Most of the tasks were aimed at speaking; therefore students were better trained to take this type of test.'

Park (2012)

This author describes a project where some of the units of EFL students' regular textbook were transformed into computer-assisted learning units. Students either followed their regular textbook throughout in the regular classroom or engaged with the modified units in a classroom equipped with computers. The target language items and features were similar in the original and the modified units, but the activities and learning conditions were different, and were called 'tasks' in the computer-assisted versions. It is hard to determine whether better outcomes brought about by the task-supported approach should be ascribed to the task-like nature of the activities or to the use of computers and the internet, or both. In any case, the results need to be interpreted with caution owing to the issue of practice-test congruency again. In one of the computer-assisted lessons, the students practiced writing emails to e-palls (p. 221) and a near-identical task was used as post-test (p. 223).

Phuong et al. (2015)

This is a study of the effects of incorporating TBLT principles in a writing course. The study found a positive effect on some aspects of writing (e.g. lexical richness) but not on some others (e.g. grammatical accuracy). We have found no grounds for questioning the eligibility of this study.

Shabani and Ghasemi (2014)

These authors report a comparison of the benefits of 'task-based language teaching' and 'content-based language teaching' on ESP learners' reading comprehension. Akin to some of the other reports, it is hard to get a clear picture of what the 'tasks' were, because all we learn about them is:

> [A] reading passage was taught based on TBLT and the class time was divided into three phases: pre-task, task cycle and post-task. In the pre-task phase, the researcher tried to activate the EFL learners' schemata related to the text and motivated the participants to read the passage. During task-phase, the students were engaged in completing different kinds of task, and in post-task phase, they gave a report through, for instance, repeating the tasks and practicing some formal and linguistic features of the text. (p. 1715)

It is also unclear whether the two treatment groups used the same texts (p. 1715). The content-based group read texts in their specialized domain of accounting, but this seems not to have been the case for the 'task-based' group. The post-test was a reading section of TOEFL, which is very unlikely to require mastery of accounting terminology to obtain a good score.

Seyedi and Farahani (2014)

These authors investigated how an approach to L2 writing instruction can improve students' L2 reading skills. In the 'task-based' program (p. 223) the students were given writing tasks borrowed from sources such as IELTS. The students were first asked to tackle the writing task without any guidance. Then they received explicit instruction about how to write the given type of text in conjunction with a model text that was read and analysed for relevant features (and so the writing classes included reading practice). Then they tackled the writing task again and received feedback, including peer feedback (which entails reading each other's texts). It is not immediately obvious why the above procedure should be considered 'task based', but it is safe to say that it was different from the comparison treatment, labelled 'traditional' in the report. However, it is impossible to tell from the report how much writing and reading was involved in this 'traditional' program, and whether the two programs involved the same amount of time in general. As to reading practice, it appears that the traditional program focused on literary texts (p. 224). The 'task-based' group outperformed the 'traditional' group on a reading post-test, but this was a section borrowed from TOEFL, which uses non-fiction texts in everyday language, not literary texts.

Tan (2016)

This author reports a comparison of the effects on EFL learners' reading comprehension of a grammar-translation method and 'TBLT'. Again, we find very little information about what the so-called task-based approach entailed. The information is limited to this vague description (p. 103), which may already look familiar:

> [T]he class time was divided into three phases, namely, pre-task, task cycle and post-task. In the pre-task phase, the researcher tried to activate the students' prior knowledge related to the reading passage and motivated them to read the passage. In task-cycle phase, the students were engaged in completing different kinds of tasks, and in post-task phase, they gave a report through, for instance, repeating the tasks and practicing some formal and linguistic features of the text.

The verbatim resemblance of this passage to the above quote from Shabani and Ghasemi (2014) is striking (and instances of such resemblance are noticeable elsewhere in the report). It is worth noting that Tan (2016) is the study yielding the largest effect size in support of 'TBLT' ($g = 1.707$!) of the 16 studies included in Xuan et al.'s (2022) meta-analysis. Excluding this report – which seems prudent, given the above observations – will naturally reduce the aggregated effect size computed in the meta-analysis.

Yang (2008)

This is an unpublished dissertation which compares a 'task-based' course with a focus on speaking skills to a grammar-translation course without any focus on speaking. In the former course, the students were also given access to computers, while in the latter course no computers were made available to the students. The post-test focused on speaking skills. Unsurprisingly, the students who had practiced speaking skills in their course outperformed those who had not. Speaking skills were simply not the objective of the other course.[3] In the absence of a comparison condition where speaking skills are practiced equally often but through activities that are less task-like, Yang's study does not demonstrate superiority of a task-supported course over other courses with the same learning objectives. It is not entirely clear why a grammar-translation course should be devoid of speaking activities, and why the students should be denied access to computers. Put differently, akin to some of the other studies reviewed here, it is unclear what the independent variable was.

Conclusion to re-examination of the reports

After this re-examination of the 16 reports which had been evaluated once before in connection with Bryfonski and McKay's (2019) meta-analysis, but which Xuan et al.'s (2022) new meta-analysis prompted us to re-visit, we must reiterate the earlier conclusion: Most of these primary studies are unsuitable for a meta-analysis that aims to determine the effectiveness of task-based/task-supported programs compared to other programs. Because Boers et al. (2021) examined three recent meta-analyses, there was insufficient space in that article for more than a handful examples to question the eligibility of the studies which were included in Bryfonski and McKay (2019). Apparently, this was not compelling enough. In the present article, we have therefore reviewed in more detail each of the primary studies from which Xuan et al. computed a new aggregated effect size, to make it easier for readers themselves to determine their suitability.

There will almost inevitably be cases of disagreement, and perhaps readers will consider one or the other candidate study eligible where we did not. For example, one might

argue that the issue of practice-test congruency in a study such as Kasap (2005) was not serious enough to compromise its results, and so one could argue that its (small) effect size should be added alongside that of Phuong et al. (2015), the one study whose eligibility we found no reason to question. Still, it is hard to imagine that a 'when in doubt, leave it in' stance would result in a large enough set of studies for the meta-analysis to be meaningful, especially because the studies are very diverse. It appears that the field is not ripe for such a meta-analysis.

It needs to be mentioned in this regard that Bryfonski and McKay's (2019) original meta-analysis was first published online in 2017 and does not include primary studies published since 2016. This invites the question whether many additional studies about the effectiveness of TBLT programs have been conducted since and whether such newer studies might be better suited. We turn to this question next.

III What's the latest?

Following steps like those used by Bryfonski and McKay (2019), we identified 14 additional reports whose titles and abstracts indicated they compared L2 learning outcomes from 'task-based' approaches (to be interpreted more broadly as task-supported programs) to learning outcomes from other instructional approaches (see Appendix 2). This collection excludes studies that only report within-participant comparisons (i.e. single-group studies) and that report only perceptions of TBLT (i.e. no actual learning outcomes). While it appears promising that the subject has kept attracting interest from researchers, the issues pointed out above with reference to the older studies unfortunately abound also in this collection of more recent ones.

For starters, the 'tasks' described in the method sections of several of these reports are not necessarily tasks as understood in TBLT circles. For example, Ni and Jingxia (2017, p. 204) offer the following examples of listening tasks: 'matching the phrases or words with correct pictures and filling in blank [sic] according to the transcript.' Yildiz and Senel (2017, p. 200) state that '[g]rammatical tasks require learners to use particular language items . . . [l]earners have to use some predetermined linguistic items.' In Wu (2018), the students were required to write sentences or paragraphs incorporating sets of pre-selected words. The writing exercises clearly served the purpose of vocabulary learning rather than communication. No communication purpose for the writing activities that students were asked to perform is mentioned in Kalifour et al. (2018) either. To be clear, we are not criticizing the use of these types of exercises per se. Research shows, for instance, that asking learners to incorporate newly learned lexical items in their own output is useful to help them retain this new knowledge (e.g. Zou, 2017). The point is that these activities are unlikely to be considered tasks in TBLT circles, and so adding studies in which they are nonetheless labelled 'tasks' to a meta-analysis about the benefits of TBLT would be questionable.

In various studies, the activities were content focused, in the sense that they concerned text comprehension (Chou, 2017), but this held true for both the 'task-based' and the comparison classes. For example, writing a summary of a text was considered a task in Madhkhan and Mousavi (2017, p. 123), but there is no mention of a purpose for doing this other than displaying text comprehension. An activity such as this could easily be

turned into a real task, for example, by asking learners to sum up a text for a peer who has not read the same text but who will find the information useful for a certain purpose. This would not be dissimilar from asking authors of a research paper to write an accessible one-page summary for non-specialists: an increasingly common real task in our discipline.

The principal reason why many of the authors considered certain activities 'tasks' is that they were done as pair work or groupwork (e.g. Kalifour et al., 2018; Page & Mede, 2018; Yegani & Jodaei, 2017; Yildiz & Senel, 2017), which echoes Zheng and Borg's (2014) findings about teachers' beliefs regarding TBLT. However, as mentioned before, tackling a language exercise collaboratively as such does not transform it into a task.

Virtually all the authors of the reports in this collection refer to Willis' (1996) and/or Ellis' (2003) models of a three-stage task-based lesson, but it is hard to tell precisely how they interpreted these models. Some certainly seem to consider the pre-task stage an opportunity for pre-teaching language items that learners are subsequently required to display knowledge of. The difference with a present–practice–produce (PPP) lesson, which also happens to be a three-stage lesson, can easily get blurred.[4] Given that the examples and descriptions of 'tasks' given by some authors cast doubt on whether the activities corresponded to what is understood to be tasks in TBLT circles, it cannot be taken for granted that the activities used by other authors did fit the bill, because they are often not described at all or described in very vague terms. Below are some examples of such vague reporting.

> The experimental group received the treatment which was using task-based language teaching based on the ongoing school program. (Dost et al., 2017, p. 249)

> Three task-based vocabulary quizzes (for experimental group) and three traditional vocabulary quizzes (for control group) were administered every 3 sessions. There were 30 multiple-choice test items in each quiz for control group and 30 vocabulary task items for the experimental group. (Hamzeh, 2016, p. 18)

> In experimental groups, the respective lesson started with a task which the students were supposed to work on so as to achieve an aim. Then, they were given the same reading text given to control groups to see how well they did. In fact, they had a purpose to read the text. (Setayesh & Marzban, 2017, p. 73)

Note that we found no additional information whatsoever about the nature of the 'tasks', the 'task-based quizzes', or the 'purpose' for reading in these studies and nor do many studies give sufficient information about the approaches to which 'TBLT' was compared. The comparison treatment is often just labelled traditional, and authors presume readers know what traditional instruction looks like in the authors' educational context. Lack of clarity about the comparison condition makes it hard to detect potential intervening variables as well. If a 'TBLT' treatment is found to bring about greater learning gains for a certain dimension of language knowledge or use, then this is not necessarily owing to the nature of the tasks, because the comparison treatment may have differed from the 'TBLT'

treatment in additional ways. For example, in a study by Page and Mede (2018) on the benefits of task-based instruction for vocabulary learning, it appears that the group who received 'traditional instruction' focused on grammar rather than vocabulary (p. 375).

Some reports are transparent enough to see that the use of tasks was not necessarily the independent variable that differing learning outcomes should be ascribed to. For example, Jaramillo Cherrez (2019) compared the use of technology-mediated tasks to a comparison treatment with neither technology nor tasks, and so it is difficult to isolate the comparative effectiveness of using tasks *per se*. In Morris (2017), the effect of task-based pragmatics instruction was compared to conditions without any pragmatics instruction (i.e. control conditions instead of comparison conditions), and so it is difficult to attribute the better learning outcomes to the precise nature of the pragmatics instruction (i.e. its task-like nature).

One study in this new collection, Harris and Leeming (2022), compares TBLT to PPP programs by manipulating the sequence of class activities. In the PPP lessons, the students first learned new language items, practiced these, and then engaged in an activity (or task) where the learned items are helpful for successful communication. In the TBLT lessons, the students tried the communicative activity (or task) before doing the language-focused learning and practice. The report is not about the overall learning outcomes from the respective programs, but instead compares the outcomes of one TBLT vs. PPP lesson, a lesson intending to improve learners' knowledge of spatial prepositions and including a two-way picture-description task. A similar picture-description task was given three months later, and students' use of prepositions during the student–student interaction was analysed. The study included only four student pairs per condition, however, and the authors understandably refrained from applying inferential statistics. Because Harris and Leeming's (2022) study did not evaluate benefits from a program but from one lesson within the program and because the study did not yield much numerical data, it is not an optimal candidate for a meta-analysis of the kind tried by Bryfonski and McKay (2019) and Xuan et al. (2022). The study could nonetheless serve as an example of how treatment conditions can be designed in ways that help to identify with greater precision what contrast between the conditions (in this case, the sequencing of the activities) makes a difference to learning gains.

This review of relevant studies since 2016 is unlikely to be exhaustive. We almost certainly overlooked some studies, and more will have become available after this article was written. It nonetheless seems safe to say that the pool of eligible candidates for a meta-analysis of the comparative effectiveness of task-supported programs has remained small.

IV Conclusions and possible directions

After re-visiting the studies which Xuan et al. (2022) decided to include in their new meta-analysis of the comparative benefits of task-supported programs and after exploring more recent research on the topic, we need to reiterate the earlier conclusion that our field is still not ripe yet for such a meta-analytic endeavour. Three recurring issues with the primary research are:

- non-transparent reporting, compromising the replicability of the studies;
- confounding variables, making it impossible to determine precise cause–effect relations, such as whether differences in learning gains should be attributed to the use of tasks or to something else; and
- a lack of clarity of what distinguishes tasks from exercises, reflected in the liberal use of the term by some authors.

Given the ambiguity of the term, it is vital for meta-analysts to be clearer as to what they themselves consider to be the characteristics of 'tasks' and include this explicitly in the list of inclusion criteria. Unless this issue is addressed in future research, an aggregated effect size is unlikely to be very informative for teachers, materials writers, and course designers.

Besides, language teachers and course designers may well find it more interesting to know what dimensions of learning (possibly vocabulary and pronunciation rather than grammar, fluency rather than complexity, and the development of effective use of interactional strategies) a certain instructional approach is particularly beneficial for, while some other aspects of L2 development may benefit from something else (e.g. Phuong et al., 2015). Xuan et al.'s (2022) use of moderator analysis is certainly a way forward in this regard. Naturally, for a moderator analysis to be able to properly compare the impact of an instructional approach on various dimensions of L2 development, the pool of eligible studies needs to be substantial. If the pool includes, for example, just three studies that measure a certain outcome (e.g. fluency), as was the case in Xuan et al., then any findings regarding this moderator variable must obviously be taken as very preliminary, because adding a few new studies may substantially alter the aggregated effect size. No, we're not there yet.

It is worth reiterating that the concern we have expressed here is not at all intended as criticism of task-based or task-supported language teaching. We have long advocated the use of tasks in our courses for pre-service language teachers and through publications (e.g. Faez & Tavakoli, 2018). However, it is important to recognize that more work needs to be done to make this advocacy truly evidence based.

Throughout this article, we have used 'comparison group' and 'comparison condition', although the authors of the primary studies included in the meta-analyses discussed here almost invariably use 'control group' and 'control condition'. Strictly speaking, a true control group does not engage in activities regarding the specific language items, patterns or skills that are the targets for learning in the experimental group (e.g. Loewen & Plonsky, 2016). Using a control group helps to estimate if (and how much) learning occurs even in the absence of an intervention, and this helps to put into perspective the amount of learning that occurs in an experimental condition. As far as we have been able to assess, the so-called control groups in most of the primary studies we examined did work toward the same targets, but in different ways from the experimental (i.e. 'TBLT') group. If so, the term comparison condition is more appropriate. A comparison condition is useful to evaluate if one instructional intervention is more effective (for a certain learning goal) than another. An effect size in favour of an experimental treatment is naturally likely to be larger if the contrast is with the outcome from a control (i.e. no-treatment)

condition than with the outcome from a comparison condition. One of the problems with the collection of studies examined here is that it is not always easy to determine whether the design involved a control condition or a comparison condition. Another problem, as repeatedly pointed out in our review of the individual studies, is that in several cases it concerned a hybrid condition, where learners did some work toward the same learning targets as the experimental group but to a much lesser extent (and sometimes only minimally). Because of this confound, it is impossible to determine if an aggregated effect size in favour of task-supported learning reflects the comparative effectiveness of the method as such or rather a difference in investment of time and effort. When designing an empirical study, researchers need to decide whether to compare two (or more) treatment conditions and whether to include a control (no-treatment) condition. In the former case, care must be taken to avoid confounding variables. Authors of primary studies should provide sufficient information so a meta-analyst can calculate separate aggregated effect sizes for three types of study design: (1) single-group, (2) treatment group vs. control group, and (3) different treatment groups.

As pointed out by Xuan et al. (2022), a lot of the classroom-based research on the usefulness of tasks is conducted by researchers who are also the teachers of the courses where the research is integrated. If lack of clarity around the nature of tasks is one of the recurring issues that makes it so hard to interpret the available body of research, then one way forward may be to better prepare teacher-researchers for this line of empirical inquiry by providing sufficient training not only about research design but also about task design. We concur with Erlam (2016) that it can be challenging to fully grasp the construct of 'task' based on how it is presented in some of the scholarly literature. Examples of how PPP lessons can be adapted to make them more task-oriented can be helpful to illustrate more concretely what distinguishes tasks from language-focused exercises (e.g. Le Diem Bui & Newton, 2022). Teachers' understanding of the construct can also be finetuned in professional development workshops by asking the participants to discuss which on a list of activities can be considered tasks instead of exercises. This could highlight the fundamental notion that tasks involve the use of language for the purposes that it serves in real life (Long, 2016), such as passing on information to someone who does not already have this information, reading a manual to figure out how to use a newly purchased device, persuading someone who does not already (fully) agree with you, brainstorming potential solutions to a problem, organizing an event, asking for help, entertaining someone, expressing how you feel about a situation or a person, and so on.

At the same time, it needs to be recognized that the situational context often matters in determining whether language use serves a 'real-life' purpose. Having learners take turns reading paragraphs of a prose text aloud in the language classroom while they all have the text before them serves no purpose apart from language practice. Neither does taking turns reading the lines of a scripted dialogue. By contrast, reading a bedtime story to a child does serve a clear purpose, and so does rehearsing lines of a play in preparation for an actual stage performance for a real audience. It is important for teachers to understand the purpose of using tasks in their classrooms and recognize that the distinction between task and exercise is not always as black and white as is portrayed in some of the

TBLT literature. There is a wide grey shade of classroom activities that can be considered more or less 'task-like'. Engaging teachers in reflective practice regarding the factors that make communicative activities purposeful will not only help them to implement a task-supported approach as teachers, but also to evaluate it as researchers, should they wish to do so.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iDs

Frank Boers https://orcid.org/0000-0001-7552-4931
Farahnaz Faez https://orcid.org/0000-0002-2823-9684

Notes

1. In the literature, this focus on the content of messages is often referred to as 'focus on meaning', while attention to the language code is called 'focus on form'. We refrain from using these terms because we have noticed that they risk being misunderstood. For example, in some of the articles we examined (e.g. Page & Mede, 2018), focus on meaning appears to have been interpreted as a focus on vocabulary, while focus on form concerns grammar. We hope that the alternative terms 'focus on content' (i.e. what is said) vs. 'focus on the language code' (i.e. how it is said) can help to avoid this confusion.
2. It is worth recognizing that we do sometimes engage in 'real-life activities' that focus on the language code. This happens, for example, when we play word games (e.g. scrabble), invent puns, and write poetry. Still, even these language-focused activities serve a purpose other than using language just for the purpose of using language: winning the scrabble game, making someone laugh, prompting an esthetic experience, and so on.
3. Associating TBLT specifically with speaking activities appears to be another common misconception of the approach (for discussion, see Ellis, 2009). Purposeful, content-focused use of language is of course not confined to speech.
4. The difference between a task-supported lesson and a PPP (present – practice – produce) lesson can certainly get blurred when language-focused instruction precedes instead of follows the actual task. For example, in an empirical study by Li et al. (2016), the following sequence was considered a task-supported lesson: (1) explicit explanations about a grammar pattern (passive voice), (2) grammaticality-judgement exercise, (3) story-reconstruction activity (Dictogloss; Wajnryb, 1990), (4) inventing an ending to the story and presenting this to class. Advocates of task-supported learning might argue that their lesson designs at least steer clear of the 'controlled' practice, that often makes up the second P ('practice') in a PPP lesson. In the above lesson, however, we do notice the presence of a find-and-correct-errors exercise (i.e. the sentence-level grammaticality-judgement exercise), and it is doubtful if a text-reconstruction activity (i.e. Dictogloss, also known as Grammar Dictation) has a clear communicative purpose. What makes the lesson task-supported is the final part, where the students aim to entertain an audience with an original story ending. Because Li et al. (2016) examined the effects of a single task-supported lesson instead of a task-supported program, it was not included in the meta-analyses. We mention the study here only to illustrate that the ambiguity of terms such as task-supported language teaching extends beyond the meta-analytic reviews discussed in this article.

References

Boers, F., Bryfonski, L., Faez, F., & McKay, T. (2021). A call for cautious interpretation of meta-analytic reviews. *Studies in Second Language Acquisition*, *43*, 2–24.

Bryfonski, L., & McKay, T.H. (2019). TBLT implementation and evaluation: A meta-analysis. *Language Teaching Research*, *23*, 603–632.

Bygate, M., Skehan, P., & Swain, M. (Eds.). (2001). *Researching pedagogic tasks: Second language learning, teaching, and testing*. Longman.

Chen, Q., & Clare, W. (2017). Contextualization and authenticity in TBLT: Voices from Chinese classrooms. *Language Teaching Research*, *21*, 517–538.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Ellis, R. (2009). Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics*, *19*, 221–246.

Ellis, R., & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition research*. Routledge.

Erlam, R. (2016). 'I'm still not sure what a task is': Teachers designing language tasks. *Language Teaching Research*, *20*, 279–299.

Faez, F., & Tavakoli, P. (2018). *Task-based language teaching*. TESOL Press.

Le Diem Bui, T., & Newton, J. (2022). Developing task-based lessons from PPP lessons: A case of primary English textbooks in Vietnam. *RELC Journal*, *53*, 203–215.

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.

Liu, Y., & Ren, W. (2021). Task-based language teaching in a local EFL context: Chinese university teachers' beliefs and practices. *Language Teaching Research*. Epub ahead of print 20 September 2021. DOI: 10.1177/13621688211044247

Loewen, S., & Plonsky, L. (2016). *An A-Z of applied linguistics research methods*. Palgrave Macmillan.

Long, M.H. (1985). A role for instruction in second language acquisition: Task-based language training. In Hyltenstam, K., & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77–100). Multilingual Matters.

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

Long, M.H. (2016). In defense of tasks and TBLT: Nonissues and real issues. *Annual Review of Applied Linguistics*, *36*, 5–35.

Loschky, L., & Bley-Vroman, R. (1993). Grammar and task-based methodology. In Crookes, G., & S.M. Gass (Eds.), *Tasks and language learning: Integrating theory and practice* (pp. 123–167). Multilingual Matters.

Nunan, D. (2004). *Task-based language teaching*. Cambridge University Press.

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

Shehadeh, A., & Coombe, C.A. (2012). *Task-based language teaching in foreign language contexts: Research and implementation*. John Benjamins.

Wajnryb, R. (1990). *Resource books for teachers: Grammar dictation*. Oxford University Press.

Willis, J. (1996). *A framework for task-based learning*. Longman.

Willis, D., & Willis, J. (2007). *Doing task-based teaching*. Oxford University Press.

Xuan, Q., Cheung, A., & Liu, J. (2022). How effective is task-based language teaching to enhance second language learning? A technical comment on Bryfonski and McKay (2019). *Language Teaching Research*. Epub ahead of print 2 November 2022. DOI: 10.1177/13621688221131127

Zheng, X., & Borg, S. (2014). Task-based learning and teaching in China: Secondary school teachers' beliefs and practices. *Language Teaching Research*, *18*, 205–221.

Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research*, *21*, 54–75.

Appendix 1

The 16 primary studies in Xuan et al.'s meta-analysis

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

Chuang, Y.Y. (2010). Task-based language approach to teach EFL speaking. *Airiti Library*. Epub 1 December 2010. DOI: 10.6425/JNHUST.201012.0037

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

Kasap, B. (2005). The effectiveness of task-based instruction in the improvement of learner's speaking skill. Unpublished Master's thesis, Bilkent University, Ankara, Turkey.

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.

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

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. [Reprinted in Li, G., & X. Ni (Eds.) (2014), *Computational linguistics: Concepts, methodologies, tools, and applications* (pp. 1374–1390). IGI Global.]

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

Park, M. (2012). Implementing computer-assisted task-based language teaching in the Korean secondary EFL context. In Shehadeh, A., & C.A. Coombe (Eds.), *Task-based language teaching in foreign language contexts: Research and implementation* (pp. 215–240). John Benjamins.

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

Seyedi, S.H., & Farahani, A.A.K. (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.

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.

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.

Appendix 2

Additional studies

Chou, M.-H. (2017). A task-based language teaching approach to developing metacognitive strategies for listening comprehension. *The International Journal of Listening*, *31*, 51–70.

Dost, I.N., Bohloulzadeh, G., & Pazhakh, A. (2017). The effect of task-based language teaching on motivation and grammatical achievement of EFL junior high school students. *Advances in Language and Literary Studies*, *8*, 243–259.

Hamzeh, A. (2016). Investigating washback effects of task-based instruction on the Iranian EFL learners' vocabulary learning. *English Language Teaching*, *9*, 16–21.

Harris, J., & Leeming, P. (2022). Speaking proficiency in EFL classrooms: Measuring the differential effect of TBLT and PPP teaching approaches. *International Review of Applied Linguistics*. Epub ahead of print 3 October 2022. DOI: 10.1515/iral-2022-0082

Jaramillo Cherrez, N.V. (2019). Examining the impact of technology-mediated oral communicative tasks on students' willingness to communicate and communicative performance. Unpublished doctoral dissertation, Iowa State University, Ames, IA, USA.

Kalifour, R., Mahmoudi, E., & Khojasteh, L. (2018). The effect of task-based language teaching on analytic writing in EFL classrooms. *Cogent Education*, *5*, 1–16.

Madhkhan, M., & Mousavi, S.M. (2017). The effect of implimentation [sic] of TBLT in reading comprehension classes of Iranian EFL learners. *English Language Teaching*, *10*, 119–128.

Morris, K.J. (2017). Learning by doing: The affordances of task-based pragmatics instruction for beginning L2 Spanish learners studying abroad. Unpublished doctoral dissertation, University of California, Davis, CA, USA.

Ni, Z., & Jingxia, L. (2017). An empirical study on task-based listening teaching mode in Junior high school of China. *Advances in Language and Literary Studies*, *8*, 203–212.

Page, M.H., & Mede, E. (2018). Comparing task-based instruction and traditional instruction on task engagement and vocabulary development in secondary language education. *The Journal of Educational Research*, *111*, 371–381.

Setayesh, M., & Marzban, A. (2017). The impact of task-based language teaching on the development of Iranian EFL learners' ESP reading comprehension skills. *Advances in Language and Literary Studies*, *8*, 70–76.

Wu, R. (2018). The effects of TBLT's strong form and weak form on ESL students' vocabulary acquisition. *Theory and Practice in Language Studies*, *8*, 785–790.

Yegani, H., & Jodaei, H. (2017). The effect of task-based and topic-based speaking activities on speaking ability of Iranian EFL learners. *International Journal of English Language and Translation Studies*, *5*, 85–93.

Yildiz, M., & Senel, M. (2017). Teaching grammar through task-based language teaching to young EFL learners. *The Reading Matrix: An international Online Journal*, *17*, 196–209.